Energy landscapes of conformationally constrained peptides

Yaakov Levy and Oren M. Becker[a)]
Department of Chemical Physics, School of Chemistry, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel

(Received 3 July 2000; accepted 9 October 2000)

Conformation constraints are known to affect the flexibility and bioactivity of peptides. In this study we analyzed the effect of conformation constraints on the topography of the energy landscapes of three analogous hexapeptides. The three analogs vary in the degree of constraint imposed on their conformational motion: linear alanine hexapeptide with neutral terminals (Ala6), linear alanine hexapeptide with charged terminals (chrg-Ala6), and cyclic alanine hexapeptide (cyc-Ala6). It was found that significantly different energy landscapes characterize each of the three peptides, leading to different folding behaviors. Since all three analogs would be encoded by the same gene, these results suggest that nongenomic post-translational modifications may play an important role in determining the properties of proteins as well as of their folding pathways. In addition, the present study indicates that the complexity of those energy landscapes that are dominated by funnel topography can be captured by one or two reaction coordinates, such as conformational similarity to the native state. However, for more complex landscapes characterized by multiple basins such a description is insufficient. This study also shows that similar views of the landscape topography were obtained by principal component analysis (based only on local minima) and by topological mapping analysis (based on minima and barrier information). Both methods were able to resolve the complex landscape topographies for all three peptides.
[DOI: 10.1063/1.1329646]

I. INTRODUCTION

Peptides and proteins are complex molecular systems with distinct stable three-dimensional structures, which are the key for understanding their biological function. These structures are determined by the underlying energy landscape, which in turn is a consequence of the molecule's composition. Therefore, the study of protein energy landscapes in the context of protein folding has long been a central topic of both theoretical and experimental investigations.[1] Experimental evidence concerning the complexity of protein energy landscapes has been obtained, for example, through the observation of a multiplicity of relaxation times, the various intermediate species, the nonexponential kinetic, and the non-Arrhenius behavior.[2] Most notable are the extensive studies of the kinetics of CO binding to myoglobin performed by Frauenfelder and collaborators.[3–5] The obtained results have been explained in terms of a hierarchy of minima that are thought to be arranged in tiers corresponding to their energy levels.

Theoretically, polypeptide energy landscapes have been studied using both simplified models and all-atom simulations, with the resulting landscapes being characterized in a variety of ways using order parameters, geometrical measures, and topological connectivity patterns. Simplified models, originally on-lattice and more recently off-lattice, have been extensively used to study the energy landscape of model proteinlike systems. In particular, in many lattice studies it was found that a single variable, which is defined as the ''fraction of native contacts'' Q, can be used as a reaction coordinate that well describes the folding process.[1] This Q variable, which has a value near zero for the highly denatured conformation and reaches unity for the native state, describes the progress of the folding reactions in models such as the 27-mer model on a cubic lattice.[6,7] However, there are cases where a single progress variable such as Q is not sufficient to distinguish trajectories that fold directly to the native state from those that go through intermediate traps.[8] For example, for a 125-residue model more than one progress coordinate was required to describe the folding process.[9] The coordinates that were found suitable to describe the folding process in that system monitor the formation of the core and the trapping of the chain in long-lived intermediate states. In cases of complex energy landscapes it was found that to study protein folding a kinetic reaction coordinate was more useful than a thermodynamic reaction coordinate (such as the above Q variable).[8,10,11]

Recently the study of polypeptide energy landscapes has shifted towards more detailed atomistic simulations. However, unlike simplified models in which extensive enumeration of all possible states is possible, such full enumeration is not practical for most atomistic polypeptide models (except for the smallest peptides). The reason is that similar to the real system the atomistic models exhibit an extremely large number of local minima (locally stable conformations) even in the vicinity of the native structure.[12] Therefore, sampling techniques must be employed to study the energy landscapes underlying atomistic models of even relatively small molecu-

