Folding of the GB1 hairpin peptide from discrete path sampling

David A. Evans and David J. Wales
University Chemical Laboratories, Lensfield Road, Cambridge CB2 1EW, United Kingdom

(Received 19 December 2003; accepted 14 April 2004)

The discrete path sampling technique is used to calculate folding pathways of the 16-amino acid β hairpin-forming sequence from residues 41–56 of the B1 domain of protein G. The folding time is obtained using master equation dynamics and kinetic Monte Carlo simulations, and the time evolution of different order parameters and occupation probabilities of groups of minima are calculated and used to characterize intermediates on the folding pathway. © 2004 American Institute of Physics. [DOI: 10.1063/1.1759317]

I. INTRODUCTION

Peptides that form elements of secondary structure when isolated in solution are important model systems for studying various aspects of protein folding. The simplest β system commonly considered is the C-terminal fragment (residues 41–56) of the B1 domain of protein G, which forms a β hairpin in the intact protein1 and also when isolated in solution.2 The folding of this “GB1” peptide was monitored experimentally using the intensity of tryptophan fluorescence as a probe, revealing that it followed two-state kinetics with a folding rate constant of 1/(6 μs) at 297 K.3 It was postulated initially by Muñoz et al. that this folding rate, which is 2–3 orders of magnitude slower than that for α helices, could be explained by a simple model where each residue alternates between a random coil and a hairpin configuration.3,4

This model predicts that formation of the turn region is the first step in hairpin folding. The GB1 hairpin has since been the subject of many simulation studies. Dinner et al. calculated a free energy surface using the EEF1 implicit solvation model and multicanonical MC.5 They found an L-shaped surface as a function of the radius of gyration and the root-mean-square deviation (RMSD) from the protein data bank (PDB) coordinates.1 This result was interpreted to mean that the peptide folded by a collapse downhill in free energy to a compact state, then rearranged to produce the hydrophobic core, followed by the formation of native hydrogen bonds. Similar conclusions on the order of folding events were reached in explicit solvent, high temperature unfolding simulations.6 This picture contrasts with the model of Muñoz et al., in that the rate-limiting step is the formation of the hydrophobic core as opposed to the hydrogen bonds in the turn region. Two replica-exchange MD studies with explicit solvent also support the former view to some extent.7,8 One study identified unfolded, molten globule, and folded states, and calculated a folding time of the same order of magnitude as experiment, based on the size of the free energy barrier.7 The other used a different potential but also reported the same basic free energy surface—the authors postulated a folding mechanism in which hydrogen bonds and the hydrophobic core form concertedly, as opposed to the core-first description mentioned above.5,6

A concerted mechanism was also found by the one study to simulate folding directly, using coupled parallel replica dynamics (PRD) in implicit solvent.9 This work produced a free energy landscape of the same shape as for previous studies, along with a folding rate constant in agreement with experiment. A recent explicit solvent study using transition path sampling (TPS) techniques10–13 calculated the rate constants for transitions along the unfolding pathway via a “frayed” and a hydrophobic intermediate.14 The transition interface sampling method15 was used in this study. The hydrophobic—unfolded transition rate was found to be about three orders of magnitude faster than the native—hydrophobic unfolding rate, in line with the experimental observation of single-exponential kinetics. The native—hydrophobic unfolding rate was found to be the same order of magnitude as the experimental value.

Another study of the hairpin system used the activation-relaxation technique (ART)16 and found three distinct pathways for hairpin formation: one corresponding to the turn-first model of Muñoz et al., one to the initial hydrophobic core formation mechanism seen in the other simulations discussed, and a third, previously unobserved, “reptation” mechanism, in which an asymmetric hairpin forms with the turn in a non-native position before the chain slides into the correct alignment.17

The above studies also consider thermodynamic properties of the peptide. In each case a hairpin was found to be the dominant conformation in solution at room temperature, except for a replica-exchange MD study with implicit solvent, where the collapsed, unfolded state was found to be lower in free energy.18 All the simulations involve a fluctuating number of hydrogen bonds at equilibrium—the structure is not rigid and only around half of the PDB structure β sheet hydrogen bonds are formed at any one time,8 in agreement with NMR data19 and constant temperature MD studies.20,21 However, discrepancies between the simulations and experimental results are apparent in the melting behavior. Experimentally, the midpoint of the folding transition is at 297 K and the hairpin population decays almost to zero once the temperature is raised to 360 K; the simulations have similar hairpin populations around 300 K, but the populations decay...
much more slowly with increasing temperature.\textsuperscript{8} This observation suggests that the commonly used potentials may overbind the secondary structure.

In the present contribution, we describe the application of the discrete path sampling (DPS) methodology\textsuperscript{22,23} to study the folding of the GB1 hairpin peptide. This approach may be less computationally expensive than the TPS and PRD methods, because of the simplified, discrete description of the kinetics. We also expect the DPS method to give more physically relevant folding trajectories than those produced by the ART,\textsuperscript{17} since the algorithm used in DPS only considers two minima to be connected if they are linked by approximate steepest-descent paths leading down from a transition state. In contrast, for the ART study any transition state found after perturbing a minimum was considered to be linked to that initial minimum,\textsuperscript{16} regardless of how far in configuration space the transition state searching procedure may have moved from the initial guess.

