A line integral reaction path approximation for large systems via nonlinear constrained optimization: Application to alanine dipeptide and the β hairpin of protein G

Ilja V. Khavrutskii
Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037

Richard H. Byrd
Department of Computer Science, University of Colorado at Boulder, Boulder, Colorado 80309

Charles L. Brooks III[a)]
Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037

(Received 22 November 2005; accepted 17 March 2006; published online 17 May 2006)

A variation of the line integral method of Elber with self-avoiding walk has been implemented using a state of the art nonlinear constrained optimization procedure. The new implementation appears to be robust in finding approximate reaction paths for small and large systems. Exact transition states and intermediates for the resulting paths can easily be pinpointed with subsequent application of the conjugate peak refinement method [S. Fischer and M. Karplus, Chem. Phys. Lett. **194**, 252 (1992)] and unconstrained minimization, respectively. Unlike previous implementations utilizing a penalty function approach, the present implementation generates an exact solution of the underlying problem. Most importantly, this formulation does not require an initial guess for the path, which makes it particularly useful for studying complex molecular rearrangements. The method has been applied to conformational rearrangements of the alanine dipeptide in the gas phase and in water, and folding of the β hairpin of protein G in water. In the latter case a procedure was developed to systematically sample the potential energy surface underlying folding and reconstruct folding pathways within the nearest-neighbor hopping approximation. © *2006 American Institute of Physics*. [DOI: 10.1063/1.2194544]

I. INTRODUCTION

One of the fundamental problems of computational chemistry is to find a transition state (first order saddle point) connecting a given reactant and a product (two local minima), or in general a connected path, comprising multiple transition states and corresponding intermediates. Traditional approaches seek a single transition state between the reactant and product through an optimization of a single molecule.[1–5] These methods are most successful for transformations involving a few atoms in the molecule, but require a good initial geometry guess for the transition state and a sound approximation to second derivatives of the potential energy (PE). The latter quickly becomes intractable with growing number of atoms in the molecule. Furthermore, transformations involving a large number of atoms likely comprise multiple intermediates and transition states. Therefore, path search for such transformations demands a preliminary exploration of the relevant region of the PE surface, which further complicates the search.[6–8] For such cases double-ended band (also called string or chain of states) methods, which attempt to locate all intermediates and transition states between the reactant and product concomitantly, have been proposed. Among these methods are elastic-band (EB) methods,[9–14] action-based (AB) methods[15–23] and line integral (LI) methods with PE-gradient-free[7,24–28] and PE-gradient-dependent[29,30] objective functions.

The EB methods approximate the steepest descent (SD) path,[31,32] but require an initial guess of the path or band. The band consists of a number of copies or replicas of the original molecule somehow interpolated between the reactant and product, with adjacent replicas connected by special harmonic springs to prevent collapse of the replicas into nearby local minima. Once generated the initial guess is then refined by minimizing the norms of the component of the PE gradient of each replica perpendicular to the band, and of the component of the harmonic restraint gradient parallel to the band. These methods need second derivatives of the PE to be efficient.[14] However, recently a variant of the EB method was proposed which exploited the variable metric limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) approximation[33,34] to second derivatives during the path refinement.[13]

In contrast to EB methods, AB methods provide dynamic trajectories connecting the two end points, and therefore parametrically depend on the total energy. Moreover, AB methods require a large number of replicas in the band to make sound approximation to the action functional. Most of the action-based methods require second derivatives of the PE.[16–23]

The last family of the double-ended methods is the LI methods, which like EB approximate the SD path, but opti-

[a)]Author to whom correspondence should be addressed. Fax: 858/784-8688. Electronic mail: brooks@scripps.edu

mize various objective functions subject to nonlinear constraints on the distance between the points in space. LI methods with PE-gradient-dependent objective functions, similar to EB methods, require evaluation of second derivatives of the PE, therefore leaving PE-gradient-free objective functions the method of choice.[24–27,35]

Because the double-ended methods provide only approximate intermediates and transition states, the pathways need to be further refined to yield exact intermediates and transition states. While it is trivial to refine the intermediates by optimization, pinpointing the transition states can be difficult. Fortunately, a conjugate peak refinement (CPR) method was developed that can easily refine transition states given an optimized path from the double-ended methods.[36,37]

Since the present paper attempts to improve of the original LI method due to Elber, we briefly outline the key features of the LI method. In the original LI method the problem is defined as follows. Given configurations of a reactant R_R and a product R_P, we find a line $l(R)$ of length L connecting the reactant and product that minimizes the corresponding line integral of the potential energy function $U(R)$, namely, $W=L^{-1}\int_{R_R}^{R_P}U(R)dl(R)$.[24] W represents the average of $U(R)$ over $l(R)$. The solution $l(R)$ provides a reaction path.[38–41] In practice the integral equation is discretized into a sum over a limited number of configurations along the line, thus forming a discretized objective function. The distances between adjacent configurations are maintained to remain identical along the discretized line with the help of a nonlinear constraint and, in case a Cartesian coordinate space is used for the path search, additional linear constraints. Ideally, by minimization of such an objective function subject to the specified constraints one aims to find the SD path. In practice, however, an objective function that would locate the SD path without explicit use of the PE gradient has not been proposed yet. Nevertheless, several functions that approximate the SD reaction path have been suggested, including the original LI formulation.[24–27] A number of papers detail why these methods only approximate the SD path.[38–41]

A paramount difficulty with the LI methods is satisfying the essential linear and especially nonlinear constraints that keep the replicas equidistant along the line, while minimizing the objective function. Earlier attempts to solve this problem utilized a penalty function approach.[24–28] However, the penalty function approach suffers from a number of limitations. Specifically, the large Lagrange multipliers that are used for the penalty functions make the optimization numerically inefficient, and, most importantly, the approach provides only an approximate solution to the problem.[42]

Contrary to the penalty function approach, recent developments in nonlinear constrained optimization (NCO) techniques allow rigorous minimization of the objective functions subject to a number of linear and nonlinear constraints.[33,34,43–45] Moreover, coupled with the variable metric L-BFGS update to approximate second derivatives NCO allows treatment of problems of very large size with linear or superlinear convergence rates and linear storage requirements.[33,34,44,45] Most notably the solution obtained through constrained optimization is exact in contrast to the solution from the penalty function approach.[42]

In the present paper we reformulate the LI method devised by Elber[24,27] to make it more suitable for the nonlinear constrained optimization ansatz. While doing so we demonstrate that the LI methods in contrast to the EB methods do not require an initial guess for the path, which presents a significant advantage for molecular mechanics (MM) and quantum mechanics (QM) applications where initial paths either cannot be constructed or do not allow energy evaluation due to unrealistic configurations. In connection to the path optimization, we provide a simple recipe to visualize various cyclic and noncyclic paths and analyze their quality for systems of high dimensionality. We then test this on conformational rearrangements of the alanine dipeptide both in the gas phase and in solution, and finally explore folding pathways of the 16-residue peptide—β hairpin of protein G—in solution.

II. FORMULATION

We define the discretized LI objective function for the chain of N molecules, including the reactant and product, as follows:

$$\Omega(R)=\sum_{i=2}^{N-1}U(R_i)+\left(\frac{\pi}{2a}\right)^{3/2}\delta^3\sum_{i=1}^{N-2}\sum_{j=i+2}^{i+2+K}\exp\left(-\frac{a|R_i^s-R_j^s|^2}{2\delta^2}\right), \tag{1}$$

where R_i and $U(R_i)$ are the all-atom Cartesian coordinate vector and respective potential energy function of the replica i in the chain. The collective coordinate R comprises coordinates of all replicas in the chain, $R=(R_1,\cdots,R_i,\cdots,R_N)$. Because geometries of the reactant and product (replicas 1 and N, respectively) remain fixed, their energies also remain constant, and hence the first term of the objective function includes the potential energy functions of the active replicas (2 through $N-1$) only. The second term represents the repulsion between the replicas other than those that are adjacent, a so-called self-avoiding walk.[27] The superscript s in R_i^s designates the representative selection of atoms used to define the path subspace.[14] The integer K controls the number of interacting replicas. The repulsive term was introduced in a slightly different form by Elber to prevent clustering of replicas at the wells of potential energy surface.[27] In our formulation every replica in the selection subspace is represented by a Gaussian of the form

$$G_i(R^s)=\exp\left(-\frac{a|R^s-R_i^s|^2}{\delta^2}\right). \tag{2}$$

The overlap between Gaussians i and j defines individual components of the repulsive term. Hence the repulsive term is simply cumulative pairwise overlap between Gaussians of interacting replicas. The parameter a controls the width of each Gaussian, and the independent scalar variable δ represents the distance between adjacent replicas in the selection subspace. Note that in the original LI method the analog of δ was simply the average of all of the distances between adjacent replicas along the discretized line.

Furthermore, in sharp contrast to the original LI method, where the equidistance between the adjacent replicas is

maintained with a single cumulative nonlinear constraint, we use $N-1$ nonlinear constraints, one for each adjacent pair, and keep them synchronized with the same scalar variable δ,

$$\{C_i = |R_i^s - R_{i+1}^s|^2 - \delta^2 = 0\}_{i=1,N-1}. \quad (3)$$

For efficiency of optimization we prefer Cartesian coordinate space, hence we need to ensure that during optimization replicas neither translate nor rotate about the selection center of mass, which would otherwise affect the Cartesian distances between replicas. Following Elber we impose six additional linear constraints per each active replica, the so called Eckart conditions,[46–49]

$$C_{\mathrm{tr},k}(R_i^s) = \sum_{j=1}^{M^s} m^j r_{i,k}^j = 0, \quad k = x,y,z, \quad (4)$$

$$C_{\mathrm{rot},k}(R_i^s) = \sum_{j=1}^{M^s} m^j [r_{i,k+1}^{0j}(r_{i,k+2}^j - r_{i,k+2}^{0j}) - r_{i,k+2}^{0j}(r_{i,k+1}^j - r_{i,k+1}^{0j})] = 0. \quad (5)$$

Equation (4) constrains the centers of mass, whereas (5) constrains the pseudoangular momenta of all active replicas. Here $r_{i,k}^j$ and $r_{i,k}^{0j}$ are the kth component of the jth selected atom vector of the ith active replica and its reference, respectively, m^j is the mass of the jth selected atom, and M^s is the total number of selected atoms in each replica. To obtain the reference coordinate set we first prepare the given endpoints by translating their coordinates into the respective centers of mass of the selected atoms. We then find the rotation matrix U^{N1} (3×3) to best fit the selected atoms of replica N onto the corresponding atoms of replica 1 in the least-squares sense,[50] and apply the resulting transformation to all atoms of replica N.

