A Quasi-Continuous Interpolation Scheme for Pathways Between Distant Configurations

David J. Wales[1]*and Joanne M. Carr[1†]

University Chemical Laboratories, Lensfield Road, Cambridge CB2 1EW, UK

June 11, 2012

*dw34@cam.ac.uk

†jmc49@cam.ac.uk

Abstract

A quasi-continuous interpolation scheme is introduced for characterising physically realistic initial pathways from which to initiate transition state searches and construct kinetic transition networks. Applications are presented for peptides, proteins, and a morphological transformation in an atomic cluster. A simple interpolating potential is first defined, which preserves the covalent bonding framework for the biomolecules. This potential is used to identify an interpolating path by minimising contributions from a connected set of images along with terms corresponding to minima in the interatomic distances between them. This procedure, combined with repulsive terms between unconstrained atoms, helps to circumvent unphysical geometries in the line segments between images. The most difficult cases, where linear interpolation would involve chain crossings, are treated by growing the structure an atom at a time using the interpolating potential. A second optimisation phase then introduces a fraction of the true potential. Permutational alignment is achieved using a shortest augmenting path algorithm based on the local environment.

1 Introduction

The first step in analysing the kinetics for interconversion of different structures is usually the construction of an initial path between the regions of interest. For large-scale changes in morphology of a system treated in atomistic detail this construction may itself become a challenging task. Obvious problems arise for methods where the initial interpolation is a straight line path between the product and reactant states, as atom clashes may occur. For systems with chains of covalently bonded atoms described by force fields such as CHARMM,[1–4] AMBER,[5–7] or coarse-grained bead models,[8,9] more subtle problems can occur. Interpolation schemes that involve discrete images of the system that bridge the product and reactant configurations can produce chain crossings where the structure jumps between discrete images of the system in an unphysical manner. In particular, we have found that coarse-grained/united atom models can exhibit transition states and pathways corresponding to such chain crossings. These unphysical pathways generally lie significantly higher in energy than the physically relevant paths, where the chain-crossing is avoided. Nevertheless, they can lie below stationary points of the potential energy surface associated with the higher energy denatured or liquid-like phase of a finite system. This energy overlap makes the unphysical paths difficult to detect in automated procedures for constructing kinetic transition networks,[10–14] such as discrete path sampling.[11,15,16]

In a recent contribution[17] we described the energy landscape and folding pathways of a reduced representation for a knotted protein, where interpolation problems are particularly acute. To avoid chain-crossings we adopted a scheme where distances were collapsed and regrown. However, in subsequent work this approach has not proved sufficiently robust, and a more general scheme to treat the interpolation problem has now been developed. The real difficulty with interpolation via discrete images is that unphysical configurations can lie between them. Increasing the number of images, or the force constants of springs that connect them,[18–23] will not change the chain-crossing topology once it has been established. What is really needed is some way to account for and avoid unphysical geometries that lie between the images. Here we seek a solution that does not involve

a fully continuous path,[24–26] but augments a discrete set of images with terms that represent the worst case energetic contributions from the intervening straight line segments between them. We therefore refer to this approach as quasi-continuous interpolation (QCI). This QCI framework enables the interpolation problem to be framed entirely in terms of geometry optimisation, where a wide range of powerful techniques can be applied.[27,28]

To further improve the efficiency of the QCI scheme, and avoid extracting specific energetic terms from an external potential such as CHARMM and AMBER, an auxiliary potential is first constructed. This potential can be based on just the two endpoints in question, or can employ data for any number of local minima previously obtained for the same system. Constraints are applied between atoms that are always separated by a fixed distance, to within a predetermined tolerance, and repulsive interactions are defined between all other atoms, with a relatively short cutoff. For these pairwise potentials the smallest and largest distances between each pair of atoms on a synchronous straight line path between neighbouring images can be calculated analytically. The total energy can therefore be augmented by terms corresponding to an internal maximum in the auxiliary interpolating potential, as described in §2.1. This auxiliary interpolation potential, which incorporates some features of an elastic network model,[29] also enables the system to be grown a single atom at a time for the most difficult chain-crossing examples. A sensible guess for the interpolated position for the next atom can then be obtained in various ways from neighbouring atoms for which constraints exist (§2.2).

The above procedure has been applied to a number of different examples, namely the twelve residue tryptophan zipper peptide, trpzip2[30,31] (§4.1), an atomic cluster (§4.2), and a coarse-grained model of protein L (§4.3). The transition investigated for the atomic cluster is a change of morphology, and a simplified procedure is employed since there is no chain-crossing issue. Finally, extensive benchmarking is conducted for 2100 pathways in the amyloidogenic GNNQQNY peptide[32,33] (NH_3^+-Gly-Asn-Asn-Gln-Gln-Asn-Tyr-COO^-) and the tryptophan zipper peptide, trpzip1[30,34] (NH_3-Ser-Thr-Trp-Glu-Asn-Gly-Lys-Trp-Thr-Trp-Lys-CH_3), for both the AMBER and CHARMM potentials. The QCI procedure is compared with a number of alternative interpolation

schemes described in previous work, which exploit specific structural information for the two peptides. Optimal permutational alignment is essential in these calculations, and efficient methods based on overall and local distance metrics are described in §3. The general QCI approach is competitive with the schemes that exploit internal coordinates, and can provide physically meaningful interpolations for the more complex rearrangements where chain-crossings would otherwise result.

2 Theory

2.1 An Interpolation Potential

Here we define the auxiliary interpolation potential, which consists of constraint and repulsive terms for pairs of atoms, α, β. The complete potential for a set of interpolating images, labelled by superscripts i and j, with configurations $\mathbf{r}^i$ and $\mathbf{r}^j$, includes all the pairwise terms evaluated at each image. The potential also includes terms corresponding to local minima in the distance that appear for pairs of atoms on straight line paths between the images, as explained below. The latter feature provides the quasi-continuous part of the interpolation.

We denote the three-dimensional position vectors for atoms α and β in configuration i as $\mathbf{r}^i_\alpha$ and $\mathbf{r}^i_\beta$, etc. The line segments joining the atomic images in two configurations i and j can be written as

$$\begin{aligned} \mathbf{r}_\alpha &= \mathbf{r}^i_\alpha \sin^2\theta + \mathbf{r}^j_\alpha \cos^2\theta, \\ \mathbf{r}_\beta &= \mathbf{r}^i_\beta \sin^2\theta + \mathbf{r}^j_\beta \cos^2\theta, \end{aligned} \tag{1}$$

so the distance between atoms α and β as a function of the interpolation angle $0 \leq \theta \leq \pi/2$ is

$$\begin{aligned}
[d^{ij}_{\alpha\beta}(\theta)]^2 &= (\mathbf{r}_\alpha - \mathbf{r}_\beta)^2 \\
&= \mathbf{r}_\alpha^2 + \mathbf{r}_\beta^2 - 2\mathbf{r}_\alpha \cdot \mathbf{r}_\beta \\
&= |\mathbf{r}^i_\alpha|^2 \sin^4\theta + |\mathbf{r}^j_\alpha|^2 \cos^4\theta + 2\mathbf{r}^i_\alpha \cdot \mathbf{r}^j_\alpha \sin^2\theta\cos^2\theta + \\
&\quad |\mathbf{r}^i_\beta|^2 \sin^4\theta + |\mathbf{r}^j_\beta|^2 \cos^4\theta + 2\mathbf{r}^i_\beta \cdot \mathbf{r}^j_\beta \sin^2\theta\cos^2\theta - \\
&\quad 2\left(\mathbf{r}^i_\alpha \cdot \mathbf{r}^i_\beta \sin^4\theta + \mathbf{r}^j_\alpha \cdot \mathbf{r}^j_\beta \cos^4\theta + \sin^2\theta\cos^2\theta \left[\mathbf{r}^i_\alpha \cdot \mathbf{r}^j_\beta + \mathbf{r}^j_\alpha \cdot \mathbf{r}^i_\beta\right]\right)
\end{aligned} \tag{2}$$

Solutions of $\partial[d^{ij}_{\alpha\beta}(\theta)]^2/\partial\theta = 0$ always exist for $\sin\theta = 0$ and $\cos\theta = 0$, i.e. $\theta = 0$ and $\pi/2$ in the range considered. The third root of the cubic equation is θ^*, where

$$\begin{aligned}
\cos^2\theta^* &= \frac{\left(\mathbf{r}^i_\alpha - \mathbf{r}^i_\beta\right)\cdot\left(\mathbf{r}^i_\alpha - \mathbf{r}^i_\beta - \mathbf{r}^j_\alpha + \mathbf{r}^j_\beta\right)}{\left|\mathbf{r}^i_\alpha - \mathbf{r}^i_\beta - \mathbf{r}^j_\alpha + \mathbf{r}^j_\beta\right|^2}, \\
\text{or} \qquad \sin^2\theta^* &= \frac{\left(\mathbf{r}^j_\alpha - \mathbf{r}^j_\beta\right)\cdot\left(-\mathbf{r}^i_\alpha + \mathbf{r}^i_\beta + \mathbf{r}^j_\alpha - \mathbf{r}^j_\beta\right)}{\left|\mathbf{r}^i_\alpha - \mathbf{r}^i_\beta - \mathbf{r}^j_\alpha + \mathbf{r}^j_\beta\right|^2},
\end{aligned} \tag{3}$$

with $\cos^2\theta^* + \sin^2\theta^* = 1$.

The solutions for $\theta = 0$ and $\theta = \pi/2$ correspond to the limits of the range with

$$\begin{aligned}
[d^{ij}_{\alpha\beta}(0)]^2 &= |\mathbf{r}^j_\alpha|^2 + |\mathbf{r}^j_\beta|^2 - 2\mathbf{r}^j_\alpha \cdot \mathbf{r}^j_\beta \equiv [d^j_{\alpha\beta}]^2, \\
\text{and} \qquad [d^{ij}_{\alpha\beta}(\pi/2)]^2 &= |\mathbf{r}^i_\alpha|^2 + |\mathbf{r}^i_\beta|^2 - 2\mathbf{r}^i_\alpha \cdot \mathbf{r}^i_\beta \equiv [d^i_{\alpha\beta}]^2,
\end{aligned} \tag{4}$$

respectively, where $d^i_{\alpha\beta}$ is the distance between atoms α and β in configuration i. If the solutions for $\cos^2\theta^*$ and $\sin^2\theta^*$ lie between zero and one then $[d_{\alpha\beta}(\theta)]^2$ exhibits a minimum in the range $0 < \theta < \pi/2$, with local maxima at the endpoints (Figure 1). Otherwise, one endpoint is a local maximum and the other is a local minimum. For the internal minimum we find

$$[d^{ij}_{\alpha\beta}(\theta^*)]^2 = \frac{\left|\mathbf{r}^i_\alpha - \mathbf{r}^i_\beta\right|^2\left|\mathbf{r}^j_\alpha - \mathbf{r}^j_\beta\right|^2 - \left[\left(\mathbf{r}^i_\alpha - \mathbf{r}^i_\beta\right)\cdot\left(\mathbf{r}^j_\alpha - \mathbf{r}^j_\beta\right)\right]^2}{\left|\mathbf{r}^i_\alpha - \mathbf{r}^i_\beta - \mathbf{r}^j_\alpha + \mathbf{r}^j_\beta\right|^2}. \tag{5}$$