[a)]Author to whom correspondence should be addressed. Current address: Bio-Information Technologies (Bio-I.T.) LTD., Israel. Electronic mail: becker@sapphire.tau.ac.il

lar systems. Atomistic simulations that focus on the analysis of energy landscapes vary both in sampling strategies and in analysis techniques. For example, Sheinerman and Brooks[13,14] have used extensive all-atom simulations to sample and to characterize the energy landscape of two small proteins. The underlying landscape was then described in terms of two order parameters: one quantifying the collapse of the protein and the other reflecting similarity to the native state (equivalent to the *Q*-order parameter used in simplified model studies). Becker and Karplus[15] and Levy and Becker[16] used a topological analysis to generate disconnectivity tree-constructs that reflect the overall topography of the energy landscapes of several peptides, highlighting basin connectivity (this method was later applied in a similar way to atomic and molecular clusters as well[17,18]). Berry and collaborators[19–23] constructed specific connectivity pathways to characterize the energy landscape of several clusters, distinguishing between good and bad structure seekers. Finally, the principal component analysis was used by several groups to visualize both molecular dynamics trajectories[24–27] and energy landscapes of peptides and proteins.[28–32]

Related to the general question of protein folding are numerous examples in which seemingly small changes in the chemical composition of a molecular system result in large structural and/or functional changes. Examples are the onset of a disease due to a single point mutation (e.g., prion diseases[33]), metabolic pathway activation following the binding of a small ligand to a receptor, and changes in the molecular bioactivity as a result of seemingly small conformational constraints. Therefore, the mechanism by which small molecular modifications, such as point mutations or geometric constraints, affect the biological properties of polypeptides is of interest for chemists and biologists alike. Since the energy landscape underlies the structural, thermodynamic, and kinetic properties of any molecule, this question can be reformulated in terms of the energy landscape theory: ''How do small molecular modifications affect the energy landscape of peptides and proteins?'' A preliminary study by Levy and Becker[16] has indicated that the overall topography of the energy landscape changes as a result of a conformation constraint. Another study by Becker *et al.*[34] showed that quantitative analysis of conformation space can be directly correlated with bioactivity of therapeutic peptides.

The goal of the present study is to characterize the effect of conformation constraints on the topography of molecular energy landscapes. To address this question the atomic-level energy landscapes of three alanine–hexapeptide analogs were reconstructed and analyzed. The three peptides studied were the linear $(Ala)_6$, the linear $(Ala)_6$ with opposite charges at the *C*- and *N*-termini, and the backbone cyclized $(Ala)_6$. Several different analysis techniques were used to highlight various aspects of the landscapes: connectivity, topography and order parameters. It has been shown that a consistent and comprehensive characterization of these energy landscapes can be obtained by integrating the different views.

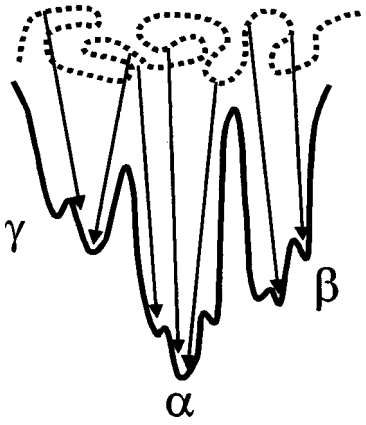

FIG. 1. Schematic illustration of conformation sampling of a molecular potential energy surface. Conformations are sampled at high temperatures where barrier crossing can easily occur and are then minimized to the nearest local minimum or slowly annealed back to room temperature before being minimized to the nearest local minimum.

II. MAPPING ENERGY LANDSCAPES

The notion of energy landscapes has an established contribution to the understanding of the reactions dynamics of small molecules.[35] In recent years the energy landscape concept, especially that of the proposed funnel topography, has become more prevalent in the discussion of protein folding, raising a controversy regarding its actual contribution, as most protein folding experiments can be described by simple two- or three-state models.[1,8,36] It should be noted, however, that the term energy landscape is sometimes used for the multidimensional potential energy surface that underlies the molecule's conformation space, and sometimes for its free energy profile. In the present study the focus is on the potential energy surface. In general, potential energy landscapes can be studied on two detailed levels. On the first detailed level the energy landscape is characterized solely by the set of locally stable conformations, i.e., by the local minima only. This approach was introduced by Stillinger and Weber[37] and later applied by many researchers.[12–14,24,25,28,31,34,38–40] The basic idea is to collect a large sample of conformations, minimize each of them to the nearest locally stable minimum and use these local minima to characterize the landscape. Figure 1 schematically represents a process by which sampled conformations are quenched to their nearest local minimum. Based on these sampled local minima the energy landscape underlying the molecule can be reconstructed. While such reconstructions can shed light on the topography of the energy landscape (basins, roughness, etc.) they miss important ingredients—the barriers or saddle points that separate the individual minima or basins. For example, in the landscape depicted in Fig. 2(a) basin β has lower energy compared to that of basin γ, but the transition from basin β to the native basin α is slower than the transition from basin γ to basin α, due to a higher barrier separating them. This important feature, which governs the system kinetics, is not reflected in the minima based picture of the energy landscape [Fig. 2(a)].