The reference coordinates for the active replicas can be generated by various interpolation procedures. The most common and simplest of all is linear interpolation between the superimposed endpoints in the full Cartesian space,[24] but other interpolation techniques, e.g., interpolation in dihedral space, can also be used. Upon generation, the reference replicas are translated into their respective selection centers of mass and then best fitted to replica 1, producing the reference path,

$$\{R_j^0\}_{j=1,N}. \quad (6)$$

To start the optimization procedure we need to initialize the coordinates of active replicas in the band. In principal we could use the reference path coordinates for that; however, interpolated structures often have multiple steric clashes. Steric clashes in the reference replicas do not present any problem for path optimization because the reference path is never used outside of Eqs. (5). Such a poor approximation to the path, if used as an initial guess, may be tolerable with inexpensive MM force fields, but precludes the use of demanding QM potentials. Therefore better initial guesses are required for QM studies.

One particularly trivial approach to construct a steric-clash-free initial guess, suited for both MM and demanding QM methods, is to place all active replicas into one of the two wells. Provided that the path is not cyclic the replicas will spread out of the wells trying to satisfy the equidistance constraint. The self-avoiding walk term will provide an additional impetus to the path optimization from such an initial point, but is not instrumental for this purpose. The main purpose of the self-avoiding walk term is to control the δ scalar variable, and hence the resolution of the path, as will become apparent later. To avoid numerical difficulties in the early steps of optimization from such an interpolation-free initial guess we translate each active replica j by $(j-1)\varepsilon$, where ε is a small number compared to δ.

We anticipate that optimization from the interpolation-free initial guesses will greatly benefit the QM community, provided the step size is controlled to avoid unrealistic geometric configurations of replicas during optimization. Optimal strategies can be devised to pass molecular orbitals along the chain minimizing the self-consistent-field (SCF) efforts. Moreover, such an interpolation-free initial path lifts the bias inherent to methods of the EB family, which, as has been noted above, can only refine a given path. In fact, there are numerous ways one can construct interpolation-free initial paths, e.g., place P active replicas into the product well and N-P-2 replicas in the reactant well. To the best of our knowledge the existence of an interpolation-free initial guess for path minimization in the LI framework has not been recognized in the literature prior to this publication. However, an analogous initial guess has been used for double-ended trajectories in the Onsager-Machlup AB method by Elber.[18]

With the new method, searching for the final path becomes equivalent to optimizing the objective function (1) subject to N-1 nonlinear constraints (3) and $6(N$-2) linear constraints (4) and (5). To solve this problem, instead of a penalty function approach as used in the original LI method we take advantage of recent developments in large scale nonlinear optimization techniques[33,34,43–45] as implemented in the program KNITRO 3.1.[51] For the purpose of the present study we implemented the new variation of the line integral method into the CHARMM program [52] and interfaced CHARMM with KNITRO 3.1.[51]

III. COMPUTATIONAL DETAILS

We utilized the all-atom CHARMM22 and united-atom CHARMM19 force fields for the alanine dipeptide and β-hairpin studies, respectively, unless noted otherwise. In the case of alanine dipeptide, the solvent water was modeled with both the generalized Born[53,54] (GBORn) and the generalized Born molecular volume[55,56] (GBMV) implicit solvent models as implemented in CHARMM version c31b1, whereas for β hairpin only the GBORn model was used due to rotational variance of the GBMV energy. In either case energy and gradient terms associated with the surface tension were neglected. In the case of GBMV, the existing code was modified to allow use with the replica path routines within CHARMM. The CHARMM19 united-atom force field was employed with the GBORn model. All nonbonded interactions were computed without cutoffs. We did, however, verify that introducing a 14 Å cutoff does not significantly affect the L-BFGS Hessian update in the case of β hairpin. To project

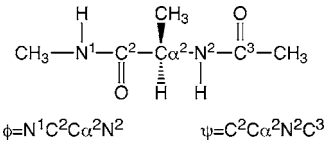

FIG. 1. Schematic representation of the alanine dipeptide, along with definition of the φ and ψ dihedral angles.

isomerization pathways of the alanine dipeptide onto the φ/ψ potential energy surface (PES), the corresponding PESs were computed through minimization with harmonic restraints with a force constant of 100 kcal/(mol deg^2) on the respective dihedral angles, sampling the PES in 10° increments in φ and ψ. Where applicable for unconstrained energy minimization of individual replicas, we used either a series of SD followed by adaptive basis Newton-Raphson (ABNR) minimization or the unconstrained KNITRO minimizer with L-BFGS Hessian update in Cartesian coordinates. To refine the transition states along the generated pathways we used the CPR method.[36,37,57–59]

IV. RESULTS AND DISCUSSION

A. Alanine dipeptide cyclic paths

In the present section we demonstrate the ability of the modified LI method to find approximate cyclic paths for conformational rearrangements of the alanine dipeptide, schematically shown in Fig. 1, both in gas phase and in water solution. The latter is simulated with the GBORn and GBMV implicit solvent models. Because in the cyclic paths the reactant is identical to the product, linear interpolation in Cartesian space cannot be used to produce a reference path. Therefore, we generate cyclic reference paths by linearly incrementing the conventional φ and/or ψ angles. The generated reference paths also serve as initial paths for optimization.

In the following paragraphs we discuss only two nontrivial cyclic paths, namely, conrotary (con-$\varphi\psi$) and disrotary (dis-$\varphi\psi$) paths, originating from the lowest energy point on the respective PESs. In the con-$\varphi\psi$ path the angle increment $\Delta\varphi$ is equal to $\Delta\psi$, whereas in the dis-$\varphi\psi$ path $\Delta\varphi$ is the negative of $\Delta\psi$. Furthermore, because alanine dipeptide has freely rotating methyl groups, which affect the distance between the replicas without significant impact on energy, the reaction subspace was chosen to comprise only heavy atoms (s=**ha**).

1. Gas phase

The gas phase PES of alanine dipeptide has three minima interconnected by four accessible transition states.[60] The minima are **1** (C_{7eq}) at [−81.4, 70.5] (0.0 kcal/mol), **2** (C_5) at [−151.4, 170.6] (0.9 kcal/mol), and **3** (C_{7ax}) at [69.7, −67.6] (2.1 kcal/mol). Both cyclic paths start from **1**, which is the global minimum on the PES.

We first evaluate the method with different numbers of replicas. For the con- and dis-$\varphi\psi$ paths we ran 10-, 20-, and 40-replica jobs with interaction number K=1 and three different a values reciprocal to the number of replicas, i.e., $a=10/N$, $20/N$, and $40/N$. The results are summarized in Table I (see also supporting information[60]).

As seen from the Table I, with ten replicas the resolution of the path is low, with the final δ falling in the range of 1.90–2.96 Å. The potential energy of the band strongly depends on the a parameter. Smaller a parameters result in greater internal band strain as reflected by the higher overlap values, which in turn causes significant deviations from the optimal path.[60] It follows from Table I that stiffer bands take many fewer iterations to optimize, and hence may be used to produce a good initial path at relatively low cost.

Another point to mention is that in low resolution paths with large δ values kinking of the band that is inherent to EB and LI methods[11,39–41] is not an issue and in principal the path may get very close to the ideal minimum energy path provided the equidistance constraints and internal band strength allow for that. In particular, for the con-$\varphi\psi$ path with a=4.0, which has very small internal band strain, one of the replicas in the band accidentally hits the transition state between points **1** and **3** (data shown in supporting information[60]). Increasing the number of replicas, while increasing the resolution, does not necessarily improve the agreement with the ideal path. For example, the same transition state between **1** and **3** on the con-$\varphi\psi$ path with 20 replicas is best approximated by intermediate value a=1.0 and this estimate is notably worse than in the 10-replica case. Furthermore, in agreement with recent analysis,[27,38–40] the use of higher numbers of replicas inflicts kinking of the band, with a number of replicas clustering around point **2** in the 20-replica case. As the number of replicas increases to 40 this problem becomes even more profound, with more points clustering near minima **2** and now **1**. The kinks can be recognized as aberrations of the main path line leading to clustering of replicas in the vicinity of local minima. Such clustering acts as weight pulling the band, which in turn causes the band to shift away from the true transition state between points **1** and **3** for example. Similar observations can be generally made for the dis-$\varphi\psi$ path by inspecting Table I [interestingly, in this case stiff bands miss even the minimum **2** (Ref. 60)].

Because kinking is an inherent problem even with the presence of the self-avoiding walk term,[27] we explored the question how does one monitor kinking, such that one is able to judge the quality of the path. Although, for the alanine dipeptide the φ/ψ map provides sufficient information about the path quality, such maps are not available for more complex systems in general. Therefore we proposed to use a simple set of coordinates to plot a path for a system of any dimensionality, which would be equally suited for both cyclic and acyclic paths,

$$\{|R_1^s| - |R_i^s|; |R_1^s - U^{i1}R_i^s|\}_{i=1,N}. \quad (7)$$

For acyclic paths alternatively one can use the following representation of the path:

$$\{|R_1^s - U^{i1}R_i^s|; |R_N^s - U^{iN}R_i^s|\}_{i=1,N}. \quad (8)$$

Here U^{i1} is the best-fit transformation of replica i onto replica 1 in the selection subspace, described in the Formulation

TABLE I. Optimization summary for gas phase cyclic paths of the alanine dipeptide. Unless otherwise specified in all entries the interaction parameter $K=1$.