In the quasi-continuous interpolation (QCI) procedure the basic idea is to add contributions to the energy corresponding to every internal minimum in the distance $[d^{i\,i+1}_{\alpha\beta}(\theta^*)]^2$ for images i and $i+1$. We note that an analogous problem arises when computing the potential between

rod-like particles using the distance of closest approach.[35] The continuous path between endpoints consists of the M images and the straight line segments between them. The associated energy functional includes contributions from all the images, along with terms corresponding to minima in the distance between pairs of atoms that occur between images. If the interpolation potential, $V_{\rm int}$, is a sum over pairwise additive functions of the interatomic distances, V_1, V_2, etc., and we have an ordered chain of M configurations $\mathbf{r}_1$, $\mathbf{r}_2$,..., $\mathbf{r}_M$, then we minimise the sum

$$V_{\rm int} = \sum_{i=2}^{M-1} \sum_{p} \sum_{\alpha<\beta} V_p[d^i_{\alpha\beta}] + \sum_{i=1}^{M-1} \sum_{p} \sum_{\alpha<\beta}^{*} V_p[d^{i\,i+1}_{\alpha\beta}(\theta^*)], \tag{6}$$

where configurations 1 and M are the fixed endpoints, and the second term is over all the distances that exhibit internal minima between consecutive images. To minimise $V_{\rm int}$ we require the derivatives with respect to all the Cartesian coordinates of atoms in the different images. Let x^i_α be the x coordinate of atom α in image i. Then

$$\begin{aligned} \frac{\partial [d^{ij}_{\alpha\beta}(\theta^*)]^2}{\partial x^i_\alpha} &= \frac{2\left(\left[\left(\mathbf{r}^i_\alpha - \mathbf{r}^i_\beta\right)\cdot\left(\mathbf{r}^j_\alpha - \mathbf{r}^j_\beta\right)\right]^2 - \left|\mathbf{r}^i_\alpha - \mathbf{r}^i_\beta\right|^2 \left|\mathbf{r}^j_\alpha - \mathbf{r}^j_\beta\right|^2\right)\left[x^i_\alpha - x^i_\beta - x^j_\alpha + x^j_\beta\right]}{\left|\mathbf{r}^i_\alpha - \mathbf{r}^i_\beta - \mathbf{r}^j_\alpha + \mathbf{r}^j_\beta\right|^4} \\ &+ \frac{2\left[\left(x^i_\alpha - x^i_\beta\right)\left|\mathbf{r}^j_\alpha - \mathbf{r}^j_\beta\right|^2 + \left(\mathbf{r}^i_\alpha - \mathbf{r}^i_\beta\right)\cdot\left(\mathbf{r}^j_\alpha - \mathbf{r}^j_\beta\right)\left(x^j_\beta - x^j_\alpha\right)\right]}{\left|\mathbf{r}^i_\alpha - \mathbf{r}^i_\beta - \mathbf{r}^j_\alpha + \mathbf{r}^j_\beta\right|^2}, \end{aligned} \tag{7}$$

with analogous expressions for y^i_α and z^i_α.

Various possibilities have been tested for the construction of the interpolation function. For example, harmonic springs were included by adding contributions proportional to $|\mathbf{r}^i - \mathbf{r}^{i+1}|^2$ for adjacent images, as for previous chain-of-states methods.[18–23,36] Aside from the all-atom representation of trpzip1 with AMBER, the spring terms were not required. However, the examples provided for benchmarking in §4.4 employ the same set of parameters for each QCI calculation, and springs were added between QCI images with a force constant of unity in each case. The parameter sets are therefore closest to optimal for the AMBER trpzip1 benchmarks, but provide acceptable performance for the simpler examples as well.

In addition to the spring terms between images, the interpolation potential, $V_{\rm int}$, consisted of two further contributions, the first to maintain the conserved structure, and the second to keep

unconstrained atoms apart. The first, constraint, term uses harmonic springs between individual atoms to fix interatomic distances that are preserved to within a given tolerance in the two endpoints. Related constructions have previously been employed in elastic network-type[29] treatments of large-amplitude motion[37–54] and double-well Gō-type models.[55–59] Constraints were applied for atoms α and β only if the distances in the endpoints, $d^1_{\alpha\beta}$ and $d^M_{\alpha\beta}$, were both less than 5Å (5σ for the atomic cluster), and the change in distance $|d^1_{\alpha\beta} - d^M_{\alpha\beta}| < \Delta$. For the peptides and proteins considered in the present work the constraints need to define a percolating network, and the input parameter Δ was increased by 10% until this condition was achieved. Percolation was diagnosed by calculating the number of steps from one selected atom to all the others using a depth first search.[60] Here each constraint was taken as an edge connecting nodes of a graph defined by all the atoms. The reference separation for constrained atoms was taken as the average value $\overline{d}_{\alpha\beta} = (d^1_{\alpha\beta} + d^M_{\alpha\beta})/2$. For each constraint the corresponding term in the interpolation potential was set to

$$V_{\rm con}(d^i_{\alpha\beta}) = \begin{cases} \dfrac{\epsilon_{\rm con}\left[\left(d^i_{\alpha\beta} - \overline{d}_{\alpha\beta}\right)^2 - \left(C^{\rm con}_{\alpha\beta}\right)^2\right]^2}{2\left(C^{\rm con}_{\alpha\beta}\right)^2}, & \left|d^i_{\alpha\beta} - \overline{d}_{\alpha\beta}\right| > C^{\rm con}_{\alpha\beta}, \\ 0, & \left|d^i_{\alpha\beta} - \overline{d}_{\alpha\beta}\right| \leq C^{\rm con}_{\alpha\beta}, \end{cases} \tag{8}$$

where parameters $\epsilon_{\rm con}$ and $C^{\rm con}_{\alpha\beta}$ determine the strength of the constraint and the width of the interval around $\overline{d}_{\alpha\beta}$ for which $V_{\rm con}$ vanishes. $C^{\rm con}_{\alpha\beta}$ is chosen large enough so that the constraint term vanishes for the endpoints, and increased if necessary when new minima are found so that $V_{\rm con}$ also vanishes for these structures. Initially we simply set $C^{\rm con}_{\alpha\beta} = \left|d^1_{\alpha\beta} - d^M_{\alpha\beta}\right|/2$. The chosen form provides an even function of $d^i_{\alpha\beta} - \overline{d}_{\alpha\beta}$, for which $V_{\rm con}(d^i_{\alpha\beta})$ and its first derivative vanish at $d^i_{\alpha\beta} = \overline{d}_{\alpha\beta} \pm C^{\rm con}_{\alpha\beta}$. These properties were found to improve the efficiency in numerical minimisations of $V_{\rm int}$; distances that are shorter or longer than $\overline{d}_{\alpha\beta}$ are penalised in a symmetrical fashion.

Constraints were restricted to atoms sufficiently close in sequence, as defined by a parameter n_1. Here the sequence corresponds to the atom ordering in the CHARMM or AMBER input files. A value of $n_1 = 15$ was used in calculations for peptides and proteins when the constraints were defined on the basis of the separations in the two end points alone. This restriction avoids con-

straining atoms distant in sequence that happen to have very similar separations in the reference minima. When constraints are defined using a larger database of local minima this condition is very unlikely to occur, and n_1 was set to the total number of atoms.

The second, repulsive, contribution to $V_{\rm int}$ was applied to prevent unphysical geometries appearing along the line segments that define the interpolating path. Repulsions were only included for atoms far enough apart in sequence, defined by the condition $|\alpha-\beta| > n_2$ for a second sequence separation parameter n_2. In fact, setting $n_2 = 0$ proved to be satisfactory for all the tests described in §4. Repulsions were not allowed for atoms involved in any of the distance constraint terms described above. A cutoff distance was defined in terms of the minimum endpoint separation minus a small displacement, $0 < \delta \ll 1$, and an input parameter, C, with $C^{\rm rep}_{\alpha\beta} = \min(d^1_{\alpha\beta}-\delta, d^M_{\alpha\beta}-\delta, C)$. For this choice the atoms in question do not repel one another in the fixed images, 1 and M. For each pair of atoms satisfying these conditions the repulsion was defined as

$$V_{\rm rep}[d^i_{\alpha\beta}] = \begin{cases} \dfrac{\epsilon_{\rm rep}\left(\left(C^{\rm rep}_{\alpha\beta}\right)^3 - 3C^{\rm rep}_{\alpha\beta}\left(d^i_{\alpha\beta}\right)^2 + 2\left(d^i_{\alpha\beta}\right)^3\right)}{\left(C^{\rm rep}_{\alpha\beta}\right)^3\left(d^i_{\alpha\beta}\right)^2}, & d^i_{\alpha\beta} \le C^{\rm rep}_{\alpha\beta}, \\ 0, & d^i_{\alpha\beta} > C^{\rm rep}_{\alpha\beta}, \end{cases} \tag{9}$$

This form again ensures that $V_{\rm rep}$ and its gradient vanish at the cutoff $C^{\rm rep}_{\alpha\beta}$, which can depend upon the pair of atoms in question. Analytical derivatives of both $V_{\rm con}$ and $V_{\rm rep}$ are easily formulated using the chain rule and equation (7). A neighbour list was constructed for $V_{\rm rep}$ to avoid referencing pairs that lie outside the cutoff in every iteration. The list included all pairs of atoms approaching within a distance of $\zeta C^{\rm rep}_{\alpha\beta}$ for any image or intervening minimum. $\zeta = 2$ proved to be generally effective, with the neighbour list updated every 20 minimisation steps and whenever the number of active atoms changed.

When only two end points were specified constraints were only included for atoms close enough in sequence by the condition $|\alpha - \beta| < n_1$ for the sequence separation parameter n_1, introduced above. However, a more general procedure makes use of previously determined stationary points or configurations to define the constraint and repulsive terms. This approach was employed for the benchmarking described in §4.4 for the peptides GNNQQNY and trpzip1. Constraints were

applied using the percolation test described above for all the minima in the test set, and the corresponding pairs and distances were saved to define the potential in all subsequent runs. The pairs of atoms corresponding to repulsive interactions were also recorded, along with the reference constraint distances $\overline{d}_{\alpha\beta}$ and cutoffs $C^{\rm con}_{\alpha\beta}$ and $C^{\rm rep}_{\alpha\beta}$. All cutoffs were defined large enough to give $V_{\rm con} = 0$ and $V_{\rm rep} = 0$ for all the local minima encountered, and increased if necessary during the run. The reference distance for constrained atoms α and β was set to $\overline{d}_{\alpha\beta} = (\max_q d^q_{\alpha\beta} + \min_q d^q_{\alpha\beta})/2$ with cutoff $C^{\rm con}_{\alpha\beta} = (\max_q d^q_{\alpha\beta} - \min_q d^q_{\alpha\beta})/2$, where $\max_q$ and $\min_q$ refer to the maximum and minimum separation over the set of reference structures, q. The cutoff for repulsive interactions between any pair of atoms was set to the smallest separation between the pair encountered in any of the reference minima, or in any new minima subsequently encountered.

Since the interpolation potential is defined with respect to specific atom-atom contacts it is essential to employ consistent permutational isomers throughout the procedure. When multiple minima were used to define the potential all structures were aligned with respect to the first reference minimum using the procedures described in §3. When only two endpoints were available, the first one was taken as the reference.

The formulation of the interpolation potential to include contributions corresponding to local minima in the interatomic distance prevents the line segments between images from cutting through unfavourable regions of the true potential. Hence the problems of corner-cutting and sliding-down[61] that occur in discrete elastic band-type representations[18–23,36] do not appear directly for the QCI potential. To prevent images becoming too close together or too far apart new images were added where large gaps appeared, and images were combined according to distance thresholds. This procedure is described in the next section.

The overall scheme is probably most closely related to the geometric targeting (GT) approach,[62–64] where the aim is to generate plausible all-atom pathways for large conformational changes in proteins. For example, pathways for nitrogen regulatory protein C were recently obtained using this approach,[64] which has also been treated using discrete path sampling.[65]

The GT framework is based on geometrical constraints, as for the present QCI method, and

employs more detailed information about protein stereochemical propensities. The aim of QCI is somewhat different, since we are seeking an initial interpolation from which to characterise a connected path in terms of local minima and transition states of the potential energy surface. This calculation represents a single step in the construction of a kinetic transition network[10–14] and the analysis of kinetics using discrete path sampling (see §4).[11,15,16] Most of the computer time required for such calculations is used in refining transition states, and good initial guesses can speed up this process significantly. The QCI procedure therefore includes a minimisation phase that incorporates the true potential corresponding to the force field in question. Furthermore, the initial minimisation phase is general enough for application to molecules without a fixed set of covalent bonds, as illustrated for an atomic cluster in §4.2. Exploiting a computationally inexpensive potential for the interpolation phase provides an initial minimisation problem, in contrast to targeted molecular dynamics approaches,[66–68] and string methods.[69,70] QCI is the only scheme for which we have managed to obtain reliable automated interpolation of folding pathways for knotted protein structures.[17] It would be interesting to compare methodologies that employ fully continuous pathways[24–26] with the present approach, but this exercise lies beyond the scope of the present investigation.