II. METHODS

The GB1 peptide was represented by the CHARMM19\textsuperscript{54} force field and the EEF1 implicit solvation potential.\textsuperscript{74} The same potential was used for the multicanonical MC study referred to above,\textsuperscript{5} which established that it provides a satisfactory description of the system, with a global free energy minimum corresponding to the hairpin structure at room temperature. A modified version of the standard CHARMM19 potential was used for the present calculations, in which permutational isomers are guaranteed to have the same potential energy (see the Appendix). The aim of the present work was to find a representative ensemble of pathways from an extended minimum to the native hairpin structure, using DPS to construct a database of connected stationary points of the potential energy.

A. Calculating thermodynamic and kinetic properties

The DPS approach provides a link between stationary points of the potential energy surface (PES) and global thermodynamic and dynamic properties of the system.\textsuperscript{25,26} Stationary points of the PES are configurations where the gradient of the potential energy vanishes. A local minimum is a stationary point where all the nonzero normal mode frequencies are positive, while a transition state has a single imaginary frequency, which corresponds to the reaction coordinate.\textsuperscript{27} To make contact with experiment we must define a local density of states for each of these minima and transition states, which enables us to calculate free energies and rate constants using statistical rate theory. This approach has been used in many previous studies\textsuperscript{25,26} and a brief overview is provided below. A fuller account was provided in our previous report on the application of DPS to metenkephalin.\textsuperscript{23} It is particularly important to distinguish between local minima and transition states of the PES and local minima and transition states of the free energy. States defined in terms of free energy generally correspond to an ensemble of stationary points of the PES, and may depend upon the choice of order parameter.

Locating stationary points of the PES involves geometry optimization. As in previous work, local minima were refined using a modified version of the (LBFGS) procedure described by Liu and Nocedal.\textsuperscript{28} Finding transition states is more difficult, and the present work employed a “double-ended” method to connect pairs of local minima using the nudged elastic band (NEB) method\textsuperscript{29-31} to produce initial transition state guesses. These candidate structures were then converged accurately using hybrid eigenvector-following.\textsuperscript{22,25,32,33} For each transition state two approximate steepest-descent paths were then obtained using LBFGS energy minimization following small displacements parallel and antiparallel to the eigenvector corresponding to the unique negative eigenvalue.

Having constructed a connected database of minima and transition states on the PES, it is possible to calculate finite temperature properties of the system,\textsuperscript{25,26} and we summarize the most important results here. As in previous work, we employ the simplest harmonic approximation to the local density of states for each stationary point of the PES. The equilibrium occupation probability of a minimum \(i\) is then given by

\[
p_i(\beta) = \frac{Z_i(\beta)}{\sum_i Z_i(\beta)},
\]

where \(Z_i(\beta)\) is the partition function of minimum \(i\) and \(Z(\beta)\) is the total partition function. For a group of minima \(I\)

\[
P_{eq}^I(\beta) = \sum_{i \in I} p_i(\beta),
\]

and the corresponding free energies are calculated as

\[
F_i = -k_BT \ln p_i(\beta) \quad \text{and} \quad F_I = -k_BT \ln P_{eq}^I(\beta),
\]

respectively.

The algorithms employed to collect local minima into groups are discussed in Sec. II C.

To obtain dynamical information we use transition state theory, again within the harmonic approximation, to calculate rate constants corresponding to each transition state of the PES

\[
k_{ij}^T(\beta) = \frac{k_BT}{\hbar} \frac{Z_j}{Z_i},
\]

where the transition state partition function \(Z_j\) does not include the reaction mode with imaginary frequency, and the partition functions are calculated relative to a common energy zero. The free energy of the thermodynamic transition state between groups of minima \(I\) and \(J\) was calculated as

\[
F_{IJ} = -k_BT \ln \left( \sum_{(i,j) \in IJ} \left( \frac{p_i(\beta)p_j(\beta)}{p_i(\beta)p_j(\beta)} \right)^{1/2} Z_j \right),
\]

where the sum is over all transition states that connect a minimum in group \(I\) with a minimum in group \(J\).

The DPS approach introduced in Ref. 22 defines a “discrete path” in terms of a sequence of local minima from the PES and the intervening transition states. To apply this theory we must be able to assign local minima of the PES to the starting and finishing states of interest, which are labeled \(A\) and \(B\). Intervening minima \(i\), which do not belong to either \(A\) or \(B\), are treated in the steady-state approximation, leading to the following expression for the rate constant between minimum \(a \in A\) and \(b \in B\):

\[
f_{ab} = \frac{k_BT}{\hbar} \frac{Z_b}{Z_a} \prod_{i \in I} \frac{p_i(\beta)}{p_i(\beta)},
\]

where the sum is over all transition states that connect a minimum in group \(I\) with a minimum in group \(J\).

The DPS approach introduced in Ref. 22 defines a “discrete path” in terms of a sequence of local minima from the PES and the intervening transition states. To apply this theory we must be able to assign local minima of the PES to the starting and finishing states of interest, which are labeled \(A\) and \(B\). Intervening minima \(i\), which do not belong to either \(A\) or \(B\), are treated in the steady-state approximation, leading to the following expression for the rate constant between minimum \(a \in A\) and \(b \in B\):

\[
f_{ab} = \frac{k_BT}{\hbar} \frac{Z_b}{Z_a} \prod_{i \in I} \frac{p_i(\beta)}{p_i(\beta)},
\]

where the sum is over all transition states that connect a minimum in group \(I\) with a minimum in group \(J\).
where \( k_{ij} \) is the rate constant from minimum \( i \) to minimum \( j \). The sums over \( \alpha_j \) include all minima directly connected to minimum \( j \), aside from degenerate isomers.\(^{22}\) Recrossings between intervening minima \( i \in I \) can be admitted using the formula

\[
k_{ba}^{\text{path}} = \frac{p_a^{\text{eq}}}{p_A^{\text{eq}}} k_{ba}^{\text{eq}} \frac{k_{ia}}{\sum_j k_{ja}} \sum_{n=1}^\infty \left[ A^p \right]_{n1},
\]