No. of replicas (No. of variables)	a	δ (Å)	ΣU (kcal/mol)	Walk (Å^3)	Objective function values	No. of iterations	No. of gradient evaluations
			con-$\varphi\psi$				
10 (529)	1.0	1.90	−131.96	20.32	−111.64	928	797
	2.0	2.02	−133.72	2.18	−131.54	2 087	1 598
	4.0	2.17	−134.19	0.09	−134.10	2 951	2 114
20 (1189)	0.5	0.95	−263.25	33.76	−229.49	718	681
	1.0	1.09	−271.11	8.27	−262.84	1 475	1 264
	2.0	1.52	−282.86	7.08	−275.78	3 206	2 558
40 (2509)	0.25	0.49	−534.93	44.69	−490.24	1 104	1 068
	0.5	0.63	−558.41	23.78	−534.63	1 190	1 118
	1.0	0.79	−569.98	10.21	−559.77	2 460	2 354
80 (5149)	0.125	0.29	−1114.02	69.36	−1044.66	2 205	2 138
	0.25	0.37	−1152.05	42.92	−1109.13	1 833	1 735
	0.5	0.47	−1180.70	26.35	−1154.35	3 240	2 892
160 (10429)	0.0625	0.18	−2325.48	113.87	−2211.61	2 591	2 221
	0.125	0.24	−2402.57	79.92	−2322.65	6 804	6 636
	0.25	0.30	−2449.38	54.66	−2394.72	5 550	5 385
	0.25[a]	0.21	−2360.14	53.36	−2306.78	12 901	10 472
	1.0[a]	0.34	−2468.23	31.96	−2436.27	10 439	9 903
			dis-$\varphi\psi$				
10	1.0	2.03	−120.22	27.64	−92.58	835	796
	2.0	2.55	−132.49	7.40	−125.09	884	852
	4.0	2.96	−137.91	2.71	−135.20	1 502	1 446
20	0.5	1.03	−240.80	43.58	−197.22	847	808
	1.0	1.25	−256.44	11.71	−244.73	2 028	1 941
	2.0	1.45	−262.55	5.10	−257.45	1 700	1 569
40	0.25	0.55	−494.63	63.52	−431.11	912	879
	0.5	0.71	−529.67	34.48	−495.19	1 136	1 115
	1.0	0.88	−545.26	12.43	−532.83	2 927	2 815

[a]$K=16$.

section. Although such representation does not yield unique points on two-dimensional (2D) plots for unique structures, it may still prove helpful to analyze different paths generated by LI methods.

To evaluate the convergence of the path with respect to the number of replicas and also to study the effect of the interaction parameter K of the walk term, we computed 80- and 160-replica con-$\varphi\psi$ paths with $K=1$ and an additional 160-replica path with $K=16$.

Minima **1** and **2** become the centers of attraction (attractors) for kinks with 40 and 80 replicas, whereas with 160 replicas the lowest energy point **1** becomes the ultimate attractor.[60] The walk interaction parameter does not affect the path in any significant way. Nevertheless, increasing the interaction number K increases the δ value from 0.30 to 0.21 Å, as seen from Table I. Such a small increase in resolution comes at great expense in terms of the number of iterations, which passes the 10 000 mark. Overall the paths obtained with 80 and 160 replicas are sufficiently close to ones with 40 replicas with final δ values in the range from 0.49 to 0.88 Å.

2. Implicit solvent

a. GBMV. To obtain con- and dis-$\varphi\psi$ paths for conformational isomerization of the alanine dipeptide in water we first used the GBMV implicit solvent model with the CHARMM22 parameter set.[55,56,61] We would like to note that the GBMV φ/ψ PES of the alanine dipeptide appears rugged in contrast to the smooth gas phase PES. This may present difficulty for second derivative updates based on the gradient, and also will render the paths less smooth. Besides, ruggedness makes it more difficult to pinpoint the actual local minima by unconstrained minimization. In accord with previous studies, the position and the number of minima on the GBMV φ/ψ map are different from those on the gas phase map,[62–66] but very similar to the free energy map obtained in explicit solvent.[67] In particular, there are four minima interconnected with seven transition states with GBMV implicit solvent.[60] The minima are approximately as follows: **1** (α_R) at [−95.7,−68.4] (0.0 kcal/mol), **2** (C_{7eq}) at [−81.3, 147.5] (0.5 kcal/mol), **3** (C_{7ax}) at [58.0, −116.7] (3.7 kcal/mol), and **4** (α_L) at [52.7, 58.7] (7.2 kcal/mol). Note that the energy ordering for minima **1** and **2** in the explicit solvent map agrees with that in GBMV solvent model. However, the energy difference between **1** and **2** in explicit solvent is only 0.3 kcal/mol.[67]

We used 40 replicas to minimize the corresponding paths, and in the case of con-$\varphi\psi$ we additionally used 80 replicas to check convergence. The results are summarized in Table II. Similar to gas phase, the GBMV bands kink around

TABLE II. Optimization summary for the GBMV implicit solvent cyclic paths of the alanine dipeptide. Interaction parameter $K=1$.

No. of replicas (No. of variables)	a	δ (Å)	ΣU (kcal/mol)	Walk (Å^3)	Objective function values	No. of iterations	No. of gradient evaluations
				dis-$\varphi\psi$			
40 (2509)	0.25	0.49	−1087.01	45.68	−1041.33	789	748
	0.5	0.62	−1115.50	23.50	−1092.00	1051	922
	1.0	0.76	−1123.89	9.68	−1114.21	1484	1244
80 (5149)	0.125	0.29	−2216.05	65.46	−2150.59	1583	1523
	0.25	0.38	−2273.64	46.78	−2226.86	1231	1167
	0.5	0.49	−2307.28	27.88	−2279.40	1579	1468
				con-$\varphi\psi$			
40	0.25	0.56	−1113.84	68.26	−1045.58	775	754
	0.5	0.66	−1139.58	27.82	−1111.76	1336	1301
	1.0	0.77	−1151.43	11.11	−1140.32	2321	2211

the lowest energy point **1**.[60] Also as expected, the path trace on the φ/ψ map has substantial ruggedness. Nonetheless, the dis-$\varphi\psi$ path and con-$\varphi\psi$ path provide accurate estimates for the **1–2** and **2–4**, and **1–2** and **1–3** transition states, respectively. Higher energy transition states are predicted with less accuracy due to kinking effects.

A clear demonstration of the ruggedness of the GBMV PES is given by minimization of individual replicas from the optimized LI paths (the results of such minimization for the 40-replica con- and dis-$\varphi\psi$ paths with $a=0.25$ and $K=1$ are shown in supporting information[60]). Due to this ruggedness even multiple applications of alternating SD and ABNR optimizers in CHARMM cannot locate the exact positions of the local minima on GBMV PES. Instead, optimization gets trapped in various places on the φ/ψ map, sometimes close to the true minima. For comparison with the CHARMM SD and ABNR duet we used the KNITRO optimizer in the absence of any constraints and obtained a similar result. Because of the ruggedness and dependence of GBMV energies on the molecule orientation seen in the alanine dipeptide case, and exacerbation of these problems for larger molecules (results not shown), we explored the alternative implicit solvent model GBORn.

b. GBORn. The GBORn implicit solvent model,[53,54] unlike GBMV, does not rely on a grid for computation of the necessary volume integrals,[55,56,61] and, furthermore, has been previously adopted for use with the replica path routines in CHARMM.[68] Because the GBORn model has been originally optimized for the united-atom CHARMM19 force field, the all-atom representation of the alanine dipeptide used in the gas phase and GBMV studies has been replaced for the united-atom representation. As in the previous cases, we optimize both con- and dis-$\varphi\psi$ cyclic paths with the GBORn model.

The GBORn PES for the alanine dipeptide is somewhat different from that of GBMV. In particular, instead of four local minima as in the GBMV case, the GBORn PES has five minima.[60] Nevertheless, the positions of the four major local minima are fairly close to those on the GBMV surface. The minima are **1** (C_{7eq}) at [−77.0, 135.1] (0.0 kcal/mol), **2** (α_R) at [−74.1, −38.1] (0.7 kcal/mol), **3** (α_L) at [54.2, 45.5] (3.8 kcal/mol), **4** (C_{7ax}) at [60.2, −71.4] (2.8 kcal/mol), and **5** (α_L') at [60.2, −166.3] (4.6 kcal/mol). Furthermore, the GBORn surface looks very similar to the free energy surface of CHARMM19 alanine dipeptide with explicit water molecules, and again the energy ordering of C_{7eq} and α_R agrees with the explicit water result, which favors C_{7eq} over α_R by 1.4 kcal/mol.[69] As expected the GBORn PES for alanine dipeptide is as smooth as the gas phase PES. Consequently, optimization of individual replicas is facile and usually gets to the very bottom of a nearby well. The relative energies of the GBORn minima differ from those of the GBMV. Contrary to GBMV, the C_{7eq} lies lower than α_R. For this reason cyclic paths optimized with GBORn were started from the point C_{7eq}.

Our preliminary findings have revealed that optimization of paths using the GBORn model from the very beginning could sometimes fail. In particular, con-$\varphi\psi$ paths cannot be located, whereas dis-$\varphi\psi$ paths have been optimized seamlessly. It is possible that the LI objective function may become non convex when the GBORn term is added to the potential, which causes the optimizer failure. To overcome this problem we have devised the following strategy for the path minimization. Given reactant and product minima optimized with GBORn we (i) optimize the path in the gas phase, (ii) turn on GBORn and gradually increase ε from 1.0 to 80.0 through a number of successive optimizations and (iii) subject the resulting path to a few rounds of optimization at final $\varepsilon=80$ until complete convergence. The strategy succeeded in the majority of attempted path optimizations. However, the final path obtained with the described strategy may differ from that optimized exclusively with GBORn.[60] Table III contains a summary of the LI optimization with GBORn.

B. Alanine dipeptide acyclic paths and interpolation-free initial guess

In the present section we demonstrate the important ability of the reformulated LI method to build a path without an initial guess, or to be more precise with an interpolation-free initial guess, as suggested in the Formulation section. For this purpose we seek three different paths interconnecting each pair of the three local minima of the alanine dipeptide

TABLE III. Optimization summary for the GBORn implicit solvent cyclic paths of the alanine dipeptide. Interaction parameter $K=1$.