On the second more detailed level of description one may add barriers to the characterization of the energy land-

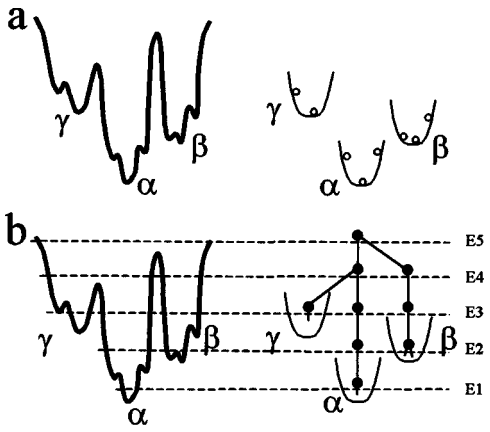

FIG. 2. (a) Illustration of the view of the energy landscape that can be obtained from information regarding the local minima only (right) compared with the full energy landscape (left). (b) Illustration of the much richer picture of the same energy landscape that can be obtained by information regarding basin connectivity added to the previous minima-only based view.

scape. Such a detailed representation, however, is computationally much more demanding. Several researchers have followed this type of energy landscape analysis for small peptides and clusters.[15–17,19–21,29,41,42] Figure 2(b) illustrates the resulting view of the energy landscape. The information retrieved this way not only reveals that basin β is less accessible than basin γ, due to the higher energy barrier separating it from the global minimum α, but it also contains all the information necessary for a complete reconstruction of the system's kinetics by means of the master equation.[15,22,41] Recently Becker has combined the two mapping approaches, integrating data from local minima as well as from barriers, to obtain a detailed energy landscape for a derivative of the alanine tetrapeptide.[29]

Clearly, not all energy landscapes will exhibit a significant difference between the view obtained when using local minima only and the view obtained via the analysis of barrier connectivity (as suggested by Fig. 2). There are likely to be many systems in which identical pictures will arise from both analyses approaches. However, at present there is not enough knowledge to predict *a priori* which energy landscape will show a discrepancy between the two viewpoints and which will not. In this work the issue of multiple viewpoints of the same potential energy landscape is explored.

III. METHODS

A. Conformation sampling

As a full enumeration of all the possible conformations of a hexapeptide is impractical, a sampling procedure must be applied in order to generate a representative sample of the molecule's conformation space. Many methods are available for sampling molecular conformations, each harboring advantages and limitations. They should therefore be applied according to their suitability to the problem at hand. The sampling procedure adopted for the present study stems from the tendency to get as broad a view as possible of the molecular energy landscape accessible to the molecule at physiological temperatures. To accomplish this goal a two-step sampling procedure was applied.[16,31,34] First, conformations are sampled from a high temperature molecular dynamics trajectory at 1000 K. Then each of the sampled high temperature conformations is gradually annealed down to 300 K (using molecular dynamics) before being quenched by direct minimization. The annealed and minimized conformations constitute the conformation sample of that molecule. The gradual annealing guarantees that the resulting conformations will indeed be on the 300 K manifold (i.e., are accessible at 300 K), while the high temperature sampling allows us to cross high-energy barriers and sample broad regions of conformation space. The choice of the 1000 K sampling temperature was based on earlier studies, which indicated that the undesirable *cis–trans* transitions of the peptide bond occur at higher sampling temperatures.[43]

Technically, each sampling procedure starts with a 500 ps molecular dynamics trajectory at 1000 K (simulated using 2 fs timesteps). Conformations are sampled along the high temperature trajectory every 1 ps, resulting in a total of 500 conformations. Short molecular dynamic trajectories (simulated at 1 fs timesteps) are then applied to cool each of the high temperature conformations down to 300 K (temperature decreases at 100 K steps). Following the cooling phase each structure is minimized by a combined protocol consisting of 200 steepest decent steps followed by adopted basis Newton–Raphson (ABNR) minimization until a total gradient of 0.01 is reached. The representation of the molecular dynamics and the various energy calculations were performed with the CHARMM program[44] and the CHARMM all atom forcefield.[45] No explicit water molecules were included, no energy cutoffs were applied and a distance dependent dielectric constant was used.