(6)

where the sums are over \( p \) odd or even for \( n \) even or odd, respectively.\(^{22}\) and \( A \) is given by

\[
A_{\alpha \beta} = \begin{cases} k_{ia}/\sum_j k_{ja}, & |i_a - i_\beta| = 1, \\ 0, & \text{otherwise}. \end{cases}
\]

The DPS approach provides an analog for the dynamical transition path sampling method\(^{11-13,35,36}\) based on stationary points of the underlying potential energy surface. Starting from an initial discrete path between \( A \) and \( B \) we systematically perturb it by seeking alternative connections between minima in different positions to produce new paths. The DPS run terminates when a predetermined maximum number of perturbations have been considered for the fastest discrete paths, as described in detail for previous work.\(^{22,23}\)

**B. Analyzing the DPS database**

One advantage of the DPS approach is that the database of stationary points created during the run can be analyzed in various ways to extract rates both during and after the DPS run itself. Within the steady-state approximation for intervening local minima of the PES, Eq. (5) provides the simplest DPS expression for the rate between two particular minima from the \( A \) and \( B \) sets. The total rate is obtained by summing over all members of these sets. Two other independent approaches were also used in the present work to obtain overall rate constants from the DPS database of stationary points, namely master equation (ME)\(^{37,38}\) and kinetic Monte Carlo (KMC) calculations.\(^{39-41}\) In a master equation analysis the coupled first-order differential equations that govern the time dependent occupation probability of each local minimum are solved, either numerically, or by casting the problem in matrix form.\(^{37,38}\) The KMC method uses the same rate constants for transitions between local potential energy minima as the ME and DPS methods, but involves a stochastic approach. In the present work we calculated a mean waiting time using KMC for transitions between different groups of local minima using the \( n \)-fold algorithm of Bortz, Kalos, and Lebowitz.\(^{39}\) We then estimated first-order rate constants by inverting the mean waiting time.

For the systems considered in Ref. 22 the total rates evaluated by the ME and KMC methods are in good agreement, and correspond closely to the sum over the fastest paths using the DPS expression in Eq. (5). Since the ME and KMC rates do not include the steady-state assumption for the intervening minima, which is implicit in the DPS expressions, these results indicate that the steady-state hypothesis works well for these systems. The ME and KMC rates also include the contributions of all possible discrete paths between regions \( A \) and \( B \) in the DPS database, not just those from the fastest paths.

In the DPS approach we typically construct thousands of discrete paths explicitly in the perturbations of the fastest paths. However, the resulting databases of stationary points may contain a large number of alternative possibilities. These discrete paths will only be considered in the DPS pathway perturbation process if they are among the fastest. Nevertheless, they still contribute to the overall rate, and in some systems the dominant contribution to this rate can come from a very large number of smaller terms in the sum over paths. This is the situation for the GB1 hairpin, and for a number of other systems, which will be analyzed in detail elsewhere.\(^{42}\) Sampling sufficient alternative pathways in such systems is likely to be particularly challenging.

Fortunately, it is possible to sum over all possible discrete paths supported by the database of stationary points in several ways. Both the ME and KMC approaches include all these terms automatically. The sum over all possible contributions in Eq. (5), which includes the additional steady-state assumption for intervening local minima, can also be evaluated quite efficiently using a matrix multiplication method. To do this we extend the weighted adjacency matrix technique in Eq. (6) for a single discrete path\(^{22}\) to include all the minima and transition states in the database. The resulting matrix \( B \) has dimension \( n_{\text{min}} \times n_{\text{min}} \), where \( n_{\text{min}} \) is the number of minima, and nonzero elements

\[
B_{ba} = \frac{p_a^{\text{eq}}}{p_A^{\text{eq}}} k_{ba},
\]

(8)

\[
B_{bi} = k_{bi},
\]

\[
B_{ij} = \frac{k_{ij}}{\sum_j k_{ji}},
\]

\[
B_{ia} = \frac{p_a^{\text{eq}}}{p_A^{\text{eq}}} \frac{k_{ia}}{\sum_j k_{ja}},
\]

where \( i \) and \( j \) refer to all minima not in sets \( A \) and \( B \), and the sums are over all minima connected to \( i \). For a single \( a \) minimum \( p_a^{\text{eq}}/p_A^{\text{eq}} = 1 \) and

\[
k_{ba} = \sum_{p \in B} \left[ B^p \right]_{ba}.
\]

(9)

The term involving \( B^p \) contains the appropriately weighted sum of contributions from all paths with \( p \) intervening minima. Results for this matrix method of summing all the DPS contributions are compared with ME and KMC calculations in Sec. III A.