2.2 Interpolation Procedure

Given two local minima as endpoints, and a true potential energy function, V_{true}, our objective is to locate a continuous, physically sensible pathway for further refinement. The first step in this procedure is to align the endpoints. Here we minimise the distance in $3N$-dimensional Euclidean space with respect to overall translation and rotation, as well as the feasible permutation-inversion operations.[36,71] The permitted permutations for a given biomolecule and force field (here AMBER[5–7] or CHARMM[1–4]) are obtained from an automated script.[71] The permutational alignment algorithms are described in detail in §3. The next step, also programmed in the `OPTIM` package,[72] is to define the percolating network of distance constraints in V_{con}. An initial guess could then be generated from a straight line interpolation between the two endpoints using all the atoms.

However, for the most difficult cases involving chain crossings it proved more effective to add one atom at a time to a growing chain. The pairwise design of $V_{\rm int}$ makes it straightforward to deal with a subset of atoms, where the constraints and repulsions are progressively turned on as atoms are added.

To start the initial interpolation procedure the two atoms involved in a distance constraint that move least between the endpoints in the fixed images $\mathbf{r}^1$ and $\mathbf{r}^M$ were identified. Having found the constrained atoms that minimise $|\mathbf{r}^1_\alpha - \mathbf{r}^M_\alpha| + |\mathbf{r}^1_\beta - \mathbf{r}^M_\beta|$, $M-2$ images were created for atoms α and β at regular intervals between the endpoints by linear interpolation. Atoms were then added sequentially when local convergence criteria for $V_{\rm int}$ were satisfied for the current set of active atoms, $N_{\rm active}$. It is usually possible to achieve paths with $V_{\rm int} = 0$, where all the distance constraints are satisfied within the tolerance $C^{\rm con}_{\alpha\beta}$, and all the repulsive interactions lie outside $C^{\rm rep}_{\alpha\beta}$. A new atom was therefore added when $V_{\rm int} < V^{\rm max}_{\rm int}$ and the root mean-square (RMS) gradient for the $3(M-2)\times N_{\rm active}$ degrees of freedom fell below a tolerance parameter $G^{\rm max}_{\rm int}$. Some element of randomness was introduced into this selection and the initial interpolation for the new atom, to avoid repeating failed interpolation attempts in the backtracking procedure described below. For each inactive atom, γ, the number of constraints to current active atoms, ϵ, and the sum of reciprocal reference distances, $s_\gamma = \sum_{\epsilon\in{\rm active}} 1/\overline{d}_{\epsilon\gamma}$, were calculated. The probability of adding atom γ to the active list was then chosen proportional to s_γ/t^4_γ, where t_γ was the number of times this atom had been tried before plus one. This scheme was chosen empirically after comparing various other possibilities.

To provide an interpolation for the next active atom several schemes were compared. For $N_{\rm active} \geq 3$ various possibilities were considered using active atoms to define a local orthogonal coordinate system. A sorted list of all the active atoms was constructed based on the average distance to the new atom for the two fixed endpoints. For three selected active atoms the components of the displacement vector for the new atom were calculated for the first endpoint. The interpolated position of the new atom was then obtained using the same components in the orthogonal coordinate system constructed from the same three active atoms in each image configuration

2, 3, ..., $M-1$. For $N_{\mathrm{active}} > 3$ active atoms were selected according to probabilities based on $1/o_\epsilon^2$, where o_ϵ was the order in which atom ϵ appeared in the sorted list of average endpoint distances. Various other selection schemes were also tested. For example, active atoms were also ordered according to how well their distance to the new atom was preserved in the endpoints, using $\left|d_{\epsilon\gamma}^1 - d_{\epsilon\gamma}^M\right|$. In addition, the V_{int} potential was also calculated using a coordinate system constructed from the three closest active atoms involved in constrained distances to the new atom. A fourth scheme simply interpolated the position based on the position of the active atom with the shortest $\overline{d}_{\epsilon\gamma}$ constraint distance in each image. The initial position of the new atom in each image was then taken as a displacement from this active atom, weighting the displacement vectors in images 1 and M according to the image position in the chain of configurations.

The initial interpolation for each new atom was chosen according to the lowest initial value of V_{int} for the four schemes described above. V_{int} was then minimised using the limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm of Liu and Nocedal,[73,74] reinitialising each time N_{active} increased. For examples involving possible chain crossings some cases were found where the convergence conditions on V_{int} and the RMS gradient were not achieved within the permitted number of refinement steps, $L_{\mathrm{max}}^{(1)}$. Such cases were treated by backtracking, removing the last N_{back} atoms added to active set, and adding one new one using the above scheme. N_{back} was increased by one for every convergence failure, but was not allowed to exceed $\min(20, N_{\mathrm{active}} - 2)$.

If N_{active} failed to reach N before a total of $L_{\mathrm{max}}^{(2)}$ LBFGS iterations then the interpolation was abandoned. Otherwise the interpolation proceeded for a further $L_{\mathrm{max}}^{(3)}$ steps using a new objective function $(1-x)V_{\mathrm{int}} + xV_{\mathrm{true}}$. The parameter x is the fraction of the true potential to use, which can be one. However, the overhead for using the relatively simple interpolating function in this stage was generally small, and retaining the fraction $1-x$ of V_{int} was helpful in preventing the interpolation from collapsing at the start of the refinement, when some of the derivatives of V_{true} could still be large.

Separate convergence parameters were also specified for the minimisation of $(1-x)V_{\mathrm{int}} + xV_{\mathrm{true}}$ in all $3N(M-2)$ degrees of freedom, with a maximum number of additional steps, $L_{\mathrm{max}}^{(3)}$. The ob-

jective here is not to converge the interpolation accurately, but to supply starting configurations for further refinement.[36,75] In fact, the value of the objective function must be allowed to rise during the minimisation procedure, since internal minima can appear (and disappear) in the interatomic distances. In the last refinement phase, where the potential considered is V_{true}, a doubly-nudged[36] elastic band[18–23] (DNEB) optimisation was conducted using starting images regularly spaced along the path obtained by the QCI procedure. The local maxima were then converged tightly to transition states using hybrid eigenvector-following,[20,76,77] and the connectivity verified by calculating approximate steepest-descent paths using energy minimisation following small displacements parallel and antiparallel to the eigenvector corresponding to the unique negative Hessian eigenvalue at the transition state.

Some further refinements of the interpolation procedure were used to improve efficiency, namely dynamic removal and addition of images and provision for frozen images. Starting from a specified number of images, M, images were removed if an image spacing became too close (less than a parameter D_{min}), and a bisecting image was added for image spacings that exceeded a specified distance, D_{max}, up to a specified maximum number of images, M_{max}. This dynamic adjustment was also permitted during the second phase minimisation of the function $(1-x)V_{\mathrm{int}} + xV_{\mathrm{true}}$. The corresponding degrees of freedom were simply removed from the LBFGS memory, and new degrees of freedom were initialised using average values from the neighbouring images. Images were frozen if the RMS gradient fell below a tolerance of $G_{\mathrm{int}}^{\mathrm{freeze}}$. The gradient terms were simply set to zero for frozen images with frozen neighbours, but calculated normally for frozen images with at least one unfrozen neighbour to check whether any images should be unfrozen.

For large systems, where the majority of atoms do not move much throughout the pathway, the interpolation can be speeded up significantly by treating these 'spectator' atoms separately. A simple linear interpolation was used for such atoms, and their interpolated positions were frozen during the initial phase, if the displacement between aligned endpoints fell below a specified cutoff. To define active constraints in this framework the depth first search employed to diagnose a percolating network was modified to admit all the frozen atoms as root nodes. Constraints and

repulsive terms were included between active and frozen atoms, but not between pairs of frozen atoms, thus reducing the computational cost of evaluating the interpolation potential. However, it was also necessary to keep a minimum number of active atoms, since some parts of the molecule can end up back almost where they started, but must be allowed to move significantly to accommodate the motion of interest. For the benchmarking in §4.4 linear interpolation was used if the aligned endpoint atomic distance for a given atom fell below 0.5 Å, and a minimum of 15 active atoms were retained. However, for the system sizes considered in the present work there is little advantage in identifying spectator atoms. A much larger gain in efficiency can result for molecules containing several thousand atoms, and we will report on pathway calculations for such systems elsewhere.

In trial calculations schemes involving a maximum number of distance constraints were considered for each active atom. In some cases the interpolation phase can succeed using all the constraints identified. However, the most efficient setting identified for the benchmarks in §4.4 used a maximum of six constraints per atom. Three constraints proved to be insufficient to prevent spurious rotation of trp groups in the benchmarking for trpzip1 (§4.4). Interpolations were also considered in which contributions from internal minima in $d^{ij}_{\alpha\beta}$ were only added for $V_{\rm rep}$. However, the results generally appear better when the corresponding terms are also included for $V_{\rm con}$.

One further refinement was employed in the present work, namely a dynamic adjustment of the force constant for the springs between images in the DNEB refinement phase. This adjustment improved all the benchmark timings reported in §4.4. The force constant was increased or decreased by 5% every five DNEB steps depending on the average spacing of the images. If the average deviation of the spacing from the mean exceeded 5% the force constant was increased, and otherwise it was decreased. Variable spring constants between different images have previously been used to increase the density of images near a transition state.[21] However, the uniform rescaling employed in the present work was employed to obtain a uniform image distribution more rapidly. For pathways involving both high energy and low energy transition states our observation is that the low-lying structure for the DNEB images can be unresolved if the spacings are not uniform enough. The rescaling helped to converge some pathways of this sort much more efficiently.

3 Permutational Alignment

The alignment of endpoints is critically important in obtaining efficient interpolation and pathways via geometry optimisation. The alignment of N atoms that define rigid body geometries to minimise the Euclidean distance in $3N$ dimensions is easily achieved using a quaternion procedure.[78] However, when permutations of atoms of the same element are also included there is no longer a deterministic solution to this problem. For a fixed centre of coordinates and orientation, the distance can be minimised using a shortest augmenting path algorithm,[79] and this is the procedure that has been used in the `OPTIM` program for previous studies of biomolecules[65,80,81] as well as clusters and condensed phases. Since it has not been described in detail before, this implementation of the *PERMDIST* keyword is outlined below. In fact, when the two endpoints are sufficiently distant in configuration space it is possible for the minimum overall distance to correspond to permutations that misalign permutable atoms. In the present work we have therefore introduced a local alignment algorithm, which ensures that unnecessary rearrangements of permutable atoms are avoided in pathway calculations.

Suitably symmetrised modifications of the CHARMM and AMBER potentials are used throughout to ensure that permutational isomers of every stationary point have the same energy.[36,71] The formulation of these force fields means that only limited subsets of atoms are actually permutable, giving a much smaller subset of the full nuclear permutation group. An auxiliary file specifying the allowed permutations is prepared via automated scripts.[71] The entries in this file contain the number of permutable atoms in the primary group and the number of other sets of permutable atoms associated with this primary set, which can be zero. The indices of the atoms in the primary set are then provided, followed by the indices of the pairs of atoms in the secondary sets, which must be permuted in correspondence with the primary atoms. Examples are given for alanine, phenylalanine and valine in Figure 2. The Euclidean distance metric is minimised for each group in turn, followed by overall translational and rotational alignment,[78] and this procedure is repeated until no further permutations result. The combination of permutational and transla-

tional/orientational alignment can produce a different result depending on the initial orientation, and it may therefore be necessary to start from a number of random initial orientations to find the global minimum for the Euclidean distance. For optimal alignment of atomic and molecular clusters we also consider the enantiomer for one of the endpoints; this cycle is generally omitted for biomolecules. Further complications occur if the system is subject to an electric or magnetic field,[82,83] a pulling potential,[84] or a central harmonic potential.[85–88] In these situations the overall translational and orientational alignment is restricted to operations that leave the Hamiltonian invariant.