No. of replicas (No. of variables)	a	δ (Å)	ΣU (kcal/mol)	Walk (Å^3)	Objective function values	No. of iterations	No. of gradient evaluations
				dis-$\varphi\psi$			
40 (1369)	0.25	0.66	−2088.53	106.35	−1982.18	892	801
	0.5	0.78	−2118.00	39.56	−2078.44	1392	1266
	0.75	0.97	−2140.20	26.58	−2113.63	3074	2702
	1.0	1.12	−2152.62	19.25	−2133.37	3324	3126
40[a]	0.25	0.66	−2088.56	106.38	−1982.18	1722	1415
	0.5	0.80	−2120.43	42.35	−2078.08	5192	4612
	0.75	0.98	−2142.18	29.62	−2112.55	7310	6656
	1.0	1.14	−2154.63	22.18	−2132.45	6228	5728
				con-$\varphi\psi$			
40[a]	0.25	0.67	−1989.39	110.41	−1878.99	2443	2140
	0.5	0.89	−2052.23	61.31	−1990.92	4702	4197
	0.75	1.10	−2086.38	44.03	−2042.35	5950	5384
	1.0	1.28	−2103.83	31.70	−2072.13	7416	6770

[a]Multistage optimization strategy has been applied (see text for description).

in the gas phase, from either the linearly interpolated or an interpolation-free initial guess. We arbitrarily use 80 replicas per path. In all three cases we used a linear reference path. The interpolation-free guess with all replicas in the reactant well is referred to as **free1**, whereas with all active replicas in the product well, as **free2**.

The results are summarized in Table IV and Fig. 2. Table IV does not reveal any discouraging trends in the number of iterations associated with the interpolation-free initial guesses: in some cases the interpolation-free guess wins over the linear guess, and in some cases the opposite is true. As seen from Fig. 2, in two out of three cases all three approaches locate similar paths; however, in case **2-3** (column 3 of Fig. 2) positioning active replicas at well **3** gives a qualitatively different path. Interestingly, this path proceeds through a bifurcation region and its two transition states [see Fig. 2(a)(3)].

As we mentioned earlier and emphasized here again putting all the active replicas in one well is not the only option for the interpolation-free initial guess. Alternatively, a varying number of active replicas could be placed at both wells at the same time in different order. It is likely that for large molecules each different interpolation-free guess combination may yield a different path. However, in this paper we did not attempt to verify this proposal.

This section also demonstrates the pitfalls of the present LI method formulation pertaining to estimates of the transition state energies. In particular, in the case of **1-3** isomerization, which has the highest barrier, all the paths are nearly identical [Fig. 2(a)(2)] and overestimate the transition state (TS) barrier by as much as 2 kcal/mol. The exact same behavior was seen for the cyclic con-$\varphi\psi$ path in gas phase (data shown in supporting information[60]). This problem is due to the fact that a significant number of the replicas slide down into the lowest energy minimum forming kinks in the path and by doing so pull the chain of states away from the true transition state. This behavior is also known as corner cutting.[11,13] Because corner cutting is observed in the exact solution of the problem it must be intrinsic to the objective function.

Despite kinking and corner cutting, the present method serves the objective of locating approximate paths for com-

TABLE IV. Optimization summary for gas phase acyclic paths of the alanine dipeptide with linear and interpolation free initial guess. The data corresponds to No. of rep=80; $a=0.0625$; $K=16$.

Path	Guess	δ (Å)	ΣU (kcal/mol)	Walk (Å^3)	Objective function values	No. of iterations	No. of gradient evaluations
1-2	**line**	0.0712	−1255.77	15.37	−1240.40	1407	1157
	free1	0.0713	−1255.80	15.40	−1240.40	1861	1602
	free2	0.0713	−1255.79	15.39	−1240.40	2167	1751
1-3	**line**	0.112	−1077.67	60.53	−1017.14	2654	1717
	free1	0.113	−1083.17	64.62	−1018.55	1120	1007
	free2	0.113	−1077.78	60.54	−1017.24	3174	2066
2-3	**line**	0.127	−1039.76	78.43	−961.34	912	877
	free1	0.127	−1039.82	78.48	−961.34	920	854
	free2[a]	0.128	−975.15	74.44	−900.71	1175	1094

[a]Converged to a path different from that starting with the line approximation.

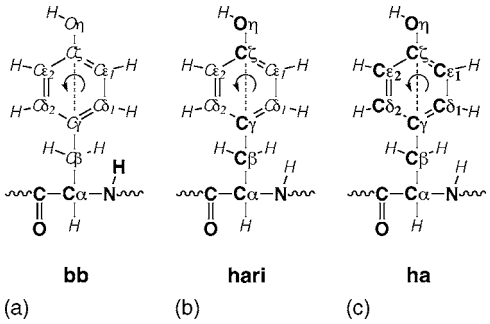

FIG. 2. Schematic representation of the **ha**, **hari**, and **bb** selection subspaces, using TYR residue with an internal rotation axis going through atoms Cγ and Cζ, as an example. The selected atoms are shown in bold.

plex systems that can later be refined with methods such as CPR, or else used as a starting point for such elaborate methods as transition path sampling (TPS).[70–76] The major advantage of the objective function method as formulated here is its ability to locate the approximate paths with interpolation-free guesses. All three examples of the present section demonstrate the viability and efficiency of this strategy as compared with linear interpolation in the Cartesian space. This advantage may become even greater in the systems with more degrees of freedom, as we will see in the following paragraphs.

C. β-hairpin folding paths

In this section we approach a much larger problem than conformational rearrangements of the alanine dipeptide. We apply the LI method to study folding mechanism of the second β hairpin of protein G, which is believed to be the key element in folding of the entire protein G. The isolated C-terminal β-hairpin (residues 40–56) is one of the smallest peptides known to form a stable β hairpin in solution. Therefore, it has been the focus of numerous experimental[77–81] and theoretical investigations.[82–102] In this section we seek reaction paths connecting the native and extended conformations of the β hairpin of protein G in water, as modeled with the GBORn implicit solvent model. We demonstrate how the reformulated LI method can be applied to obtain approximate reaction pathways for β-hairpin folding. We also demonstrate the advantage of having the interpolation-free initial guess options in addition to standard linear interpolation initial guess. Furthermore, we address the choice of a representative coordinate subspace in this section, using all backbone heavy atoms and amide hydrogen atoms (**bb**), all heavy atoms (**ha**), and only those heavy atoms whose positions remain unaltered after internal rotations by 180° that preserve overall sidechain conformation (**hari**) for the selection subspace. Thus, for the β hairpin, the **hari** subspace is derived from **ha** by removing Oδ1 and Oδ2 for the ASP, Oε1 and Oε2 for the GLU, and Cδ1, Cδ2, Cε1, and Cε2 atoms for both TYR and PHE residues. Figure 3 illustrates the three selections using tyrosine residue as an example.

1. Reactant and product setup

A model of the folded state of the β hairpin is obtained from the NMR structure of the full protein G (PDB code 2gb1). From this structure we extract the 16 residues (from 40 to 56) corresponding to the C-terminal β hairpin. The N terminus of the extracted peptide is capped with the neutral

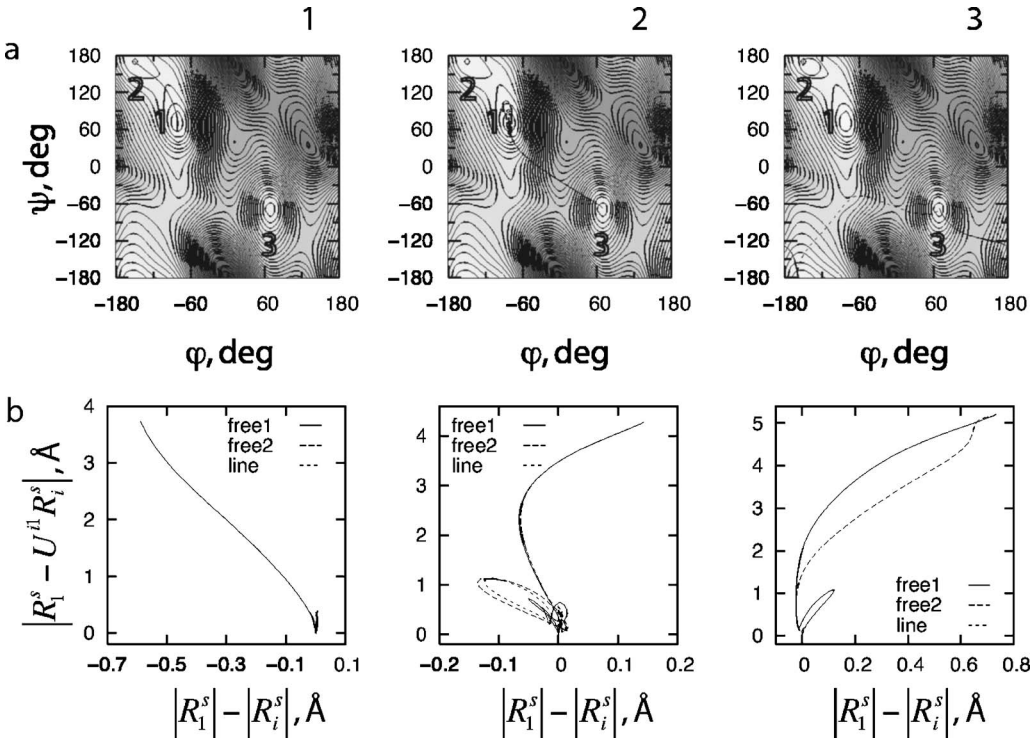

FIG. 3. Acyclic isomerization paths of the alanine dipeptide in the gas phase: comparison of the line interpolation and two interpolation-free guesses. Columns 1 through 3 correspond to paths **1-2**, **1-3**, and **2-3**, respectively. Rows are as follows: (a) projection of the paths on φ/ψ PES, in degrees; (b) projection of the paths in the representation (7) as described in the text $x_i=|R_1^s|-|U^{i1}R_i^s|$, $y_i=|R_1^s-U^{i1}R_i^s|$, in Å.

acetyl (ACE) group, leaving the native C terminus at residue 56 negatively charged. Care must be taken to make sure that the added ACE cap is in the proper *trans* conformation about its amide bond. The resulting β-hairpin model provides a charge distribution at the termini similar to that in the full protein G. Because of the neutral charge of the N terminus, our model prevents formation of the stable intertermini salt bridge, which could otherwise alter the behavior of the β hairpin from that in the context of protein G and in the dansylated β-hairpin alone.[79]

To obtain coordinates of the native and extended states of the β-hairpin model suitable for computational experiments, we utilize the following protocol. Starting with the extracted coordinates of the native state, initially, only coordinates of the ACE cap atoms are optimized, while keeping the rest of the coordinates in place with positional harmonic restraints. Next, the harmonic restraints on all the atoms are gradually relaxed providing the fully optimized native state coordinates. The extended state of the protein has been obtained by restraining the span between centers of mass of N and C terminal residues by means of harmonic distance restraints, while minimizing all other degrees of freedom. The span has been gradually increased in small increments starting from the native structure. At around 65 Å the energy increases sharply, revealing backbone strain of the fully extended peptide conformation. From such a strained extended conformation the geometry is relaxed by unconstrained minimization producing the final product coordinates.