A related matrix-based approach can be used to locate the fastest discrete path in the DPS database. Whenever new transition states and local minima are located during perturbations of the current path they are all added to this database. When combined with existing stationary points the new tran-
sition states may produce fast pathways that have not been sampled explicitly. To ensure that such paths were considered in the DPS procedure we employed a modification of Dijkstra’s algorithm for locating the shortest path between any two vertices in a network.\textsuperscript{33,34} If we treat our database of stationary points as a network, with the minima as vertices and the connecting transition states as bidirectional edges (the rate in one direction is different from the rate in the opposite direction), we can construct a log-weighted adjacency matrix $\tilde{A}$ to represent the kinetic information. The logarithms are required because Dijkstra’s method locates the shortest path in terms of the sum of weights, whereas the rate expression in Eq. (5) involves products. To identify the fastest path from minimum $a$ to minimum $b$ (if $b$ is part of the native state of the peptide, for example), we define

$$
\tilde{A}_{ba} = \ln \frac{P_{eq}^a k_{ba}}{P_{eq}^b},
$$

$$
\tilde{A}_{bi} = \ln k_{bi},
$$

$$
\tilde{A}_{ij} = \ln \frac{k_{ij}}{\sum \gamma_{kl} k_{li}},
$$

$$
\tilde{A}_{ia} = \ln \frac{P_{eq}^a k_{ia}}{P_{eq}^a \sum \gamma_{kl} k_{li}},
$$

where $i$ and $j$ denote intervening minima. Comparison with Eq. (5) then gives

$$
\ln k_{ba} = \tilde{A}_{pi} + \sum_{k \in 2n} \tilde{A}_{ik} - \tilde{A}_{i^*} + \tilde{A}_{ia}.
$$

Expressing the rate as a sum of contributions from edges of the graph enabled us to use Dijkstra’s algorithm to find the discrete path with the largest value of $\ln k_{ba}$, aside from recrossing contributions. This fastest path was used to seed subsequent DPS runs in the iterative refinement scheme described in Sec. II E.

C. A grouping algorithm

The stationary point databases produced by DPS runs generally contain thousands of structures. In previous work we have examined schemes for grouping local minima together to identify regions of configuration space that are likely to be in local equilibrium.\textsuperscript{33,34,45} The relative free energies of these sets and the barriers between them can then be compared, which enables us to construct disconnectivity graphs based on free energy,\textsuperscript{34,46} and to identify possible kinetic intermediates on the folding pathway.

In the present study a grouping strategy was used in the iterative refinement of the database, which we refer to as $G_R$. In this scheme minimum $a$ was included in a group if the fastest rate from $a$ to the lowest free energy minimum in the group, as calculated by the approach described below in §2.2, was greater than a specified threshold. In execution, the algorithm is very similar to the method described for barrier heights in Ref. 23, which we refer to as $G_B$.

1. GR grouping algorithm

(1) Choose the rate threshold $r$.

(2) Start at the minimum of lowest free energy. This is the lowest minimum $L M_n$ for group $n = 1$, and has free energy $F(L M_n)$.

(3) Search all ungrouped minima $a$ and include them in group $n$ if they are connected to $L M_n$ by a path with rate $k_{LM_n} > r$.

(4) Set $n = n + 1$. Find the lowest potential energy minimum not yet included in a group. Go to step 3, or terminate if all minima are now grouped.

The grouping method $G_R$ is different from $G_B$, since $G_B$ will include minimum $a$ in a group even if it is high in energy above $L M_n$, provided that there are no large barriers on the fastest path from $a$ to $L M_n$. Hence it collects together minima separated by relatively small downhill barriers, whereas the $G_B$ procedure is based on uphill barriers. The $G_R$ grouping procedure is used in Sec. III A to reduce the dimension of the matrices involved in a master equation analysis of the rates. In contrast, the $G_R$ algorithm produces a smaller number of groups, which was more convenient in the iterative refinement phase of the calculations.

D. Constructing an initial path

The first step in studying hairpin formation was the construction of an initial connected path of minima and transition states from a minimized extended structure (denoted $i_{ex}$) to the folded, hairpin structure (taken as the minimized PDB structure). We attempted to do this using a “double-ended” search algorithm,\textsuperscript{22} which involves multiple applications of the nudged elastic band (NEB) technique.\textsuperscript{29–31} This initial step proved technically difficult, as the double-ended search algorithm (described in Ref. 22) consistently failed to connect these two endpoints in a practical number of attempts. Too many intervening minima and transition states that were not connected to either endpoint were found without mapping out a realistic path. Varying the number of NEB images and the size of the spring force constant was also unproductive.\textsuperscript{37}

A different procedure was found to be necessary: An initial NEB run of 100 steps with 99 images was carried out, and rather than selecting one candidate image for a transition state search, every resulting image was minimized. Starting from endpoint $a$, these 99 minima were placed in an ordered list according to their image number, and the one with the closest RMSD to $a$ was chosen as a putative intermediate minimum $i$. The same procedure was then carried out finding the minimum $j > i$ closest to $i$ according to its RMSD, and continuing until the endpoint $b$ was reached. The selected intermediates $i, j, k, \ldots$, were then connected using the double-ended search algorithm ($a$ connected to $i$, $i$ to $j$, $j$ to $k$, etc.), to give a complete path. In fact, it was necessary to perform this procedure more than once, carrying out another 99 image NEB run followed by minimization between two of these putative intermediates, before a final connected path was eventually found. This approach is similar to the adaptive NEB technique,\textsuperscript{47} but differs in the minimization of the images before the “focusing” of the NEB algorithm. Such
iterative procedures have the advantage that subsections of the path can be generated simultaneously in parallel, and our results suggest that generating intermediate minima in this way is also more efficient in overall computational time. The initial path used to seed the DPS run contained 77 transition states between $i$ and a folded minimum, while for the fastest folding path that we located this number is reduced to 38.