In constructing kinetic transition networks via geometry optimisation it is also necessary to recognise permutation-inversion isomers whenever they are encountered. To achieve this goal we employ standard orientations before the shortest augmenting path procedure and diagnose equivalent isomers via 'zero' overall distance within a given tolerance. The atom that is furthest from the centre of coordinates is placed on the z axis and the atom that is then most distant from the z axis is moved into the xz plane by overall rotation about this axis. Due regard must be paid to the presence of external potentials. For example, only rotation about the axis of an applied electric or magnetic field is allowed in preparing the corresponding standard orientations. There is also a complication due to symmetry, which can be approximate symmetry in the local alignment procedure described below. In general, there could be more than one atom that is approximately the same distance from the centre of coordinates or from the z axis. Permutational alignment must therefore be attempted for standard orientations based on all the atoms in each of the corresponding orbits for the two endpoints, and a cutoff is required to decide whether the distances are considered the same or not. A tolerance of 0.2 Å generally seemed to be satisfactory.

Unfortunately, minimising the total Euclidean distance in $3N$ dimensional space can actually result in incorrect local permutational alignment. For example, the terminal NH_3^+ groups of lysine residues that undergo a large spatial displacement in space can sometimes give a slightly shorter overall distance if they are permuted cyclically relative to the rest of the residue. Worse still, the optimal overall distance could correspond to swapping a pair of hydrogen atoms in

the NH_3^+ group, which entails a very large barrier on the corresponding pathway. To overcome these problems a local permutational alignment procedure has been implemented, corresponding to the *LPERMDIST* keyword in `OPTIM`. The standard orientation and shortest augmenting path procedures (including loops over atoms in the same orbit) are performed locally for each group of permutable atoms and n nearest neighbours. The neighbours are chosen from the atoms outside the permutable group (i.e. excluding all primary atoms in the group, as defined above) inside a cutoff distance from the centre of coordinates of the permutable set for the two endpoints. A tolerance is also specified in terms of the minimal Euclidean distance for the permutable and neighbour atoms for an acceptable alignment. Atoms were added to a trial neighbour list one at a time in increasing order of the mean distance from the centre of coordinates of the permutable set, and accepted if the alignment threshold was achieved. Permutable neighbours were added together with the rest of the atoms in their primary permutable set. This procedure was continued until a specified number of neighbours was reached, or no more candidate atoms remained within the cutoff distance. For the peptide benchmarks described in §4.4 a maximum of ten neighbour atoms, a threshold of 1.0 Å for the overall alignment, and a cutoff distance of 5Å were found to give good results for the 2,100 pathways considered.

Although the permutational alignment procedures entail a considerable amount of bookkeeping, they are generally fast compared to operations that involve evaluating the potential, especially when only small subsets of atoms are permutable. This effort can greatly reduce the computational expense of finding connected pathways, especially for large systems with many equivalent atoms, such as atomic clusters and condensed phases. Further complications arise for systems with periodic boundary conditions, where the cell symmetries and the centre of coordinates must be accounted for. Our alignment procedures for condensed phases described by periodic supercells will be described in detail elsewhere.

4 Results

Several examples that employ the QCI procedure to obtain initial pathways are described in the subsections below. These initial paths would generally require further refinement using discrete path sampling[11,15,16] to build up a kinetic transition network.[10–14] For distant endpoints we usually find that the initial path can be improved significantly by additional pathway searches. Nevertheless, the efficiency with which kinetically relevant paths are located depends strongly on the initial path.

The QCI images and straight line segments connecting them were used to generate equispaced images for a doubly-nudged[36] elastic band[18–23] optimisation. Local maxima in the DNEB profile were then converged tightly to transition states using hybrid eigenvector-following.[20,76,77] The connectivity of each transition state was defined in terms of local minima obtained by energy minimisation using a modified version of the limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) algorithm[73,74] coded in `OPTIM`.[72] To produce a complete connected path may require further connection attempts. Pairs of unconnected minima were chosen for subsequent pathway searches using our missing connection algorithm,[89] which is based on Dijkstra's algorithm.[90] To define a metric for the missing connection algorithm we normally employ a function of the distance, minimised with respect to overall translation and rotation, as well as permutation of equivalent atoms. However, in the present work the metric was calculated as the Euclidean distance in $3N$ dimensional space, minimised with respect to translation, overall rotation, and local permutational alignment, plus $V_{\rm int}/10^3$, with the result exponentiated or raised to a power such as 10 or 20. The latter rescaling tends to disfavour single connection attempts between minima separated by large metric values relative to multiple connection attempts between 'closer' minima. The value of $V_{\rm int}$ was calculated using the two minima in question as the only images. This choice proved to be particularly important for pathways where a straight line interpolation would result in chain crossing. In this case the minimised distance for one chain passing through another chain could be quite small, whereas $V_{\rm int}$ should be large. Hence it is possible to work around the crossing via a

longer pathway that corresponds to lower energy and physically reasonable transition states. True transition states corresponding to unphysical chain crossings seem to exist for all the biomolecular force fields we have investigated, and cannot generally be eliminated on the basis of V_{true}. Dividing V_{int} by 10^3 means that the metric is dominated by V_{int} for unphysical straight line interpolations, but correlates mainly with the Euclidean distance for more realistic interpolations.

4.1 Trpzip2

As a straightforward test case we first considered the twelve residue tryptophan zipper peptide, trpzip2[30,31] (NH_3-Ser-Thr-Trp-Glu-Asn-Gly-Lys-Trp-Thr-Trp-Lys-CH_3). The force field employed was a symmetrised[36,71] version of CHARMM19[2] together with the implicit solvation potential EEF1.[91] Eleven different pairs of local minima were selected from a large database, which will be described elsewhere in terms of a detailed kinetic analysis. These pairs were selected as cases where relatively high barriers remained between low-lying minima for endpoint alignment schemes involving internal coordinates.[92] The QCI calculations therefore served to check whether pathways with lower barriers exist.

The QCI procedure either reproduced the pathway with the lowest barrier found using interpolation in internal coordinates,[92] or, in a few cases, improved upon it. The number of transition states in the resulting paths varied from one to 40, depending on the endpoints in question. Two of the shorter paths are illustrated in Figure 3 and Figure 4. The first path is obtained via a single interpolation and involves a conformational change in the lys side chain, which is involved in a salt bridge to glu (Figure 3). The second path consists of a more complicated reorganisation of the trp side-chains, with eight transition states linking nine local minima (Figure 4). A complete path was obtained in three cycles of the missing connection procedure,[89] which required 700 s of cpu time on a laptop computer (using one core of a 2.50 GHz Intel T9300 processor), including tight convergence to an RMS gradient of 10^{-6} kcal/mol and normal mode analysis for all stationary points. The overall path involves rearrangements on rather different length and energy scales, which can cause problems with automated geometry optimisation procedures. For example, the

first step corresponds to rotation of the hydroxyl group in ser, while the second step involves a much higher barrier for restacking trp side chains.

4.2 An Atomic Cluster

The QCI scheme was also implemented for an atomic cluster, to test whether useful interpolations could be obtained for systems without a fixed connectivity defined by covalent bonds. For clusters bound by the Lennard-Jones potential[93] $V_{\rm int}$ was defined as

$$V_{\rm int} = \sum_{i=2}^{M-1} \sum_{\alpha<\beta} V_{\rm LJ}[d^i_{\alpha\beta}] + \sum_{i=1}^{M-1} \sum_{\alpha<\beta}^{*} V_{\rm rep}[d^{i\,i+1}_{\alpha\beta}(\theta^*)], \tag{10}$$

where $V_{\rm LJ}$ is the usual Lennard-Jones form and

$$V_{\rm rep}[d^{i\,i+1}_{\alpha\beta}(\theta^*)] = \frac{\epsilon_{\rm rep}\left(d^i_{\alpha\beta} - d^{i\,i+1}_{\alpha\beta}(\theta^*) - \delta\right)^2 \left(d^{i+1}_{\alpha\beta} - d^{i\,i+1}_{\alpha\beta}(\theta^*) - \delta\right)^2}{\left[d^{i\,i+1}_{\alpha\beta}(\theta^*)\right]^4}, \tag{11}$$

for $d^i_{\alpha\beta} - d^{i\,i+1}_{\alpha\beta}(\theta^*) \geq \delta$, and $d^{i+1}_{\alpha\beta} - d^{i\,i+1}_{\alpha\beta}(\theta^*) \geq \delta$, and zero otherwise. Reduced units for the LJ potential are employed throughout this section. The form of $V_{\rm rep}$ was chosen to provide a smooth function and first derivatives. The parameter δ was included to prevent numerical instabilities that might occur in the limits where the internal minimum is very shallow and close to one of the images. The results reported below employed $M = 11$ initial images, a maximum of $L_{\rm max} = 2000$ LBFGS steps, $G^{\rm max}_{\rm int} = 10^{-3}$, $\epsilon_{\rm rep} = 1$, and $\delta = 10^{-6}$. However, sensible pathways were also obtained with $\delta = 0$. No constraints were applied, and the complete $V_{\rm int}$ including all atoms was used throughout the minimisation. This procedure is therefore simpler than the two-phase scheme employed for the biomolecules considered in this report. Dynamic addition, deletion, and freezing of images was included, as described in §2.2, with $D_{\rm min} = 0.1$, $D_{\rm max} = 2.5$, and $G^{\rm freeze}_{\rm int} = 10^{-10}$. The QCI images were again used to seed DNEB calculations followed by single-ended hybrid eigenvector-following transition state refinement.

The pathway illustrated in Figure 5 corresponds to a change in morphology from the face-centred-cubic global minimum of the 38-atom cluster $\rm LJ_{38}$ bound by the Lennard-Jones potential.[93] This system exhibits a double-funnel potential energy landscape,[94–99] and has served as a benchmark for global optimisation,[100,101] thermodynamics,,[95–98] and rare event dynamics.[15,16,102,103]

The pathway calculation based on QCI required 6 s of CPU time (on one core of a 2.50 GHz Intel T9300 processor), and four cycles of the missing connection procedure.[89] For comparison, when a DNEB interpolation was used via an initial straight line guess the CPU time was practically the same, although a slightly different pathway was produced. The precise sequence of transition states and minima is generally quite sensitive to parameters of the interpolation procedure in all these calculations, and small changes can significantly affect the number of cycles required to obtained a complete connected path.

4.3 Protein L

Protein L exhibits a ubiquitin fold with a central α helix packed against a mixed four-strand β sheet.[104] This system was modelled using the sequence-dependent BLN potential of Brown and Head-Gordon,[8,9] where each of the 56 amino acids is represented by a hydrophobic (B), hydrophilic (L), or neutral (N) bead. A detailed analysis of the energy landscape for this protein as a function of static pulling force is presented elsewhere.[84] The QCI procedure proved to be very effective in providing pathways between distant structures that avoid chain crossings, and has been tested in millions of connection attempts in this work.[84] One example will be illustrated here, for which an initial straight-line interpolation is very inefficient. The two minima correspond to the global minimum for zero force and the global minimum that first appears when a static force is applied to residues 1 and 56, relaxed for zero force.

The pathway in Figure 6 involves 41 transition states and required 12 cycles of the QCI procedure, which took 3045 s CPU time on an Intel Xeon E5404 processor (running at 2.0 GHz). The folding path involves initial formation of the N-terminal hairpin, followed by association of the hairpin with the helix. Strand β_1 then dissociates and strand β_3 associates with the helix and strand β_2. Strands β_1 and β_4 then form a connection and associate with the helix in a concerted fashion. For reference, the path that makes the largest contribution to the overall rate constant for transitions between these two minima after refinement of the kinetic transition network involves 22 transition states and a similar overall barrier. In the folding direction the N-terminal hairpin

forms first, followed by association with the helix and strand β_3. Finally, strand β_4 joins the β sheet. Hence the fastest path corresponds to the same mechanism as the longer initial path, but with some unnecessary motion cut out.