2. Exploring the potential energy surface with LI method

Because of the size of the β-hairpin model and the long range of the conformational changes underlying folding, straightforward application of the LI method can be challenging. In particular, resolving very long paths in detail requires a large number of replicas, which may be computationally prohibitive. Therefore a "divide and conquer" approach has been devised to study folding at the most detailed level, yet with a limited number of replicas. As a compromise between resolution and computational costs, the maximum number of replicas used in any given computation does not exceed 150, which corresponds to a constrained optimization problem with a total of 71 485 variables. We note that in a recent study on β-hairpin folding as few as 40–80 minima were sufficient to produce a connected pathway between native and extended states.[100]

Initially for each of the three coordinate subspaces, namely **bb**, **ha** and **hari**, we ran exploratory 150-replica paths with $K=1$, and three different a values, specifically, 1.0, 0.5, and 0.25. For each set of parameters we employed three initial guess approaches, namely, linear interpolation (**line**) and two interpolation-free initial guesses (**free1** and **free2**), thus providing a total of 27 exploratory paths. In the case of **free1** all active replicas were placed in the native folded state well, whereas in the case of **free2** they were started in the extended state well. All of the exploratory paths were minimized with the multistage strategy described in Sec. IV A 2, converging 26 out of 27 attempted paths. The only failure at this stage corresponds to the path **free2** with $a=0.25$ in the **bb** representation.

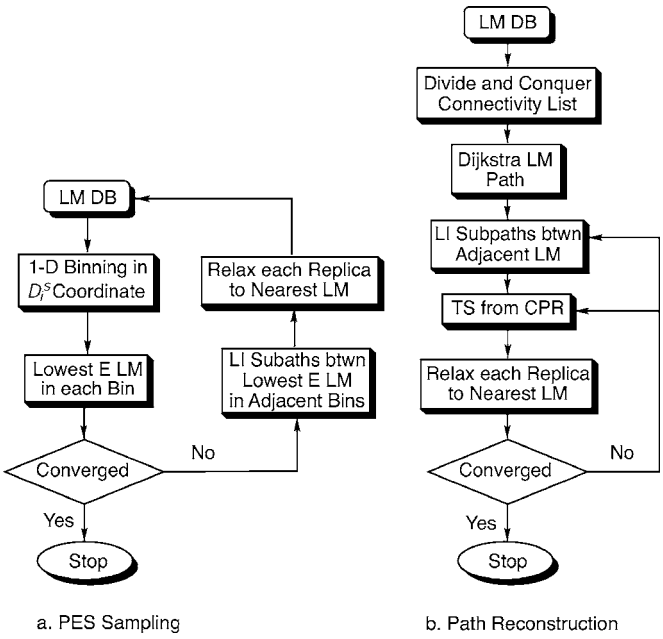

FIG. 4. Flow chart for exploration of the (a) PES and (b) path reconstruction. LM, TS, and DB stand for local minima(um), transition state and database, respectively.

Each of the three different path initialization methods locates a unique final path within the given subspace,[60] as suggested by the studies mentioned in Sec. IV B.

Paths with low resolution provide limited information if any about the folding landscape and are unlikely to sample the lowest energy path in such complex systems. To increase the resolution and explore the PES in greater detail we have attempted to refine the initial 26 paths following a procedure that is a hybrid of those devised by Wales and Elber.[7,100]

Our refinement procedure is summarized in Fig. 4(a) and is as follows. First, we minimize every active replica from the 26 successful exploratory paths to its nearest local minimum. Then we compute the progress coordinate based on representation (7), using the $s=$**hari** subspace for all the minima;

$$D_i^s = \sqrt{(|R_1^s| - |R_i^s|)^2 + |R_1^s - U^{i1}R_i^s|^2}. \tag{9}$$

All the local minima combined cover the range of 0.0–171.2 Å in the D_i^s coordinate. We split the resulting range into 32 equal bins and identify the lowest energy local minimum within each bin, thus forming a sequence of reactant and product pairs along the progress coordinate (9). Each consecutive pair is used to construct a new path, a subpath, for LI optimization in the **hari** selection subspace. The number of replicas for each such subpath is determined according to the distance between the end points in the **hari** subspace and assigned 3 replicas/Å density not to exceed 150 for any given path. For each such path we used one value of a (initially 0.25 and later 0.5) and three initial guesses, **line**, **free1** and **free2**. After the LI optimizations converged, each active replica was minimized and added to the original data set, containing previously identified local minima. This procedure, of spawning multiple subpaths with

up to 150 replicas, was repeated until convergence, meaning that for each bin in the D_i^s coordinate with s=**hari** no new local minimum with energy lower than already registered in this bin can be found (see supporting information[60] for detail). To further increase the resolution of the LI paths starting with iteration 4 we used 64 bins instead of 32 to spawn the subpaths. Although at each iteration in very few cases path optimizations could fail, in each section at least one attempt succeeds.

Upon convergence the sampling procedure provides a database of approximately 6780 unique (based on the root-mean-square displacement (RMSD) in the **hari** space and the total energy) points, each of which is assumed to represent a local minimum. The size of our database is modest compared to that reported recently for the same β hairpin, which comprised on the order of 25 000 local minima.[100] Therefore, by no means is our sampling procedure exhaustive. Furthermore, our sampling is clearly biased by the initial pathways. Despite this, we believe that we systematically covered a substantial area of the PES in the vicinity of the initial paths, which enables us to obtain high-resolution folding paths.

Because the progress coordinate (9) we use is isotropic in space, and thus only controls the distance from a given structure to the central reactant and not the direction, the final subpaths often connect fairly remote, in terms of the actual distance, structures. This prohibits simple assembly of the final lowest energy paths connecting native and extended configurations of the β hairpin. Therefore a strategy is needed to reconstruct reaction paths given the database of local minima identified through biased LI sampling. Below we describe the path reconstruction strategy without *a priori* knowledge of the underlying transition states. Note that if the transition states were known the path reconstruction would have been significantly simplified (see the following paragraphs).

3. Path reconstruction

To reconstruct reaction pathways from a data set of points spanning the configuration space between reactant and product, and without *a priori* knowledge of transition states interconnecting these points, we exploit the following ideas. First of all, any real transformation in the configuration space must map onto a sequence of adjacent local minima, separated by respective transition states.[103–110] Second, for local minima that have multiple nearest neighbors, the choice of transition is restricted by the Hammond postulate,[111] and Evans-Polanyi-Semenov rule,[112,113] of which the former has already been shown to apply to conformational rearrangements.[37] The Hammond postulate suggests that shorter hops that supposedly require less reorganization energy would be preferred over longer ones.[114] On the other hand, the Evans-Polanyi-Semenov rule allows one to further limit the choice of local minima by considering the energy change upon the hop. In particular, if the energy increase for the chosen transition is greater than a certain energy threshold the transition is disallowed. The energy change threshold can be chosen on the basis of the available thermal energy at which the transformation occurs. When the rate-limiting step of the rearrangement is known, the energy threshold should not exceed the corresponding activation energy. For the β-hairpin the estimated activation energy for the rate-limiting step is about 8 kcal/mol.[79,82] From these basic principles the minimum energy path can be reconstructed by locating a sequence of nearest-neighbor jumps of allowed energy between reactant and product configurations. Such a sequence might represent the reaction course over time as well. Locating the nearest-neighbor-hop sequence requires connectivity information between the local minima in the database.

a. Building connectivity information. To build sufficient connectivity information or a graph from the database of local minima we use a simple Euclidean distance metric. Because of the large number of unique data points in the database we did not attempt to compute a complete pairwise distance or adjacency matrix, which would have been costly. Instead we devised a two-step divide-and-conquer approach. The first step is exploratory and provides limited connectivity information at the lowest possible cost, while the second step exploits the obtained information to build the final sufficient adjacency list. The idea behind the divide-and-conquer approach is to partition the database of local minima on a mesh using some characteristic mapping, and then collect all distances between minima within each bin and across the adjacent bins. Clearly, this procedure is computationally less intensive than building complete adjacency matrix or graph of $K(K\text{-}1)/2$ distances, where K is the total number of minima in the database. Here we chose a 2D mapping procedure. In particular, we project all the data points, following RMSD best fitted to the reactant, onto the vector connecting the original reactant and product points and use the resulting parallel and transverse projections to perform 2D binning. Note that both best fit and projection were performed in the **hari** space.

In the first step of building an exploratory adjacency matrix we choose the smallest bin size that ensures that all occupied bins are connected by adjacency (each nonempty bin is connected to any other nonempty bin through a series of adjacent nonempty bins), i.e., to form a connected-bin graph. With the overall data set enclosed by a box of dimension $[-15.2\ \text{Å}, 152.2\ \text{Å}] \times [0.0\ \text{Å}, 71.5\ \text{Å}]$ ([parallel] $\times$[transverse]); the dimension of the bins has to be at least 6×6 Å for the occupied bins to be connected. The resulting mesh consists of 29 bins in the parallel and 13 bins in the transverse dimension, with the largest bin occupation of 305, and hence allows fast computation of the pair-wise distances (see supporting information[60] for additional details). We sweep across the bins collecting all intrabin distances for the reference bin (i,j) and all interbin distances between the bin (i,j) and its adjacent bins $(i,j+1)$, $(i+1,j)$, $(i+1,j-1)$ and, lastly, $(i+1,j+1)$. The pairwise distances are computed following RMSD best fit, all in the **hari** space. This procedure provides partial connectivity information, which we analyze to obtain the required bin dimensions for the final step.