### E. Optimizing the path ensemble

Once a connected path had been found linking the extended and minimized PDB structures, it was optimized using a single DPS run, with parameters as shown in Tables I and II. This run was terminated once 1000 minima had been found, and the resulting database of stationary points was then refined using an iterative approach similar to that used in Ref. 23. The procedure was accelerated by employing the fastest path in the database, located using Dijkstra’s algorithm (Sec. II B) to “seed” each DPS run:

1. Choose $n_{\text{jobs}}$, the number of DPS runs to carry out at each iteration.

2. Group minima in the database together with the $G_R$ algorithm as described in Sec. II C. $r$ was chosen as $10^{-8} \text{ s}^{-1}$, as this gave a manageable number of groups.

3. Select $n_{\text{jobs}}$ pairs of groups to connect. The groups were chosen randomly, with the probability of selecting each group $I$ weighted by its equilibrium occupation probability $P_{i}^{\text{eq}}$. This procedure ensures, given enough iterations, that the thermodynamically relevant regions of the landscape are connected via kinetically relevant paths. More than one job connecting the same two groups in the same iteration was not allowed, as this would be a duplication of effort, and groups were not further connected amongst themselves.

4. Carry out $n_{\text{jobs}}$ DPS runs connecting the selected group pairs; ideally these jobs run in parallel on $n_{\text{jobs}}$ CPU’s. The initial path for each DPS run is chosen to be the fastest path in the database between the pair of minima of lowest energy in the two groups as opposed to generating the path from scratch using the double-ended search algorithm as was done for metenkephalin in Ref. 23.

5. When each of the jobs in Step 4 finishes, start a new

---

### Table I

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_{\text{max}}^{\text{cyt}}$</td>
<td>Maximum number of iterations</td>
<td>30</td>
</tr>
<tr>
<td>$n_{\text{min}}^{\text{cyt}}$</td>
<td>Minimum number of NEB images</td>
<td>3</td>
</tr>
<tr>
<td>$n_{\text{max}}^{\text{cyt}}$</td>
<td>Maximum number of NEB images</td>
<td>15</td>
</tr>
<tr>
<td>$\Delta n_{\text{cyt}}$</td>
<td>Increment in the number of NEB images</td>
<td>4</td>
</tr>
<tr>
<td>$n_{\text{max}}^{\text{EB}}$</td>
<td>Maximum number of minima in the path</td>
<td>1000</td>
</tr>
<tr>
<td>$n_{\text{min}}^{\text{EB}}$</td>
<td>Maximum number of transition states in the path</td>
<td>100</td>
</tr>
<tr>
<td>RMS force convergence criterion for NEB iterations</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>Maximum step length in LBFGS minimisations</td>
<td>0.1Å</td>
<td></td>
</tr>
<tr>
<td>Maximum iterations in tangent space near convergence</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Maximum iterations in tangent space away from convergence</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Pushoff in pathway calculations</td>
<td>0.1Å</td>
<td></td>
</tr>
</tbody>
</table>

### Table II

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_{\text{max}}^{\text{pert}}$</td>
<td>Maximum number of perturbations per path</td>
<td>10</td>
</tr>
<tr>
<td>$n_{\text{fast}}$</td>
<td>$n_{\text{max}}^{\text{pert}}$ perturbations are performed for the $n_{\text{fast}}$ fastest paths</td>
<td>10</td>
</tr>
<tr>
<td>$N_{\text{max}}^{\text{pert}}$</td>
<td>Maximum number of perturbations for a local minimum</td>
<td>50</td>
</tr>
<tr>
<td>$\delta x$</td>
<td>Maximum dihedral angle selection probability, initial value</td>
<td>0.9</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Adjustment factor for $\delta x$</td>
<td>1.02</td>
</tr>
<tr>
<td>$n_{\text{shift}}$</td>
<td>Connections are sought from a new minimum to minima $n_{\text{shift}}$ steps away</td>
<td>4</td>
</tr>
<tr>
<td>RMS force convergence criterion for optimisations</td>
<td>$10^{-5}$ kcal mol</td>
<td></td>
</tr>
<tr>
<td>Smallest number of connections per minimum permitted</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Maximum difference in energy for identical stationary points</td>
<td>$10^{-2}$ kcal mol</td>
<td></td>
</tr>
<tr>
<td>Maximum difference in inertia tensor trace for identical stationary points</td>
<td>10.0 Å²</td>
<td></td>
</tr>
</tbody>
</table>
DPS run connecting the initial extended minimum to one of the two groups in the pair.
(6) When all jobs in Step 4 have finished, immediately terminate all the jobs of Step 5.
(7) Add new minima and transition states found in the DPS runs to the database.
(8) Test for convergence of the property of interest. Terminate or go back to Step 2 for the next iteration.

The set of runs in Step 5 was added to sample the folding process from the extended state to the conformations relevant at equilibrium. The immediate termination once all the pairwise jobs are completed ensured an even use of CPU time across all nodes throughout each iteration, which had not always been the case for met-enkephalin in our previous study.23