4.4 Benchmarks for the Trpzip and GNNQQNY Peptides

In previous work[92] we tested seven different combinations of internal and Cartesian coordinates for interpolation and alignment of the amyloidogenic GNNQQNY peptide[32,33] (NH_3^+-Gly-Asn-Asn-Gln-Gln-Asn-Tyr-COO^-) and the tryptophan zipper peptide, trpzip1[30,34] (NH_3-Ser-Thr-Trp-Glu-Asn-Gly-Lys-Trp-Thr-Trp-Lys-CH_3). Details of these calculations can be found in the original reference;[92] the abbreviations are summarised in Table 1.

For each peptide we chose the same pairs of minima from existing databases for connection attempts as in previous work. Pairs known to be connected by pathways involving at least one, two and three transition states were selected to give test sets denoted TS1, TS2, and TS3. The test sets for CHARMM involved connections between 150 pairs of local minima, except for GNNQQNY TS1 and TS2, where 300 examples were chosen, giving 1200 in total. For the AMBER benchmarks 150 pairs were considered in each case, i.e. 900 paths in total. The CHARMM calculations employed the united-atom force field CHARMM19[2] with the implicit solvation potential EEF1,[105] while the AMBER calculations used the ff03 parameters[106] and the generalised Born solvation model GB^{OBC}.[107] Both potentials were symmetrised so that permutational isomers would have the same energy.[36,71] The mean cpu time required to achieve a reconnection for each procedure and test set are summarised in Table 2 and displayed graphically in Figure 7. A consistent set of parameters was employed for the `OPTIM` program throughout these tests, the only differences being the keywords associated with the interpolation procedure. The neighbour list for nonbonded interactions in CHARMM was updated every 100 geometry optimisation steps, after noting an apparent discontinuity that occurred in one test case when the check was every 1000 steps.

The timings for these 2,100 connection attempts, referred to hereafter as the MSB test set, are summarised in Table 2 and plotted in Figure 7. These timings refer to the local permutational

alignment procedure, which improved the overall performance in virtually every case. In particular, the QCI approach is generally as good as the methods that use internal coordinates, and even DNEB searches starting from straight line interpolations are competitive when the permutational alignment is optimal. The overhead for the QCI procedure is visible in the simple connections that run very quickly with all the interpolation procedures. For the most difficult connections, namely TS3 with AMBER and trpzip1, the QCI interpolation is the fastest among the methods that give zero failures. The principal advantage of QCI is that it does not require prior identification of internal coordinates, which may be problematic for systems containing separate molecules, for example. To facilitate future comparisons all the necessary input files, along with the `OPTIM` output in each case, will be made available for download from the `OPTIM` web site.[72]

5 Conclusions

The quasi-continuous interpolation (QCI) procedure can identify physically realistic initial paths between fixed end point configurations in a variety of systems. QCI is a chain-of-states method, involving geometry optimisation of all the coupled images of the system to locate a pathway that satisfies various constraint conditions. During the minimisation procedure for the chain of states additional terms are added to the potential energy function, which account for constraint violations between the images. For synchronous motion between images, local minima in the distance between any given pair of atoms can be identified analytically. The auxiliary interpolation potential is evaluated for these pairwise local minima, and corresponding penalty terms are added to the total energy and gradient. Hence we account for atom clashes or chain crossings that occur between images.

The pairwise interpolation potential can be evaluated for any subset of atoms, which enables paths to be constructed by adding one or more atoms at a time. Once a path that satisfies all the bond distance and atom-atom repulsion constraints has been obtained, further refinement is allowed using a combination of the interpolation function and the real potential. Regularly spaced

configurations along the path between images are then used as the starting guess for a doubly-nudged[36] elastic band[18–23] calculation, and local maxima in the DNEB profile are tightly converged to transition states using hybrid eigenvector-following.[20,76,77] This approach has been tested and benchmarked for an atomic cluster, selected pathways of trpzip2, a coarse-grained model of protein L, and atomistic representations of two peptides using both AMBER[5–7] and CHARMM[1–4] force fields. For the most difficult peptide connections the QCI approach was somewhat faster than schemes that use internal coordinates. However, the main benefit is realised in the construction of kinetic transition networks for folding and unfolding of protein L and protein G, where chain crossings were frequently obtained using straight line initial interpolations.[84]

Efficient construction of pathways using geometry optimisation for atomistic representations of peptides and proteins depends upon proper treatment of permutational isomers. In previous work we have shown how to automate the symmetrisation of CHARMM and AMBER force fields so that consistent energies are obtained for equivalent configurations.[36,71] It is also necessary to align the end points in double-ended calculations, to avoid unnecessary and potentially unphysical rearrangements. The shortest augmenting path procedure employed in previous work[65,80,81,92] is improved in the present contribution using local distance metrics by defining suitable neighbourhoods. The local permutational alignment solves the initial interpolation problem for cases where the overall minimum distance for all atoms produces additional permutational rearrangements on the pathway. For difficult cases involving distant end points local permutational alignment can produce successful connections where previous procedures have failed or required excessive computer time.

Acknowledgements

We are grateful to Dr Thomas Hofer and Dr Christopher Whittleston for supplying the trpzip2 endpoints, and Dr Tomas Oppelstrup for his interface to the shortest augmenting path procedure.

References

[1] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, J. Comp. Chem **4**, 187 (1983).

[2] E. Neria, S. Fischer and M. Karplus, J. Chem. Phys. **105**, 1902 (1996).

[3] A. D. MacKerell Jr., D. Bashford, M. Bellott, R. L. Dunbrack Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus, J. Phys. Chem. B **102**, 3586 (1998).

[4] N. Foloppe and A. D. MacKerell, Jr., J. Comput. Chem. **21**, 86 (2000).

[5] D. Case, T. Darden, T. Cheatham, C. Simmerling, J. Wang, R. Duke, R. Luo, K. Merz, D. Pearlman, M. Crowley, R. Walker, W. Zhang, B. Wang, S. Hayik, A. Roitberg, G. Seabra, K. Wong, F. Paesani, X. Wu, S. Brozell, V. Tsui, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, P. Beroza, D. Mathews, C. Schafmeister, W. Ross and P. Kollman, AMBER 9, University of California (2006).

[6] D. Pearlman, D. Case, J. Caldwell, W. Ross, T. Cheatham, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman., Comp. Phys. Commun. **91**, 1 (1995).

[7] D. Case, T. Cheatham, T. Darden, H. Gohlke, R. Luo, K. Merz, A. Onufriev, C. Simmerling, B. Wang and R. Woods, J. Computat. Chem. **26**, 1668 (2005).

[8] S. Brown, N. J. Fawzi and T. Head-Gordon, Proc. Natl. Acad. Sci. USA **100**, 10712 (2003).

[9] S. Brown and T. Head-Gordon, Protein Sci. **13**, 958 (2004).

[10] F. Noé, D. Krachtus, J. C. Smith and S. Fischer, J. Chem. Theory Comput. **2**, 840 (2006).

[11] D. J. Wales, Int. Rev. Phys. Chem. **25**, 237 (2006).

[12] F. Noé and S. Fischer, Curr. Op. Struct. Biol. **18**, 154 (2008).

[13] D. Prada-Gracia, J. Gómez-Gardenes, P. Echenique and F. Fernando, PLoS Comput. Biol. **5**, 1 (2009).

[14] D. J. Wales, Curr. Op. Struct. Biol. **20**, 3 (2010).

[15] D. J. Wales, Mol. Phys. **100**, 3285 (2002).

[16] D. J. Wales, Mol. Phys. **102**, 891 (2004).

[17] M. C. Prentiss, D. J. Wales and P. G. Wolynes, PLoS Comput. Biol. **6**, e1000835 (2010).

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

[19] R. Czerminski and R. Elber, J. Chem. Phys. **92**, 5580 (1990).

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

[21] G. Henkelman, B. P. Uberuaga and H. Jónsson, J. Chem. Phys. **113**, 9901 (2000).

[22] G. Henkelman and H. Jónsson, J. Chem. Phys. **113**, 9978 (2000).

[23] G. Henkelman and H. Jónsson, J. Chem. Phys. **115**, 9657 (2001).

[24] C. N. Nguyen and R. M. Stratt, J. Chem. Phys. **133**, 124503 (2010).

[25] P. Faccioli, J. Phys. Chem. B **112**, 13756 (2008).

[26] P. Faccioli, A. Lonardi and H. Orland, J. Chem. Phys. **133**, 045104 (2010).

[27] H. B. Schlegel, in *Encyclopedia of Computational Chemistry*, edited by P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer and P. R. Schreiner, vol. 2, p. 1136, John Wiley and Sons, New York (1998).

[28] O. Farkas and H. B. Schlegel, J. Mol. Struct. **666-667**, 31 (2003).

[29] M. M. Tirion, Phys. Rev. Lett. **77**, 1905 (1996).

[30] A. G. Cochran, N. J. Skelton and M. A. Starovasnik, Proc. Nat. Acad. Sci. USA **98**, 5578 (2001).

[31] C. D. Snow, L. Qiu, F. Gai, S. J. Hagen and V. S. Pande, Proc. Natl. Acad. Sci. USA **101**, 4077 (2004).

[32] M. Balbirnie, R. Grothe and D. S. Eisenberg, Proc. Natl. Acad. Sci. USA **98**, 2375 (2001).

[33] R. Nelson, M. R. Sawaya, M. Balbirnie, A. O. Madsen, C. Riekel, R. Grothe and D. Eisenberg, Nature **435**, 773 (2005).

[34] C. D. Snow, L. Qiu, D. Du, F. Gai, S. J. Hagen and V. S. Pande, Proc. Nat. Acad. Sci. USA **101**, 4077 (2004).

[35] C. Vega and S. Lago, Computers and Chemistry **18**, 55 (1994).

[36] S. A. Trygubenko and D. J. Wales, J. Chem. Phys. **120**, 2082 (2004).

[37] F. Tama and Y. H. Sanejouand, Prot. Eng. **14**, 1 (2001).

[38] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demeril, O. Keskin and I. Bahar, Biophys. J. **80**, 505 (2001).

[39] M. K. Kim, G. S. Chirikjan and R. L. Jernigan, J. Mol. Graph. Model **21**, 151 (2002).

[40] D. Ming, Y. Kong, M. A. Lambert, Z. Huang and J. Ma, Proc. Natl. Acad. Sci. USA **99**, 8620 (2002).

[41] M. Delarue and Y. H. Sanejouand, J. Mol. Biol. **320**, 1011 (2002).

[42] M. Ikeguchi, J. Ueno, M. Sato and A. Kidera, Phys. Rev. Lett. **94**, 078102 (2002).

[43] W. Zheng and S. Doniach, Proc. Natl. Acad. Sci. USA **100**, 13253 (2003).

[44] N. Reuter, K. Hinsen and J. J. Lacapère, Biophys. J. **85**, 2186 (2003).

[45] C. Xu, D. Tobi and I. Bahar, J. Mol. Biol. **333**, 153 (2003).

[46] F. Tama, M. Valle, J. Franks and C. L. Brooks, Proc. Natl. Acad. Sci. USA **100**, 9319 (2003).

[47] P. Maragakis and M. Karplus, J. Mol. Biol. **352**, 807 (2005).

[48] I. Bahar and A. J. Rader, Curr. Op. Struct. Biol. **15**, 586 (2005).

[49] P. C. Whitford, O. Miyashita, Y. Levy and J. N. Onuchic, J. Mol. Biology **366**, 1661 (2007).

[50] W. Zheng, B. R. Brooks and G. Hummer, Proteins **69**, 43 (2007).