The analysis of the connectivity information was performed with agglomerative clustering (AC), which exploits the Kruskal algorithm for finding minimum spanning tree (MST) of a connected graph.[115,116] AC verifies that the resulting graph is connected.[60] Furthermore, AC provides in-

formation on the key link, which we call the topological bottleneck, required for a basin of unfolded configurations to link to the folded state basin. In percolation theory language the topological bottleneck is equivalent to the percolation threshold. The percolation threshold provides a lower bound for the size of the bin to be used during the final construction of the adjacency list. Another useful parameter, available from the clustering, is the largest link in the MST. We can use the size of the largest link in the MST as the lower limit for the bond size allowed in the final adjacency list, which will substantially reduce the size of the final adjacency list without loss of critical information.

To find the topological bottleneck for folding we pick 20 local minima furthest from and nearest to the native structure, and using agglomerative clustering determine at what level these basins become interconnected. It turns out that potentially 400 different paths lie mainly on a common percolating cluster and therefore the largest link in this cluster, which is 19.6 Å long, corresponds to the topological bottleneck. The largest link in the cluster of the highest hierarchy is 37.3 Å. Thus, in the final round we used bins of size 20×20 Å and the maximum allowed link of size 40 Å. The resulting final adjacency list is used to reconstruct the folding pathway connecting the furthermost structure to the native state. Note that because we superimpose structures with the native state to build the projection, and then perform pairwise superposition when we compute the distances for the adjacency list, some structures in the 2D map appear further away then they actually are after superposition. This disagreement increases with the separation from the native structure. However, with the final mesh size 20×20 Å we feel confident to collect all critical links. With the connectivity information at hand we can now proceed to constructing a folding path.

b. Constructing a folding path. To construct the paths between the reactant and product starting from the adjacency list we used Dijkstra shortest path algorithm.[117] In the standard implementation of the Dijkstra algorithm the path is characterized by a cumulative cost, which is the sum of costs for all links comprising the path. For the folding problem, however, such cost definition is inappropriate since the shortest path would then comprise the longest possible links, missing a multitude of intermediate points, and thus will be no different than the problem we started with in the very beginning. Using the Hammond postulate[111–113] we model the folding process as jumping between the nearest local minima.[17,118–121] Hence we attempt to build a path of the shortest links.[122] Constructing the path out of the shortest links would also keep the number of yet unidentified intermediate structures between the connected points in the path at a minimum. To build such a path we can use the original Dijkstra single-source algorithm for finding the MST.[117] Using the structure furthest from the native as a source the Dijkstra algorithm finds the shortest-link paths to every other point, including the native state, in a single pass. For protein folding, however, we have introduced an energy test in the relaxation stage of the Dijkstra algorithm. This test prevents jumps that result in an energy increase higher than a set maximum. Without the energy test the forward and reverse paths would be identical. Also we would obtain exactly the same MST as that given by the hierarchical clustering. Adding the energy test causes the paths in forward and reverse directions to differ and also alters the appearance of the energy allowed MST. In fact, the database may fail to remain connected because some links would be prohibited by energy. Here we used an energy threshold of 5.0 kcal/mol to obtain the shortest-link path. For this pathway the largest link is 19.6 Å long in the **hari** subspace and is the same as the topological bottleneck identified earlier. The two structures corresponding to this link are depicted in the Figs. 5(e) and 5(f). Going through this link in the folding direction the energy decreases by 5.0 kcal/mol, and the side chain of TYR5 forms a hydrogen bond with the side chain of THR13, which brings the ends of the two strands closer together and causes the sidechain of the THR4 residue to rotate about the $C\alpha-C\beta$ bond by 95°. We will return to this link later when we refine the corresponding transition states.

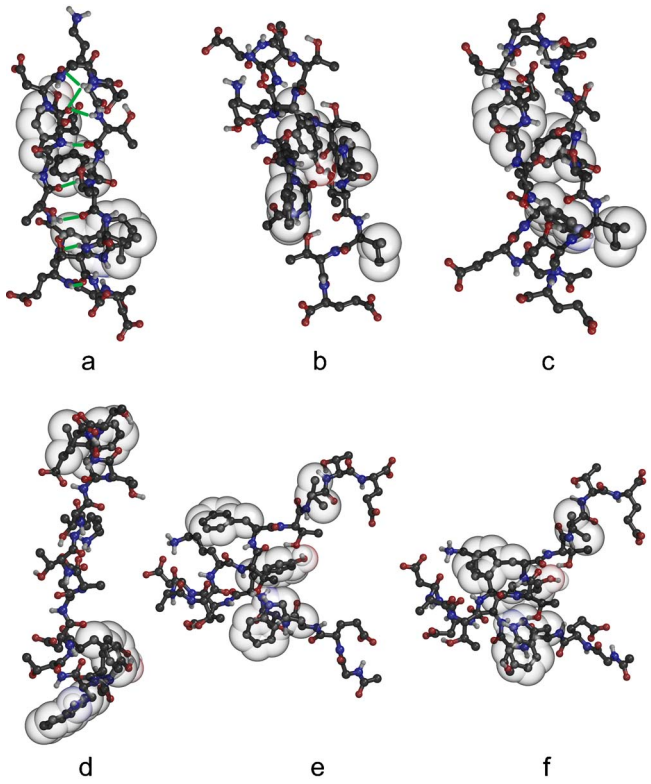

FIG. 5. (Color online) Representative peptide structures: (a) the optimized native β hairpin from protein G, (b) the most stable structure from sampled PES ($E=-1157.439$ kcal/mol), (c) the lowest energy structure in the basin near the native structure ($E=-1154.212$ kcal/mol) from the shortest-link path, (d) the extended structure, and [(e) and (f)] the two minima in the topological bottleneck (see text for description). In (a) the eight native hydrogen bonds are shown with solid lines. All atoms and bonds are shown in ball and stick. The side chains of the four hydrophobic residues are also shown with van der Waals spheres.

c. Final path refinement with LI and CPR methods. Now the folding path is defined as a sequence of local minima closest to each other in the **hari** space and such that at any link going in the folding direction the energy does not increase by more than 5.0 kcal/mol. There is only one shortest-link path defined by the Dijkstra algorithm, which consists of 481 individual local minima found in the **hari**

space. We use these minima to build 480 LI subpaths for the final refinement. Because at the final stage we intend to use CPR method, which in contrast to our LI method is defined in the complete Cartesian space, it is necessary to ensure that all possible mismatches in the order of atoms not in the **hari** space are removed before the final stage. Furthermore, it might be advantageous at this stage to construct LI pathways in the full space, since some transition states may occur exactly in the space perpendicular to **hari**, for example, rearrangements of the internal hydrogen bonds. Thus for each reactant-product pair in the final path we performed LI path optimizations in the **hari** and additionally in the complete (**all**) reaction spaces. In both cases we used $a=0.25$, $K=1$, and all three available guesses, the number of replicas was decided on the basis of distance in the **hari** space and a 5 replica/Å density with the minimum of three replicas in the path. Upon completion the LI paths were subjected to CPR refinement, which was in most cases very fast and straightforward. Having potentially six different LI paths increases the chance to locate the path with the lowest activation energy for any given pair. In some cases paths found in the **all** space have lower energies, but in other cases the **hari** space provides the lowest energy paths. In particular, the LI method in the **hari** space located 88, 89, and 69 lowest barrier paths with **line**, **free1** and **free2** initial guesses, respectively, while in the **all** space it located 59, 112, and 63 such paths for the same respective guesses.

Because most of the LI work was performed in the **hari** space at the final stages the LI paths in the **all** space as well as the final CPR runs locate a number of new intermediates, which differ by the configuration in the space orthogonal to **hari**. The presence of multiple intermediates along many of the subpaths creates some difficulties for CPR refinement. In particular, CPR sometimes fails to identify some of the intermediates and connecting transition states. Furthermore, because CPR is designed to pinpoint transition states[36,37,57–59] and not intermediates, one has to refine the energies of the intermediates to accurately compute corresponding transition state barriers. Therefore, to locate all intermediates and compute accurate transition states one has to minimize all the points in the CPR band to the nearest local minima using a gradually vanishing harmonic harness on the molecules to prevent sudden geometry changes during optimization. Following the minimization one needs to identify unique minima and then setup new CPR runs between consecutive minima. The information from the previous CPR run should be used at that stage to provide good initial guess for locating the TS between the new local minima if possible, otherwise new LI paths may be required [refer to the Fig. 4(b)].