III. RESULTS AND ANALYSIS
A. Measuring the folding time

The kinetic property we are primarily interested in is the length of time, $\tau_f$, required on average to reach a folded, hairpinlike structure starting from an initial extended minimum $i_{ex}$. All calculations were performed at 298 K. It was apparent that folded structures did not all lie in one group of minima, as defined by either grouping algorithm $G_B$ or $G_R$, for any sensible value of the grouping parameters. Physically, this result may mean that the folded “state” is not entirely in ergodic self-equilibrium on the experimental time scale, or it could also indicate under-sampling of connections within the folded region. Therefore, it was necessary to define a minimum as folded or unfolded according to structural order parameters. We chose three order parameters, drawing from previous work.8,14 The first was the number of native cross-chain hydrogen bonds $N_{nhb}$, where the six hydrogen bonds between backbone amide groups present in the PDB structure are considered (see Fig. 1). For each minimum, a hydrogen bond is defined to be formed if the N and O atoms are within 3.4 Å and the N–H–O angle is greater than 110°.5 The second parameter was $D_{rmsd}$, the RMSD of all atoms from the minimized PDB structure. The third was $R_h$, the radius of gyration of the hydrophobic core of the peptide, which is defined as the heavy (nonhydrogen) atoms in the sidechains of the tryptophan, tyrosine, phenylalanine, and valine residues (see Fig. 2). Preliminary free energy plots (see Sec. III C) indicated that suitable values of the cutoffs for a folded system were $N_{nhb} \geq 2$, $D_{rmsd} \leq 4.0$ Å and $R_h < 6.0$ Å.

The iterative scheme was run with $n_{jobs} = 10$. The intention was to investigate folding using the master equation technique, treating each minimum as a separate state.25,26,48,49 However, it was not possible to solve the full master equation for the large databases of minima collected by the iterative scheme, even after dead-end minima with $P^{eq} < 10^{-6}$ had been pruned away. The diagonalization of the master equation transition matrix was very slow and did not always produce a clear separation between the one zero eigenvalue and the other (negative) eigenvalues. Instead we therefore coarse-grained the sample further using the grouping algorithm $G_B$ of Ref. 23. A value of $b = 5$ kcal mol$^{-1}$ in $G_B$, followed by pruning of groups with $P^{eq} < 10^{-6}$, enabled the probability to be plotted as a function of time for the first five iterations, as shown for the points labeled “ME group” in Fig. 3.

As mentioned in Sec. II B, the KMC approach provides an alternative stochastic treatment of the dynamics. At each iteration one thousand independent KMC runs were carried out starting from $i_{ex}$, both for the full database of stationary points, and for the grouped, pruned sample used for the mas-

![FIG. 1. The backbone of the minimized PDB structure for the GB1 hairpin. The six native, cross-chain hydrogen bonds used to define $N_{nhb}$ are shown as dashed black lines.](Image)

![FIG. 2. The minimized PDB structure of the GB1 hairpin, drawn with only heavy atoms from the backbone along with the hydrophobic sidechain heavy atoms shown at an exaggerated size. This figure was prepared using MOLMOL.56](Image)

![FIG. 3. Folding time $\tau_f$ for the GB1 hairpin, determined at ten iterations of the refinement algorithm. “KMC full” represents the average over 1000 KMC runs of the time taken to reach a folded minimum (defined as described in the text) using KMC on the full database starting from an extended minimum $i_{ex}$. “KMC group” is the same quantity for the database after it has been grouped and pruned, where the system starts in the group containing $i_{ex}$, and a group is defined as folded if the mean values of the order parameters lie within the limits described in the text. The error bars for the KMC runs represent one standard deviation. “ME group” represents the master equation results for the same grouped and pruned database as “KMC group.” These calculations correspond to a folding time $\tau_f$, defined as the time at which the probability of being in a folded group reaches 0.5. The results for “ME group” and “KMC group” are in good agreement.](Image)
ter equation. The results are plotted in Fig. 3. Good agreement is obtained between the KMC and ME results for the grouped database, but the folding times are systematically longer when the full sample is considered. This difference probably arises from the presence of “dead end” minima in the full database, which act as traps because their connections have not been fully sampled. Hence, although the pruned database contains fewer minima, it may give more accurate results.

Figure 3 indicates that convergence to within about an order of magnitude was obtained in ten iterations of the refinement algorithm (taking around four weeks of CPU time on ten UltraSparc III processors). The folding time predicted by full KMC is roughly 30–90 μs, which is around an order of magnitude slower than the experimental value of 6 μs. This result represents reasonable agreement, given the approximations in the model, including force field inaccuracies and the statistical description of the kinetics. We note, however, that better agreement has been obtained with similar potentials and more intensive computation. Given that the transition state theory rates are an upper bound to the reaction rate,50 we might speculate that undersampling of the reaction pathways is responsible for the somewhat slower folding rate in the present work. In fact, the rates predicted from both the full and the pruned databases lie closest to the experimental value at the first iteration. This agreement is probably fortuitous, but it is possible that the additional sampling may have increased the number of off-pathway minima in the database at the expense of faster paths. An alternative time-resolved iterative procedure may help to clarify the situation, as discussed in Sec. V.

B. Validity of the steady-state approximation

The DPS technique was introduced as a method for finding the rate constants involving two sets of minima, with all other intervening minima treated in steady state.22 We have compared the results calculated using this particular DPS approximation with those obtained from the same database using the KMC and ME approaches to detect if any significant, long-lived, intermediates are in fact present, since the latter techniques do not involve the steady-state approximation. We analyzed the final database from the end of the tenth iteration of the refinement algorithm, which contains 24 802 minima and 36 009 transition states.

The single fastest path from \( i_{ex} \) to the folded state, found using Dijkstra’s algorithm as described in Sec. II B, produces a contribution to the folding rate of \( k_{Ba} = 3.7 \times 10^{-48} \text{s}^{-1} \), where \( a = i_{ex} \) and \( B \) is the set of all folded minima. This is more than 50 orders of magnitude slower than the total folding rate, as obtained by KMC, indicating that a very large number of paths must be involved in the folding process.