[51] A. Korkut and W. A. Hendrickson, Proc. Nat. Acad. Sci. USA **106**, 15667 (2009).

[52] M. Lu and J. Ma, in *Energy Flow in Proteins*, edited by D. Leitner and J. Straub, pp. 229–245, CRC Press, Boca Raton (2009).

[53] C. Peng, L. Zhang and T. Head-Gordon, Biophysical J. **98**, 2356 (2010).

[54] P. Batista, C. Robert, J. D. Marechal, M. Hamida-Rebai, P. Pascutti, P. Bisch and D. Perahia, Phys. Chem. Chem. Phys. **12**, 2850 (2010).

[55] R. B. Best, Y.-G. Chen and G. Hummer, Structure **13**, 1755 (2005).

[56] K.-i. Okazaki, N. Koga, S. Takada, J. N. Onuchic and P. G. Wolynes, Proc. Nat. Acad. Sci. USA **103**, 11844 (2006).

[57] K.-i. Okazaki and S. Takada, Proc. Nat. Acad. Sci. USA **105**, 11182 (2008).

[58] Q. Lu and J. Wang, J. Am. Chem. Soc. **130**, 47724783 (2008).

[59] Z.-Z. Lai, Q. Lu and J. Wang, J. Phys. Chem. B **115**, 4147 (2011).

[60] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, Massachusetts (2003).

[61] H. Jónsson, G. Mills and K. W. Jacobsen, in *Classical and quantum dynamics in condensed phase simulations*, edited by B. J. Berne, G. Ciccotti and D. F. Coker, Singapore (1998), World Scientific.

[62] S. Wells, S. Menor, B. Hespenheide and M. F. Thorpe, Phys. Biol. **2**, S127 (2005).

[63] D. W. Farrell, K. Speranskiy and M. F. Thorpe, Proteins: Struct. Func. Bioinfo. **78**, 2908 (2010).

[64] D. W. Farrell, M. Lei and M. F. Thorpe, Phys. Biol. **8**, 026017 (2011).

[65] M. Khalili and D. J. Wales, J. Phys. Chem. B **112**, 2456 (2008).

[66] J. Schlitter, M. Engels, P. Krüger, E. Jacoby and A. Wollmer, Mol. Simul. **10**, 291308 (1993).

[67] J. Schlitter, M. Engels and P. Krüger, J. Mol. Graph. **12**, 84 (1994).

[68] J. Ma and M. Karplus, Proc. Natl Acad. Sci. USA **94**, 11905 (1997).

[69] W. E, W. Ren and E. Vanden-Eijnden, Phys. Rev. B **66**, 052301 (2002).

[70] B. Peters, A. T. Bell and A. Chakraborty, J. Chem. Phys. **121**, 4453 (2004).

[71] E. Małolepsza, B. Strodel, M. Khalili, S. Trygubenko, S. Fejer and D. J. Wales, J. Comp. Chem. **31**, 1402 (2010).

[72] D. J. Wales, *Optim: A program for optimising geometries and calculating pathways* ().

[73] J. Nocedal, Mathematics of Computation **35**, 773 (1980).

[74] D. Liu and J. Nocedal, Math. Prog. **45**, 503 (1989).

[75] D. Sheppard, R. Terrell and G. Henkelman, J. Chem. Phys. **128**, 134106 (2008).

[76] L. J. Munro and D. J. Wales, Phys. Rev. B **59**, 3969 (1999).

[77] Y. Kumeda, L. J. Munro and D. J. Wales, Chem. Phys. Lett. **341**, 185 (2001).

[78] S. K. Kearsley, Acta Cryst. A **45**, 208 (1989).

[79] R. Jonker and A. Volgenant, Computing **38**, 325 (1987).

[80] J. M. Carr and D. J. Wales, J. Phys. Chem. B **112**, 8760 (2008).

[81] J. M. Carr and D. J. Wales, Phys. Chem. Chem. Phys. **11**, 3341 (2009).

[82] T. James, D. J. Wales and J. H. Rojas, J. Chem. Phys. **126**, 054506 (2007).

[83] D. Chakrabarti and D. J. Wales, Soft Matter **7**, 2325 (2011).

[84] D. J. Wales and T. Head-Gordon, J. Phys. Chem. B **000**, 0000 (2012).

[85] R. Rafac, J. P. Schiffer, J. S. Hangst, D. H. E. Dubin and D. J. Wales, Proc. Natl. Acad. Sci. USA **88**, 483 (1991).

[86] D. J. Wales and A. M. Lee, Phys. Rev. A **47**, 380 (1993).

[87] E. Yurtsever, F. Calvo and D. J. Wales, Phys. Rev. E **72**, 026110 (2005).

[88] F. Calvo, E. Yurtsever and D. J. Wales, J. Chem. Phys. **136**, 024303 (2012).

[89] J. M. Carr, S. A. Trygubenko and D. J. Wales, J. Chem. Phys. **122**, 234903 (2005).

[90] E. W. Dijkstra, Numerische Math. **1**, 269 (1959).

[91] T. Lazaridis and M. Karplus, Science **278**, 1928 (1997).

[92] M. S. Bauer, B. Strodel, S. N. Fejer, E. F. Koslover and D. J. Wales, J. Chem. Phys. **132**, 054101 (2010).

[93] J. E. Jones and A. E. Ingham, Proc. R. Soc. A **107**, 636 (1925).

[94] D. J. Wales, M. A. Miller and T. R. Walsh, Nature **394**, 758 (1998).

[95] J. P. K. Doye, M. A. Miller and D. J. Wales, J. Chem. Phys. **110**, 6896 (1999).

[96] J. P. Neirotti, F. Calvo, D. L. Freeman and J. D. Doll, J. Chem. Phys. **112**, 10340 (2000).

[97] F. Calvo, J. P. Neirotti, D. L. Freeman and J. D. Doll, J. Chem. Phys. **112**, 10350 (2000).

[98] D. D. Frantz, J. Chem. Phys. **115**, 6136 (2001).

[99] D. J. Wales, *Energy Landscapes*, Cambridge University Press, Cambridge (2003).

[100] D. J. Wales and J. P. K. Doye, J. Phys. Chem. A **101**, 5111 (1997).

[101] D. J. Wales and H. A. Scheraga, Science **285**, 1368 (1999).

[102] M. A. Miller, J. P. K. Doye and D. J. Wales, J. Chem. Phys. **110**, 328 (1999).

[103] M. Picciani, M. Athenes, J. Kurchan and J. Tailleur, J. Chem. Phys. **135**, 034108 (2011).

[104] M. Wikstrom, T. Drakenberg, S. Forsen, U. Sjobring and L. Bjork, Biochem. **33**, 14011 (1994).

[105] T. Lazaridis and M. Karplus, Proteins: Struct., Func. and Gen. **35**, 133 (1999).

[106] Y. Duan, C. Wu, S. Chowdhury, M. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo and T. Lee, J. Comput. Chem. **24**, 1999 (2003).

[107] A. Onufriev, D. Bashford and D. A. Case, Proteins **55**, 383 (2004).

[108] W. Humphrey, A. Dalke and K. Schulten, J. Molec. Graphics **14**, 33 (1996).

TABLE 1: Summary of the different alignment and interpolation schemes used in previous work.[92]

method	alignment	interpolation
BCSC	Cartesian	Cartesian for backbone and sidechains
BCSI	Cartesian	Cartesian for backbone, CHARMM internal for sidechains
BISI	Cartesian	CHARMM internal for backbone and sidechains
NONI	Cartesian	natural internal
NI	natural internal	natural internal
NIS	natural internal	natural internal, using non-equispaced images
NIC	natural internal	natural internal or Cartesian, according to the lowest energy

TABLE 2: Timings for the MSB test set introduced in previous work.[92] QCI and DNEB refer to interpolation using the QCI approach described in the text and to images placed along an initial straight line path for the doubly-nudged[36] elastic band[18–23] procedure, respectively. The other abbreviations are defined in Table 1. The average cpu time (s) for an Intel Xeon E5404 processor (running at 2.0 GHz) is given in each case; a number in brackets indicates that there were one or more failures for this combination when a maximum of 50 cycles of the missing connection algorithm[89] was allowed. The number of tests in each category is 150, except for sets TS1 and TS2 for CHARMM, where 300 paths were considered. All the input files required to run these `OPTIM` benchmarks will be made available from the `OPTIM` web site,[72] together with the corresponding output.

	GNNQQNY			trpzip1		
interpolation	TS1	TS2	TS3	TS1	TS2	TS3
	CHARMM					
QCI	6.3	14.0	20.4	29.6	57.8	91.1
DNEB	3.0	8.6	14.7	30.0	62.9	115.0
BCSC	3.1	8.8	14.7	30.7	62.0	112.4
BCSI	3.1	9.4	13.1	30.6	50.8	90.8
BISI	3.7	12.0	19.7	47.9	96.9(1)	116.8
NONI	3.4	9.1	13.6	30.2	78.0	92.9(1)
NI	3.7	9.1	13.5	30.7	78.1	90.9(1)
NIS	3.7	8.8	14.4	30.9	74.9(1)	94.9(1)
NIC	3.7	8.5	13.6	29.9	55.9	106.1(1)
	AMBER					
QCI	37.5	79.7	92.2	214.2	412.4	562.8
DNEB	31.5	77.2	132.0(1)	251.5	516.2	576.7(3)
BCSC	31.6	76.1	131.4(1)	243.8	511.7	590.2(3)
BCSI	31.5	77.4	131.2(1)	245.2	518.2	590.5(3)
BISI	31.5	76.2	131.5(1)	244.8	524.2	584.6(3)
NONI	31.9	82.7	79.6(1)	177.4(2)	439.2	703.7
NI	32.1	82.4	79.9(1)	175.4(2)	445.7	690.7
NIS	31.8	72.2(1)	80.0	197.4	420.5	730.5
NIC	30.4	78.4	75.7	205.2(1)	360.9	614.2

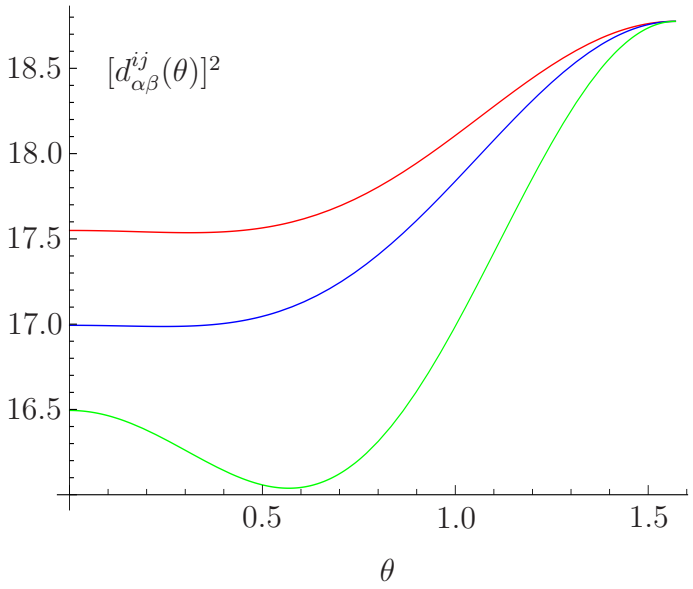

FIGURE 1: Squared distance (arbitrary units) between two atoms $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ as a function of θ (radians). As the z coordinate of $\mathbf{r}^j_\beta$, atom β in configuration j, varies, an internal minimum develops.