The final CPR refinement provides 1680 nonzero barriers, with the highest barrier of 13.0 kcal/mol and the second highest barrier of 11.5 kcal/mol. Out of 1680 nonzero barriers in the final path 1557 barriers (92.6%) are below 5 kcal/mol. Note that the highest barrier for the topological bottleneck described above is 8.2 kcal/mol. Given the presence of even higher barriers in the path, it is unlikely that in the case of the β hairpin the topological bottleneck is related to the rate-limiting transition in the folding process. Despite

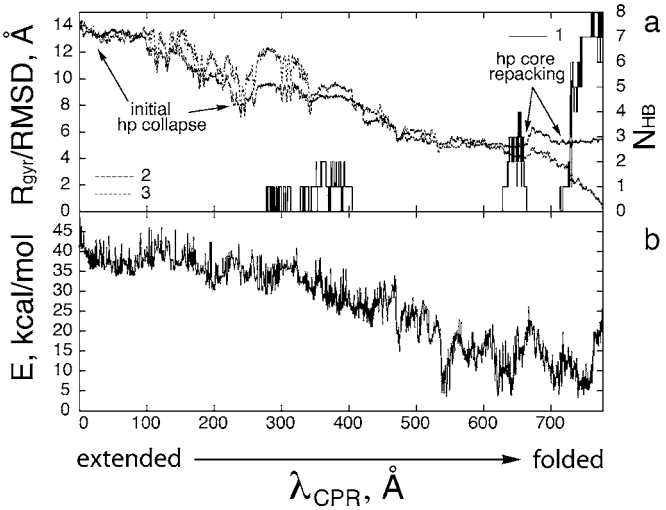

FIG. 6. Summary of the structural and energetic parameters for the CPR refined folding path. In (a) 1 represents the N_{HB} with the scale on the right y axis, 2 the radius of gyration for the hydrophobic core, and 3 the RMSD to the optimized structure of the native β hairpin from protein G; in (b) the solid line represents the energy profile along the path including all intermediates and transition states.

the substantial rate limiting activation energy, we attempted to look into the mechanism of folding to see whether the shortest-link path model can provide any insight into the folding process.

d. Folding mechanism. To discuss the folding mechanism based on the obtained path, we compute the following parameters along the CPR refined path [Fig. 6(a)]: RMSD to the native structure, radius of gyration of the hydrophobic core (residues TRP4, TYR6, PHE13, and VAL15) R_{gyr}^{hphb}, and the number of native hydrogen bonds N_{HB}. For RMSD we use the **hari** subspace. When computing the number of hydrogen bonds we consider a total of eight hydrogen bonds: two pairs of backbone hydrogen bonds between GLU3:THR16 and THR5:THR14, two bifurcated hydrogen bonds, namely, ASP7(O):THR12(H) and its partner ASP7(O):LYS11(H), and lastly a lone bond ASP8(O):LYS11(H). The hydrogen bond is said to exist if the H···O distance is less then or equal to 2.5 Å. In addition we plot the energy profile obtained after the CPR refinement [Fig. 6(b)].

From the Fig. 6(b) it becomes clear that the native configuration of the β hairpin from protein G (the rightmost point), depicted in Fig. 5(a), has relatively high energy. However, it has been found from NMR experiments that the β hairpin in the context of protein G and by itself in solution have somewhat different structures due to the lack of certain stabilizing interactions provided by the protein in the free hairpin.[78,79] Thus the native configuration is expected to have high energy. Furthermore, the presence of several rugged wells with comparable energies in the region between 500 and 800 Å of the cumulative CPR coordinate λ_{CPR} [Fig. 6(b)] suggests that a number of conformations of the peptide can coexist in solution. From experiment the estimated population of the β hairpin is ca. 30%, which is consistent with coexistence of other configurations.[78,79,81]

Our mechanism agrees with the mechanisms inferred

from a number of simulations in explicit[88,92,93,102] and implicit water,[87] as derived from free energy surfaces and transition path ensembles.[88,92,93,102] In particular, it follows from Figs. 6(a) and 6(b) that the folding of the β hairpin proceeds via initial hydrophobic collapse, as judged from $R_{\mathrm{gyr}}^{\mathrm{hphb}}$, where a precursor to a hydrophobic intermediate is formed (with $R_{\mathrm{gyr}}^{\mathrm{hphb}}$ about 9.5 Å), which also registers a couple of native hydrogen bonds. The collapse continues further, breaking the registered native hydrogen bonds, and then plateaus at about 5.0–5.5 Å $R_{\mathrm{gyr}}^{\mathrm{hphb}}$. This flat $R_{\mathrm{gyr}}^{\mathrm{hphb}}$ area between 500 and 670 Å in λ_{CPR} contains about three relatively deep energy basins, which might correspond to distinct intermediates. The two furthermost basins form up to four native hydrogen bonds and have $R_{\mathrm{gyr}}^{\mathrm{hphb}}$ of about 5.0 Å. Before going into the native configuration the hydrophobic residues unpack increasing the $R_{\mathrm{gyr}}^{\mathrm{hphb}}$ to about 6.3 Å and breaking all four of the native hydrogen bonds, an event that also coincides with a hump on the potential energy profile, and then finally repack reducing the value of $R_{\mathrm{gyr}}^{\mathrm{hphb}}$ to about 5.4 Å concomitant with registering the hydrogen bonds to reach the final maximum value of eight. The region following the barrier, namely, between 690 and 800 Å in λ_{CPR}, has two basins. The first basin registers only up to one native hydrogen bond, whereas the second basin has between five and eight native hydrogen bonds. A representative structure from the latter basin is shown in Fig. 5(c). Interestingly, the shortest-link path between the extended and native states of the hairpin does not visit the lowest energy states identified during PES exploration. The absolute lowest structure is illustrated in Fig. 5(b). It is likely that these lowest energy states do not belong to the percolating cluster; however, we did not attempt to check if such states were kinetically accessible to interfere with the folding process.

V. CONCLUSIONS

In summary, we have presented a modified line integral method with self-avoiding walk for searching approximate reaction pathways in large molecules. Use of the state-of-the-art nonlinear constrained optimization procedure with the L-BFGS Hessian update renders the new path optimization robust and additionally provides a superlinear convergence rate. Moreover the method no longer requires an initial guess, or more precisely it can exploit an interpolation free initial guess, which makes it particularly useful for studying complex molecular rearrangements. The accurate reaction barriers from the paths can be trivially obtained with the use of the complementary conjugate peak refinement method.[36,37] The method has been applied to conformational rearrangements of alanine dipeptide in gas phase and in water, as well as folding of the β hairpin of protein G in water. In the latter case a procedure was developed for limited systematic sampling of the potential energy surface underlying folding and reconstruction of the folding pathways within the nearest-neighbor hopping approximation. The mechanism for the β-hairpin folding derived with this procedure agrees well with the best simulations available in the literature.[87,88,92,93,102]

ACKNOWLEDGMENTS

The authors would like to acknowledge the NIH for support of this work (GM48807). The authors are grateful to Dr. R. Waltz for his assistance with the KNITRO code.

[1] A. Komornicki, K. Ishida, K. Morokuma, R. Ditchfield, and M. Conrad, Chem. Phys. Lett. **45**, 595 (1977).
[2] T. A. Halgren and W. N. Lipscomb, Chem. Phys. Lett. **49**, 225 (1977).
[3] C. J. Cerjan and W. H. Miller, J. Chem. Phys. **75**, 2800 (1981).
[4] H. B. Schlegel, J. Comput. Chem. **3**, 214 (1982).
[5] S. Bell and J. S. Crighton, J. Chem. Phys. **80**, 2464 (1984).
[6] R. Czerminski and R. Elber, Proc. Natl. Acad. Sci. U.S.A. **86**, 6963 (1989).
[7] R. Czerminski and R. Elber, J. Chem. Phys. **92**, 5580 (1990).
[8] C. Guilbert, D. Perahia, and L. Mouawad, Comput. Phys. Commun. **91**, 263 (1995).
[9] C. Choi and R. Elber, J. Chem. Phys. **94**, 751 (1991).
[10] A. Ulitsky and R. Elber, J. Chem. Phys. **92**, 1510 (1990).
[11] G. Henkelman and H. Jonsson, J. Chem. Phys. **113**, 9978 (2000).
[12] B. Peters, A. Heyden, A. T. Bell, and A. Chakraborty, J. Chem. Phys. **120**, 7877 (2004).
[13] S. A. Trygubenko and D. J. Wales, J. Chem. Phys. **120**, 2082 (2004).
[14] J.-W. Chu, B. L. Trout, and B. R. Brooks, J. Chem. Phys. **119**, 12708 (2003).
[15] R. E. Gillilan and K. R. Wilson, J. Chem. Phys. **97**, 1757 (1992).
[16] A. E. Cho, J. D. Doll, and D. L. Freeman, Chem. Phys. Lett. **229**, 218 (1994).
[17] M. Berkowitz, J. D. Morgan, J. A. McCammon, and S. H. Northrup, J. Chem. Phys. **79**, 5563 (1983).
[18] R. Olender and R. Elber, J. Chem. Phys. **105**, 9299 (1996).
[19] V. Zaloj and R. Elber, Comput. Phys. Commun. **128**, 118 (2000).
[20] D. V. Berkov, J. Magn. Magn. Mater. **186**, 199 (1998).
[21] R. Elber, J. Meller, and R. Olender, J. Phys. Chem. B **103**, 899 (1999).
[22] D. Passerone and M. Parrinello, Phys. Rev. Lett. **87**, 108302 (2001).
[23] D. Passerone, M. Ceccarelli, and M. Parrinello, J. Chem. Phys. **118**, 2025 (2003).
[24] R. Elber and M. Karplus, Chem. Phys. Lett. **139**, 375 (1987).
[25] S. Huo and J. E. Straub, J. Chem. Phys. **107**, 5000 (1997).
[26] S. Huo and J. E. Straub, Proteins **36**, 249 (1999).
[27] R. Czerminski and R. Elber, Int. J. Quantum Chem. **24**, 167 (1990).
[28] H. L. Woodcock, M. Hodoscek, P. Sherwood, Y. S. Lee, H. F. Schaefer III, and B. R. Brooks, Theor. Chem. Acc. **109**, 140 (2003).
[29] R. Olender and R. Elber, J. Mol. Struct.: THEOCHEM **398–399**, 63 (1997).
[30] G. Domotor, L. Stacho, and M. I. Ban, J. Mol. Struct.: THEOCHEM **501–502**, 509 (2000).
[31] H. F. Schaefer III, Chem. Br. **11**, 227 (1975).
[32] K. Fukui, S. Kato, and H. Fujimoto, J. Am. Chem. Soc. **97**, 1 (1975).
[33] J. Nocedal, Math. Comput. **35**, 773 (1980).
[34] J. Nocedal, Math. Program. **45**, 503 (1989).
[35] S. S.-L. Chiu, J. J. W. McDouall, and I. H. Hillier, J. Chem. Soc., Faraday Trans. **90**, 1575 (1994).
[36] S. Fischer and M. Karplus, Chem. Phys. Lett. **194**, 252 (1992).
[37] S. Fischer, P. D. J. Grootenhuis, L. C. Groenen, W. P. van Hoorn, F. C. J. M. van Veggel, D. N. Reinhoudt, and M. Karplus, J. Am. Chem. Soc. **117**, 1611 (1995).
[38] L. Stacho, G. Domotor, and M. I. Ban, J. Math. Chem. **29**, 169 (2001).
[39] L. Stacho, G. Domotor, and M. I. Ban, Chem. Phys. Lett. **311**, 328 (1999).
[40] L. Stacho, G. Domotor, and M. I. Ban, J. Math. Chem. **26**, 87 (1999).
[41] R. Elber and M. Karplus, Chem. Phys. Lett. **311**, 335 (1999).
[42] R. Fletcher, *Practical Methods of Optimization* (Wiley, New York, 1981).
[43] M. Lalee, J. Nocedal, and T. Plantega, SIAM J. Optim. **8**, 682 (1998).
[44] R. H. Byrd, M. E. Hribar, and J. Nocedal, SIAM J. Optim. **9**, 877 (1999).
[45] R. H. Byrd, J. C. Gilbert, and J. Nocedal, Math. Program. **89**, 149 (2000).
[46] C. Eckart, Phys. Rev. **47**, 552 (1935).
[47] A. Y. Dymarsky and K. N. Kudin, J. Chem. Phys. **122**, 124103/1 (2005).
[48] A. Y. Dymarsky and K. N. Kudin, J. Chem. Phys. **122**, 227102/1 (2005).
[49] K. N. Kudin and A. Y. Dymarsky, J. Chem. Phys. **122**, 224105/1 (2005).
[50] W. Kabsch, Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. **32**, 922 (1976).
[51] R. H. Byrd, J. Nocedal, and R. A. Waltz, *KNITRO: An Integrated Pack-*