Using the matrix multiplication approach of Sec. II B, we obtained \( k_{Ba} = 1.3 \times 10^9 \text{s}^{-1} \). The two alternative approaches to calculating \( k_{Ba} \) from the database of stationary points, namely, master equation and KMC calculations, also correctly account for the full multiplicity of paths. In the present case the full DPS sum obtained by matrix multiplication gives a value for \( k_{Ba} \) that is several orders of magnitude faster than the KMC and master equation results. This difference indicates that there are stable intermediates somewhere on the folding pathway from extended to folded structures. The waiting time in these minima is substantially underestimated by the steady-state approximation used to derive the DPS rate expressions, but is treated correctly by the KMC and master equation techniques.

It would be interesting to compare the rates obtained from Eq. (9) with those calculated by the minimum cut/maximum flow (MF) approach used by Krivov and Karplus.46 The elements of the B matrix consist of individual rate constants, whereas the edges of the network considered by Krivov and Karplus were weighted by the absolute free energy of the transition state. Our approach should be more accurate than the MF expressions for circumstances where the steady-state approximation is appropriate, as Eq. (9) includes the effect of path length and off-pathway intermediates in the rate, whereas the MF method simply considers the highest barrier height on each path.

C. Free energy surfaces

Given a database of minima and their energies and vibrational frequencies, we can calculate the relative equilibrium occupation probabilities of sets of minima using the harmonic superposition approximation (HSA),25,26,51,52 and plot free energy surfaces as functions of any given order parameters. Figure 4 represents the free energy surface as a function of \( N_{nhb} \) and \( R_h \), calculated using the harmonic superposition approximation at 298 K.

![Figure 4. Free energy for the GB1 hairpin as a function of the order parameters \( N_{nhb} \) and \( R_h \), calculated using the harmonic superposition approximation at 298 K.](image-url)

Downloaded 07 Jun 2007 to 131.111.115.149. Redistribution subject to AIP license or copyright, see http://jcp.aip.org/jcp/copyright.jsp
PDB structure is around 3 Å. Comparison with the free energy surfaces plotted in Refs. 5 and 8 does, however, reveal a significant undersampling of the non-native, unfolded state in this work, since there is a far more rapid rise in free energy moving away from the folded structures. This result is not surprising, given that our procedure was only designed to sample the pathways seeded from a path between one extended structure and the native state.

D. Sequence of folding events

Using our final database we carried out a further 1000 independent KMC simulations at 298 K, each lasting $10^{2.3}$ s and starting from $i_{ex}$. Various averages were calculated as a function of time, and these data were analyzed to assess the sequence of folding events, along with the main driving forces toward folding.

Figure 6 shows the time evolution of the three order parameters described in Sec. III C, plotted as the relative deviation from the equilibrium value at 298 K, according to the HSA. Initially, all three deviations fall rapidly, but then reach a “plateau” region where they are roughly constant until around $10^{-6}$ s. At this time, $N_{nhb}$ actually decreases slightly, away from its equilibrium value, apparently to allow some rearrangement of the hydrophobic sidechains, since $R_h$ decreases more rapidly at the same time. In a second and final step, all three order parameters tend to their equilibrium, folded, values at around $10^{2.4}$ s.

Also plotted in Fig. 6 is the time evolution of the folded population, as defined in Sec. III A. It is clear that relaxation into the folded state also occurs around $10^{-4}$ s, the same time as the final decay of order parameters, described above.

Figure 7 shows the values of various components of the potential energy as a function of time, averaged over the same 1000 KMC simulations used for Fig. 6. The difference between the total potential energy $E_{all}$ and $E_{elec} + E_{vdw} + E_{eef1}$ is roughly constant, indicating that these nonbonded terms are mainly responsible for stabilizing the hairpin. In fact, it appears that the main energetic driving force toward folding is electrostatic, given that the change in $E_{all}$ mirrors almost exactly the change in $E_{elec}$. In contrast, $E_{vdw} + E_{eef1}$, which one might loosely think of as the hydrophobic interaction, remains roughly constant throughout the folding process. “Hydrophobic collapse” is therefore not evident for the present potential using implicit solvent. Instead formation of hydrogen bonds to minimize the electrostatic energy seems to be the dominant process, with little intrinsic preference for the “hydrophobic core” sidechains to be either packed together or widely separated.

We can attempt to visualize the folding process by calculating the occupation probabilities of different groups of minima over time. Using the algorithm $G_R$ and $r = 10^{-8}$ s$^{-1}$, as for the DPS refinement process, we grouped...
the final database, and then used the same 1000 KMC simulations described above to calculate the group occupation probabilities, \( P_I(t) \). The most significant occupation probabilities are plotted in Fig. 8. For each of these groups, the minimum lowest in free energy is shown in Fig. 9. It appears that the “plateau” region observed in Fig. 6 is dominated by the relatively loose structures \( F \) and \( G \), which have quite large numbers of hydrogen bonds but no packing of the hydrophobic core. The more compact structures \( C \), \( D \), and \( E \) form a second intermediate state, whose probability peaks at around \( 10^{-5} \) s. The formation of this state occurs at the same time as the dip in \( N_{\text{nhb}} \) discussed for Fig. 6. The analysis was repeated using the \( G_B \) algorithm with \( b = 10 \) kcal mol \(^{-1} \), and very similar results were obtained, indicating that the outcome is not sensitive to the particular grouping method.