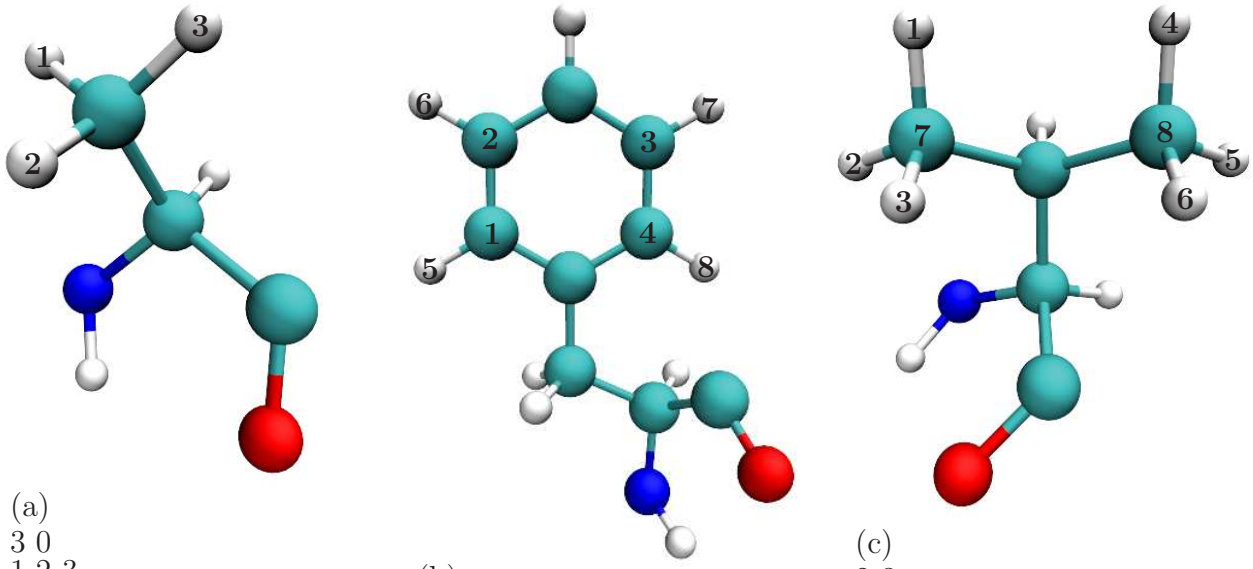

FIGURE 2: Allowed permutations and corresponding entries in the auxiliary file, illustrated for three amino acid sidechains (graphics generated with VMD[108]). (a) Alanine: the three hydrogens are permutable, and comprise a single set. No other atoms are required to change places at the same time. (b) Phenylalanine: the aromatic ring can rotate, exchanging exactly four pairs of atoms. If atoms 1 and 4 comprise the primary set, then three other pairs 2–3, 5–8 and 6–7 define the secondary sets of atoms that must also be swapped. The definition of particular sets of permutable atoms as primary is arbitrary. (c) Valine: the two methyl groups can be exchanged via four simultaneous pair swaps (here atoms 7 and 8 comprise the primary set). Additionally the three hydrogens within each methyl group can be permuted as in alanine.

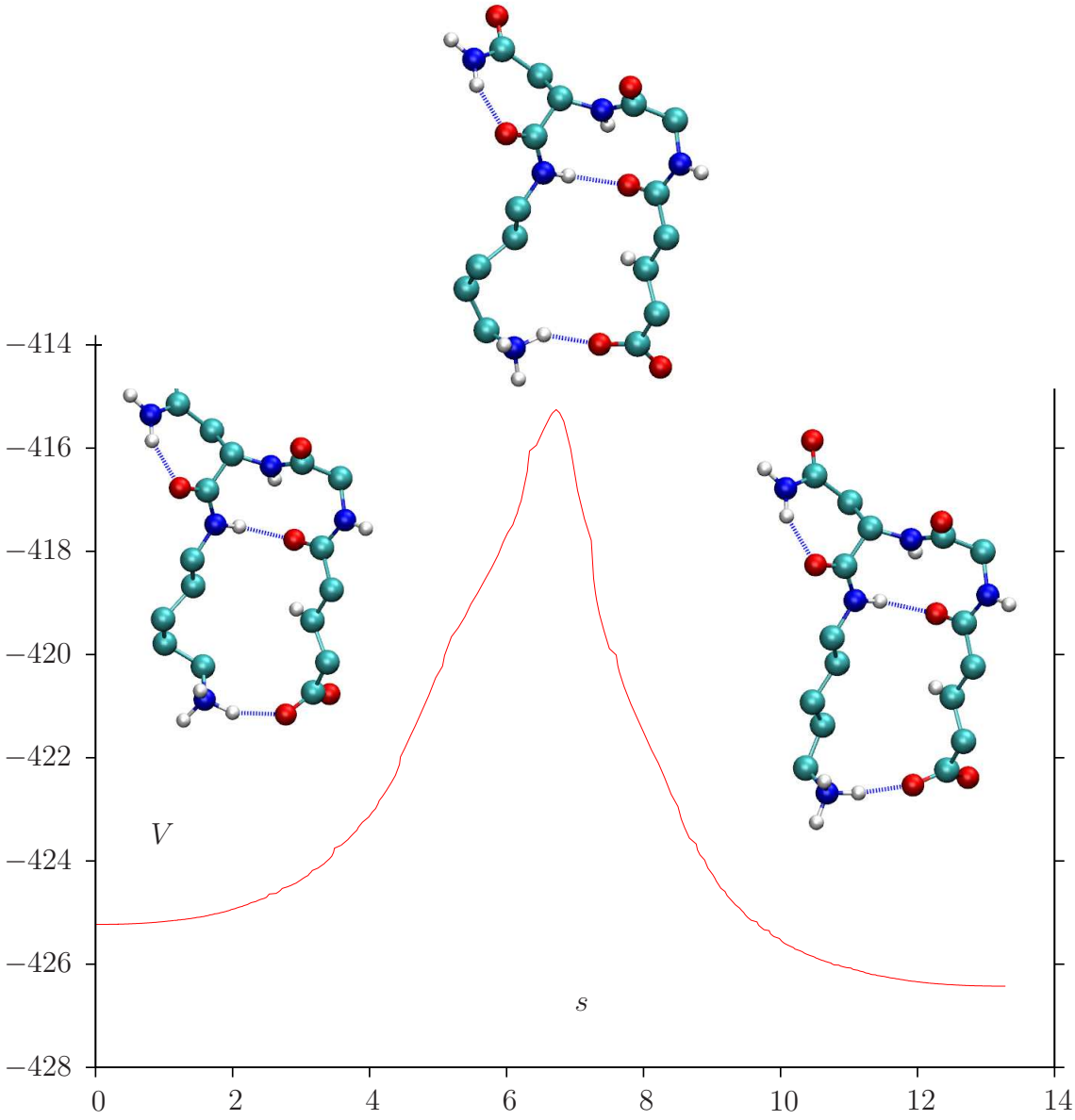

FIGURE 3: Potential energy, V (kcal/mol), plotted as a function of the integrated path length, s (Å), for a single transition state pathway of the trpzip2 peptide. The two local minima and the transition state are superimposed above the path. This pathway corresponds to an internal rotation of the lys sidechain constrained by the hydrogen-bonding contact in the glu/lys salt bridge. The snapshots (graphics generated with VMD[108]) include the C_α carbon atoms of glu and lys, along with the sidechains and the intervening gly and asn residues.

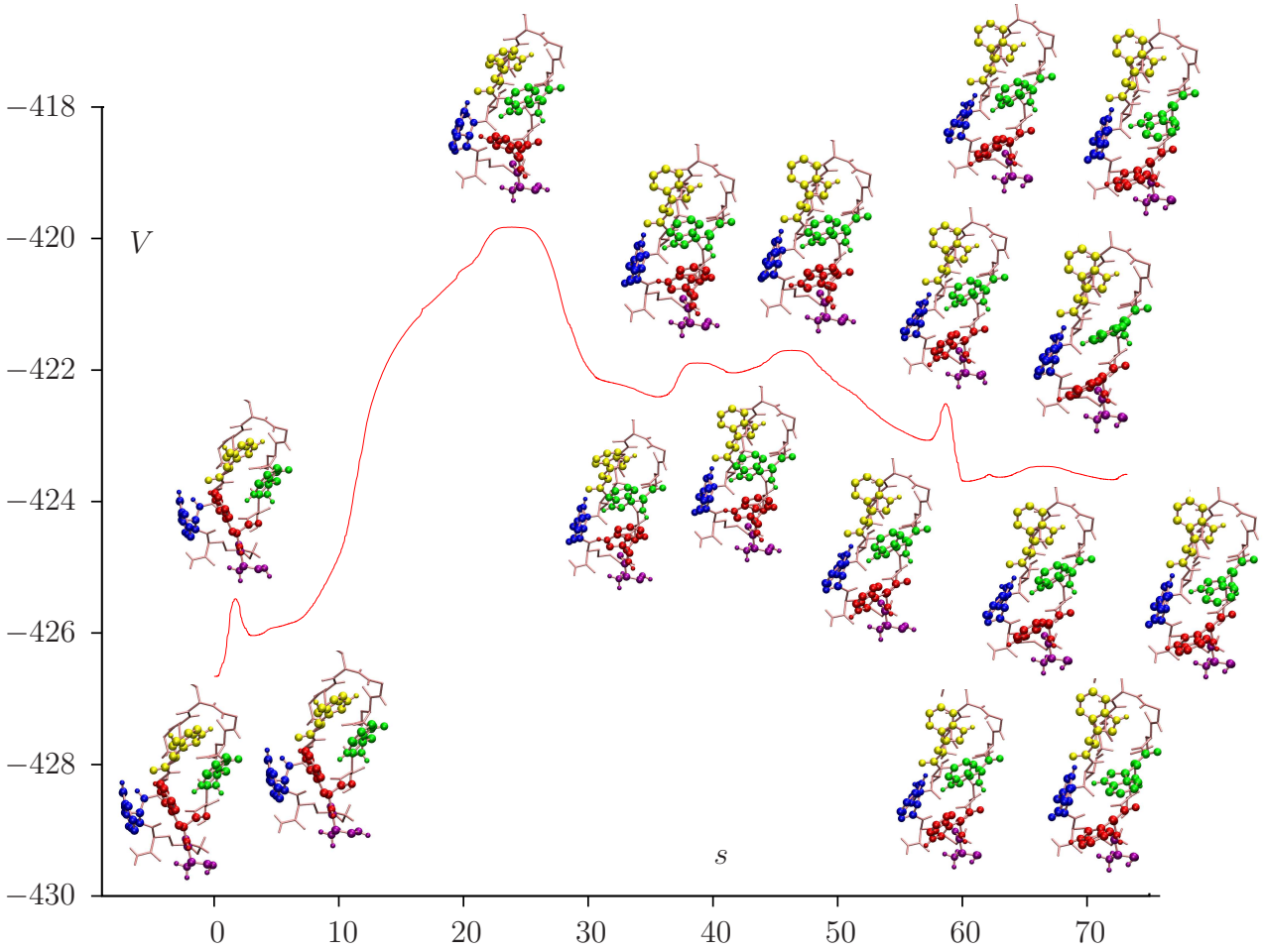

FIGURE 4: Potential energy, V (kcal/mol), plotted as a function of the integrated path length, s (Å), for a path involving trp side chain reorganisations in the trpzip2 peptide. The relevant atoms of the nine local minima and the eight transition states are illustrated below and above the path, respectively (graphics generated with VMD[108]), at approximately the corresponding path length. The trp residues are all coloured differently to distinguish them, and the ser residue is also highlighted, since the first step involves a conformational change in the corresponding OH group.