age for Nonlinear Optimization. (Kluwer Academic, Erice, 2004).
[52] M. Karplus, CHARMM Harvard University, Boston, MA, 2004.
[53] B. N. Dominy and C. L. Brooks III, J. Phys. Chem. B **103**, 3765 (1999).
[54] B. N. Dominy, Mol. Simul. **24**, 259 (2000).
[55] M. S. Lee, M. Feig, F. R. Salsbury, Jr., and C. L. Brooks III, J. Comput. Chem. **24**, 1348 (2003).
[56] M. S. Lee, F. R. Salsbury, Jr., and C. L. Brooks III, J. Chem. Phys. **116**, 10606 (2004).
[57] A.-N. Bondar, M. Elstner, S. Suhai, J. C. Smith, and S. Fischer, Structure (London) **12**, 1281 (2004).
[58] A. D. Gruia, A.-N. Bondar, J. C. Smith, and S. Fischer, Structure (London) **13**, 617 (2005).
[59] R. Dutzler, T. Schirmer, M. Karplus, and S. Fischer, Structure (London) **10**, 1273 (2002).
[60] See EPAPS Document No. E-JCPSA6-124-505617 for Figs. S1–S10. This document can be reached via a direct link in the online article's HTML reference section or via the EPAPS homepage (http://www.aip.org/pubservs/epaps.html).
[61] M. Feig, A. Onufriev, M. S. Lee, W. Im, D. A. Case, and C. L. Brooks III, J. Comput. Chem. **25**, 265 (2004).
[62] M. Feig, A. D. MacKerell, Jr., and C. L. Brooks III, J. Phys. Chem. B **107**, 2831 (2003).
[63] A. D. MacKerell, Jr., M. Feig, and C. L. Brooks III, J. Am. Chem. Soc. **126**, 698 (2003).
[64] C. M. Cortis, J.-M. Langlois, M. D. Beachy, and R. A. Friesner, J. Chem. Phys. **105**, 5472 (1996).
[65] I. R. Gould and I. H. Hillier, J. Chem. Soc., Chem. Commun. **1993**, 951.
[66] I. R. Gould, W. D. Cornell, and I. H. Hillier, J. Am. Chem. Soc. **116**, 9250 (1994).
[67] P. E. Smith, J. Chem. Phys. **111**, 5568 (1999).
[68] C. L. Brooks III (unpublished).
[69] J. Apostolakis, P. Ferrara, and A. Caflisch, J. Chem. Phys. **110**, 2099 (1999).
[70] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, J. Chem. Phys. **108**, 1964 (1998).
[71] C. Dellago, P. G. Bolhuis, and D. Chandler, J. Chem. Phys. **108**, 9236 (1998).
[72] P. G. Bolhuis, C. Dellago, and D. Chandler, Faraday Discuss. **110**, 421 (1998).
[73] C. Dellago, P. G. Bolhuis, and D. Chandler, J. Chem. Phys. **110**, 6617 (1999).
[74] P. L. Geissler, C. Dellago, and D. Chandler, J. Phys. Chem. B **103**, 3706 (1999).
[75] P. G. Bolhuis, C. Dellago, and D. Chandler, Proc. Natl. Acad. Sci. U.S.A. **97**, 5877 (2000).
[76] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).
[77] A. M. Gronenborn, D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield, and G. M. Clore, Science **253**, 657 (1991).
[78] F. J. Blanco, G. Rivas, and L. Serrano, Nat. Struct. Biol. **1**, 584 (1994).
[79] V. Munoz, P. A. Thompson, J. Hofrichter, and W. A. Eaton, Nature (London) **390**, 196 (1997).
[80] A. G. Cochran, N. J. Skelton, and M. A. Starovasnik, Proc. Natl. Acad. Sci. U.S.A. **98**, 5578 (2001).
[81] R. M. Fesinmeyer, F. M. Hudson, and N. H. Andersen, J. Am. Chem. Soc. **126**, 7238 (2004).
[82] V. Munoz, E. R. Henry, J. Hofrichter, and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. **95**, 5872 (1998).
[83] V. S. Pande and D. S. Rokhsar, Proc. Natl. Acad. Sci. U.S.A. **96**, 9062 (1999).
[84] A. R. Dinner, T. Lazaridis, and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **96**, 9068 (1999).
[85] A. Kolinski, B. lkowski, and J. Skolnick, Biophys. J. **77**, 2942 (1999).
[86] D. K. Klimov and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **97**, 2544 (2000).
[87] B. Zagrovic, E. J. Sorin, and V. S. Pande, J. Mol. Biol. **313**, 151 (2001).
[88] R. Zhou, B. J. Berne, and R. Germain, Proc. Natl. Acad. Sci. U.S.A. **98**, 14931 (2001).
[89] J. Lee and S. Shin, Biophys. J. **81**, 2507 (2001).
[90] R. Zhou and B. J. Berne, Proc. Natl. Acad. Sci. U.S.A. **99**, 12777 (2002).
[91] B. Ma and R. Nussinov, Protein Sci. **12**, 1882 (2003).
[92] R. Zhou, Proteins **53**, 148 (2003).
[93] P. G. Bolhuis, Proc. Natl. Acad. Sci. U.S.A. **100**, 12129 (2003).
[94] G. Wei, P. Derreumaux, and N. Mousseau, J. Chem. Phys. **119**, 6403 (2003).
[95] A. K. Felts, Y. Harano, E. Gallicchio, and R. M. Levy, Proteins **56**, 310 (2004).
[96] R. Zhou, G. Krilov, and B. J. Berne, J. Phys. Chem. B **108**, 7528 (2004).
[97] W. C. Swope, J. W. Pitera, F. Suits *et al.* J. Phys. Chem. B **108**, 6582 (2004).
[98] S. Brown and T. Head-Gordon, Protein Sci. **13**, 958 (2004).
[99] S. Y. Lee, Y. Fujitsuka, D. H. Kim, and S. Takada, Proteins **55**, 128 (2004).
[100] D. A. Evans and D. J. Wales, J. Chem. Phys. **121**, 1080 (2004).
[101] M. Andrec, A. K. Felts, E. Gallicchio, and R. M. Levy, Proc. Natl. Acad. Sci. U.S.A. **102**, 6801 (2005).
[102] P. G. Bolhuis, Biophys. J. **88**, 50 (2005).
[103] R. E. Kunz and R. S. Berry, J. Chem. Phys. **103**, 1904 (1995).
[104] F. H. Stillinger and T. Weber, Science **225**, 983 (1984).
[105] O. M. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).
[106] F. H. Stillinger, Science **267**, 1935 (1995).
[107] J. M. Troyer and F. E. Cohen, Proteins **23**, 97 (1995).
[108] R. Elber and M. Karplus, Science **235**, 318 (1987).
[109] B. W. Church and D. Shalloway, Proc. Natl. Acad. Sci. U.S.A. **98**, 6098 (2001).
[110] G. T. Barkema and N. Mousseau, Phys. Rev. Lett. **77**, 4358 (1996).
[111] G. S. Hammond, J. Am. Chem. Soc. **77**, 334 (1955).
[112] M. G. Evans and M. Polanyi, Trans. Faraday Soc. **34**, 11 (1938).
[113] N. N. Semenov, *Some Problems of Chemical Kinetics ad Reactivity (in Russian)*, 1st ed. (Academic Science USSR, Moscow, 1954).
[114] Note, however, that there are exceptions to that rule, and high barriers may still separate structures that are close spatially. We discovered that one such exception is rotation about amide bond at the N terminus, which has high activation barrier (10–16 kcal/mol), but little effect on the distance, merely swapping two atoms. However, this appears to be an artifact of the model, and has nothing to do with folding.
[115] J. B. Kruskal, Jr., Proc. Am. Math. Soc. **7**, 48 (1956).
[116] T. Kurita, Pattern Recogn. **24**, 205 (1991).
[117] E. W. Dijkstra, Numer. Math. **1**, 269 (1959).
[118] S. H. Northrup and J. A. McCammon, J. Chem. Phys. **78**, 987 (1983).
[119] J. T. Bartis and B. Widom, J. Chem. Phys. **60**, 3474 (1974).
[120] B. Widom, J. Chem. Phys. **55**, 44 (1971).
[121] B. Widom, J. Chem. Phys. **61**, 672 (1974).
[122] Provided with a knowledge of transition states, connected intermediates and their energies, the final path can be constructed by simply using forward and reverse barrier heights as the cost in the shortest-link path.