IV. DISCONNECTIVITY GRAPH

The final database of stationary points is represented as a free energy disconnectivity graph in Fig. 10. The same groupings of minima were used as in the previous section, enabling us to visualize the location of the structures that were visited in the folding simulations in the context of the entire sampled configuration space. It is clear that our sample of the PES is a reasonably effective funnel of population from \( i_{\text{ex}} \), both from the global appearance of the graph and from the fact that only a few groups trap population to a measurable extent during folding. However, it is noticeable that there are other possible “traps” apparently separated from the main funnel by larger barriers than those visited in the simulations (such as Group \( X \) on Fig. 10). The reasons why such groups do not act as traps in folding from \( i_{\text{ex}} \) will be investigated further in future work.

The graph also illustrates that the iterative DPS procedure has succeeded in sampling a wider range of configura-
tions with non-native secondary structures, such as helices and ion pairs, are sampled significantly in some previous simulations, our database contains only a few minima of this kind. Since the present sampling scheme is based on dynamics this result probably indicates that such structures have little effect on the folding rate.

V. CONCLUSIONS

We have shown how multiple applications of the DPS algorithm can sample an ensemble of paths from an extended local potential energy minimum to the folded state of a β-hairpin forming peptide. The folding times, as predicted by KMC simulations for the database of minima and transition states created by DPS, are about an order of magnitude slower than the experimental results. We find a folding mechanism with two major kinetic intermediates—the first is stable up to around $10^{-6}$ s and mainly consists of loosely hydrogen bonded structures, and the second is dominant at around $10^{-3}$ s and is more compact, with a smaller radius of gyration of the hydrophobic core sidechains. The final folding into the hairpin state happens at around $10^{-4}$ s, when there is a concerted movement of all the order parameters towards their equilibrium values. Therefore, it would appear that this model of the GB1 peptide is not two-state in the usual sense, since there are on-pathway intermediates. While the kinetics are single-exponential at this temperature, with the rate-determining step being passage from the minima in groups C, D, and E to the folded state, there might be some other, more denaturing conditions for which the rate-determining step becomes the transition from F and G to the folded state.

Compared to other studies, our results assign less importance to the hydrophobic core residues in the folding process. According to the force field we used, there is little overall preference in terms of the solvation energy of the molecule for these residues to be packed together, although it is not clear to what extent this is an artifact of the EEF1 implicit solvation potential. The multicanonical MC study of Dinner et al., which used the same potential as our work, located a free energy minimum with no native hydrogen bonds but a well-packed hydrophobic core, of the type often labeled as “H.” The authors concluded that this was a significant folding intermediate. The difference between the two sets of calculations could either arise from incomplete sampling in the present work, or from the sampling scheme itself, which is specifically designed to probe the folding dynamics in the case of DPS. Here we have attempted to locate the folding pathways explicitly, rather than inferring them from free energy surfaces. If the “H” state is a free energy minimum, but has little effect on the folding rate, then the corresponding configurations will not be sampled extensively in the DPS approach. However, although we did not find a dominant population in H state minima, the second intermediate kinetic phase described above does correspond to a slight decrease in $R_h$. Hence, although there are some native hydrogen-bonds present in the structures we associate with this intermediate, the configuration space in question may actually overlap with that identified for the H state in previous works.

Other calculations of free energy surfaces by Zhou et al. showed that the choice of order parameters could affect the conclusions about the order of folding events—these authors found an L-shaped free energy surface as a function of $R_h$ and $N_{hh}$, which would indicate a hydrophobic collapse followed by rearrangement, but the results obtained using other parameters suggested a more concerted mechanism, as the stable unfolded structures had neither hydrogen bonds nor a well-formed hydrophobic core.

Zhou et al. did not find a significant “H” intermediate, but other folding and unfolding studies have identified such structures as important. The PRD study of Zagrovic et al. shows an intermediate “H” state with some key hydrophobic contacts formed, but few backbone residues in hairpin configurations. The final folding process then involves concerted hydrogen bond formation and packing of hydrophobic residues, as in our work and as proposed by Zhou et al. The TPS unfolding study of Bolhuis also located a metastable “H” intermediate, and concluded that the rate determining step for unfolding was motion from the folded state to “H.”

It remains to be seen whether the discrepancies between our work and some other studies are due to differences in force fields, undersampling of configuration space or simply due to difficulties in comparing the structures. In particular, it would be worthwhile to repeat the DPS sampling from a series of extended structures, perhaps obtained from a high-temperature simulation. Alternatively, since the extended structures are not significantly populated at the temperature of interest, future sampling could focus on pathways between the kinetic intermediates and the native state. We might also improve our database refinement procedure by connecting groups of minima selected according to their occupation probability at different times, rather than all at equilibrium, i.e., replace the algorithm of Sec. II E with one that uses information such as that plotted in Fig. 8. It may be that a better sampling procedure will bring the folding rate closer to the experimental one, or that the difference is more systematic and due to the potential or the transition state theory (TST) approximation used to calculate the individual rate constants for barrier crossing. Nevertheless, we find it encouraging than this initial application of DPS to peptide folding gives results that are at least comparable to experiment and to more established methods.

ACKNOWLEDGMENT

D.A.E. is grateful to the EPSRC for financial support.

APPENDIX: MODIFICATION OF THE CHARMM POTENTIAL

As noted in the Appendix of Ref. 23, there are asymmetries in the way the potential is defined in CHARMM, meaning that some permutational isomers of the same struc-
ture have different energies. We modified the CHARMM19 force field to eliminate these asymmetries, and used this potential for the calculations presented here. The difference in energy between our potential and the original CHARMM19 is of the order of $10^{-3}$ kcal mol$^{-1}$, so it is not significant for the calculation of finite temperature properties. However, tightly converged stationary points are required in the present study in order to distinguish between structures with similar energies.