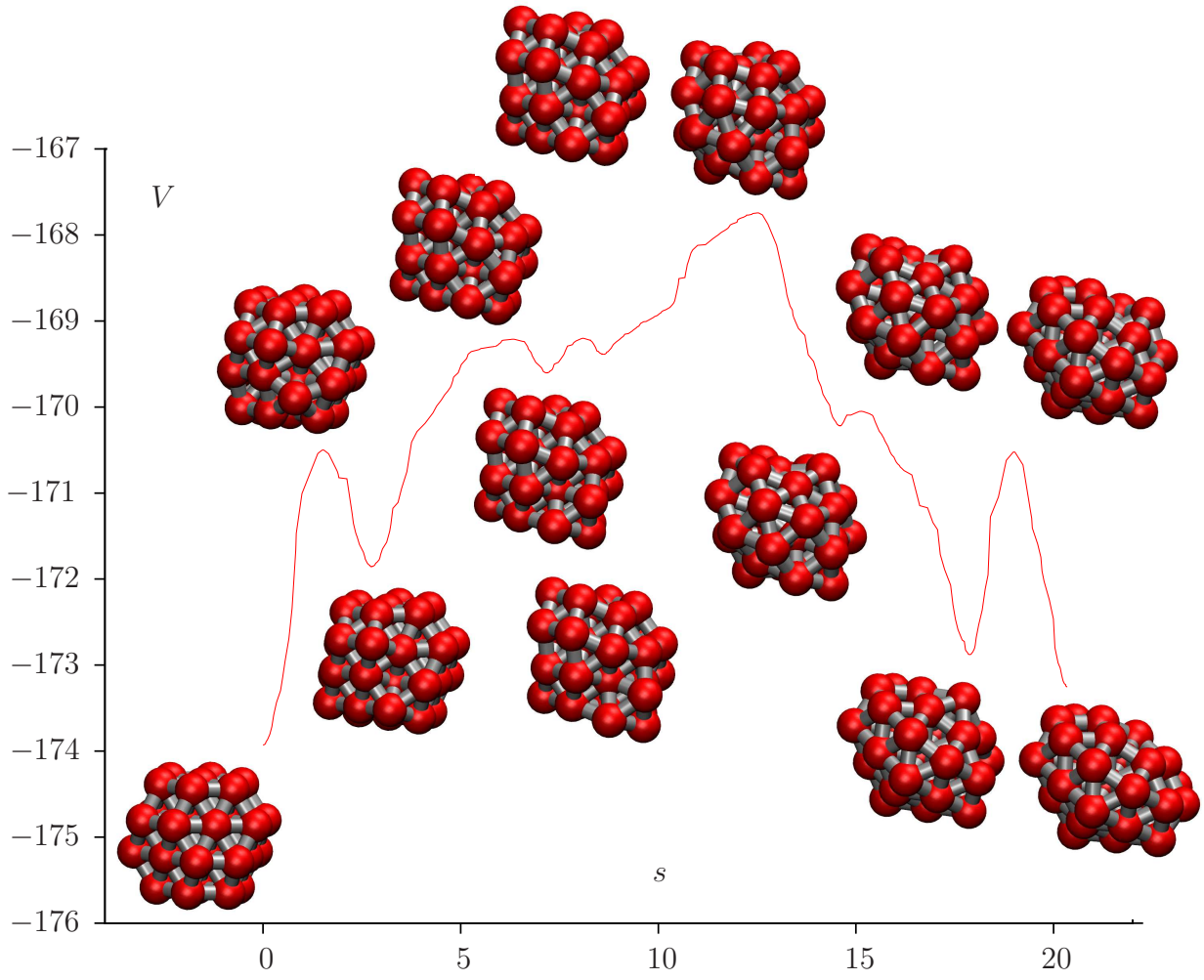

FIGURE 5: Potential energy, V (ϵ), plotted as a function of the integrated path length, s (σ), for a path connecting the global minimum and second-lowest minimum of the LJ_{38} cluster. This path corresponds to an overall change in morphology from face-centred-cubic to icosahedral packing, and involves six transition states and seven minima. These stationary points are illustrated above and below the path, respectively (graphics generated with VMD[108]), at approximately the corresponding path length.

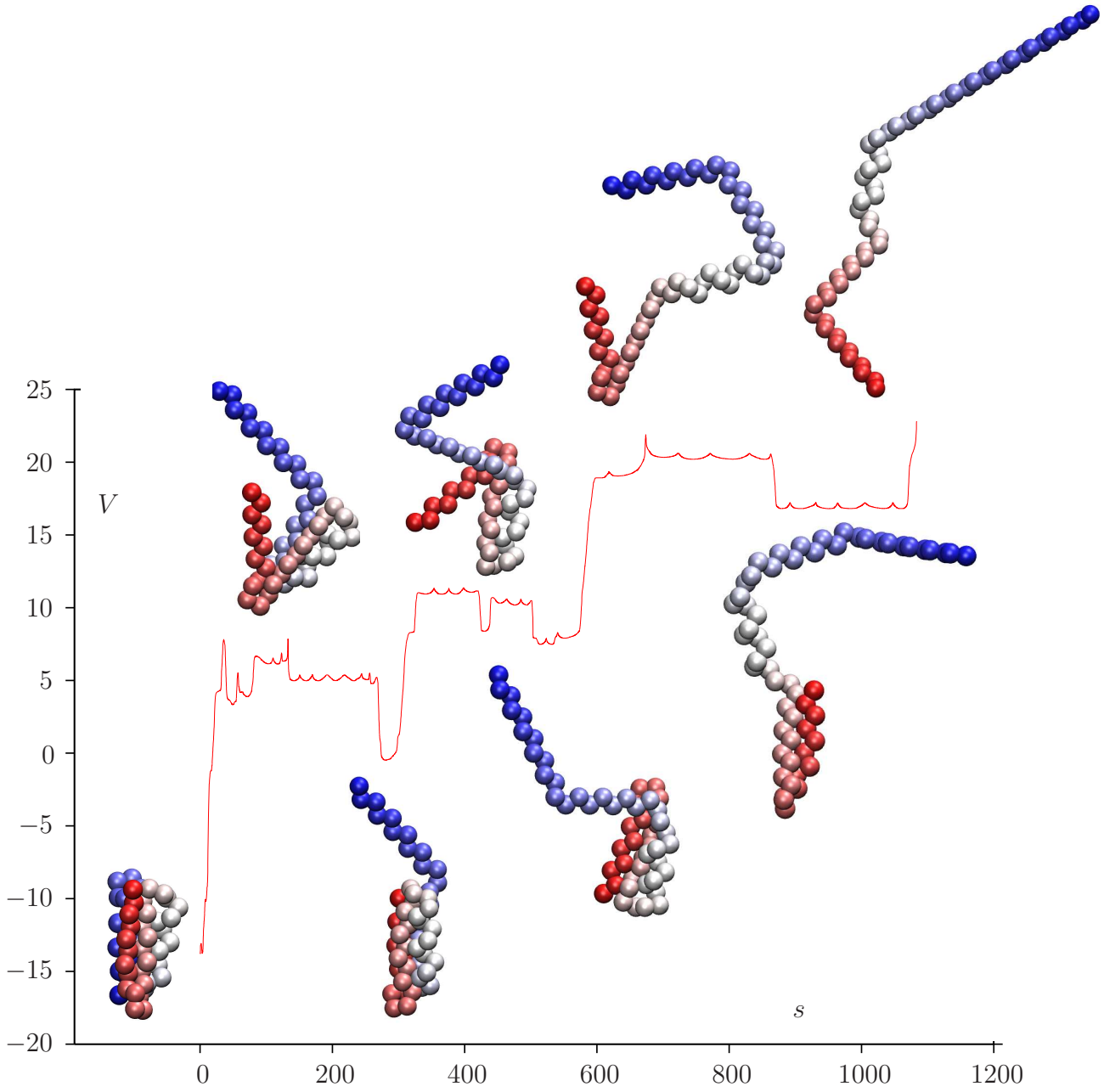

FIGURE 6: Potential energy, V (ϵ), plotted as a function of the integrated path length, s (σ), for a path connecting the global minimum and an extended minimum for protein L represented by a coarse-grained potential. Selected local minima are illustrated using the `VMD` program[108] to generate representations coloured from red to blue (N-terminus, β_1, to C-terminus, β_4) according to the position in the chain. These structures are positioned at approximately the corresponding path length.

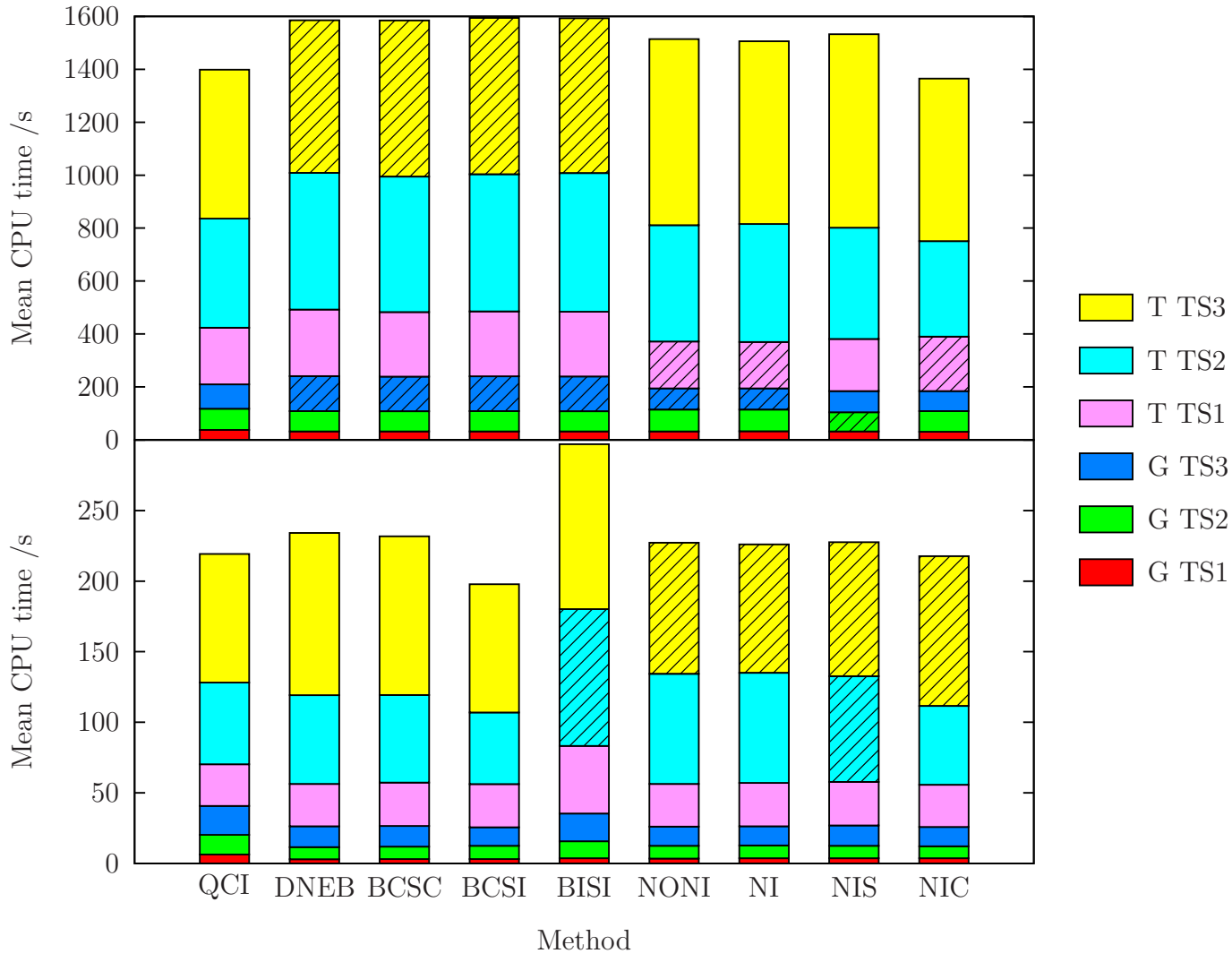

FIGURE 7: Stacked bar charts displaying the results from Table 2. The average cpu time (s) using an Intel Xeon E5404 processor (running at 2.0 GHz) is plotted for the pathways in the MSB test set[92] and each interpolation method. The results with AMBER are shown in the upper panel, and those with CHARMM are in the lower panel. In the legend, G and T refer to GNNQQNY and trpzip1, respectively. Stripes within a box indicate that there were one or more failures for this combination within 50 cycles of the missing connection algorithm,[89] and therefore that the box height shown is a lower bound.