Conformation samples constructed in the above way are likely to include some similar conformations. While these may be important for reflecting conformational probability distributions, they are redundant as far as the energy landscape's topography and topology are concerned. Therefore, in the present study the conformation samples are pruned, removing conformations exhibiting high similarity. As will be discussed below, this pruning allows for a more efficient landscape analysis without affecting the results.

B. Molecular systems

Three alanine hexapeptide analogs were studied. (i) Ala6–alanine hexapeptide with neutral terminal groups, (ii) chrg-Ala6–alanine hexapeptide with a positive charge at the *N*-terminus and a negative charge at the *C*-terminus, (iii) cyc-Ala6—a backbone cyclized alanine hexapeptide. The initial conformations used in the sampling process of Ala6 and chrg-Ala6 were the fully extended conformations. The initial conformation for cyc-Ala6 was an extended conformation that was backbone cyclized and minimized until it assumed a reasonable starting cyclic conformation.

This set of three alanine hexapeptide analogs was selected because of the dramatic difference in the degree of conformational constraints imposed on their internal flexibility. In particular, these molecules span a broad spectrum ranging from a completely unconstrained analog, linear Ala6, to a maximally constrained analog, cyclic Ala6, in

which the covalent bond between the two terminals restricts its flexibility. The third analog, chrg-Ala6, reflects an intermediate point between these two extremes. This spread of analogs is expected to allow one to study the effects of constraints on the energy landscape of this hexapeptide. In addition, the inclusion in this study of the highly constrained cyclic analog, along side the unconstrained analog, allows us to address an issue relevant to drug discovery. It is well known that cyclization is often employed in drug discovery to reduce the flexibility and increase the bioactivity of peptide drug candidates.

C. Principal coordinate analysis

A problem inherent in polypeptide conformation spaces is their extremely high dimensionality. A molecule of N atoms has $3N$ degrees of freedom, and its corresponding conformation space is $3N$-6 dimensional. As a result, even relatively small molecules have very large conformation spaces (more than 100 dimensions for the alanine hexapeptide analogs studied here). It is clearly not practical to chart molecular energy landscapes in the full $3N$-6 dimensional space. Luckily, in practice, a much smaller number of dimensions is sufficient to capture the essential information required for energy landscape map making. This is achieved by projecting the full multidimensional space on an appropriately low dimensional subspace. Reducing the dimensionality of multidimensional conformational spaces can be obtained by principal component analysis (PCA).[24–26,28–31,34,46–50]

The PCA variant applied in this study was the so-called principal coordinate analysis (PCoorA) originally developed by Gower.[51] In general, PCA projects the $n\times m$ data matrix $\mathbf{M}$ (a distribution of n points in an m variable space) on a transformed axes set in which a low-dimensional subspace containing most of the relational information about the original distribution can be identified. In the context of conformational analysis this matrix holds a set of n conformations described by the points $P_i(q_{i1},q_{i2},\ldots,q_{im})$ in an m-dimensional conformation space. However, while the standard PCA operates on the square $m\times m$ $\mathbf{M^TM}$ matrix, known as the ''covariance matrix'' $\mathbf{C}$, reflecting the relationships between the *coordinates*, the PCoorA operates on the square $n\times n$ $\mathbf{MM^T}$ matrix known as the ''distance matrix'' Δ, reflecting the relationships between *conformations*. This matrix is transformed into a centered matrix which is then diagonalized. The resulting eigenvalues (normalized) give the percentage of the projection of the original distribution on the new set of coordinates, and the eigenvectors (scaled by their corresponding eigenvalues) give the coordinates of the original data points in the new axes frame.[30,31,51]

One of the advantages of PCA in general is that the normalized λ_i eigenvalues, associated with each principal axis (eigenvector), are directly related to the average error associated with the projection. Principal axes are sorted according to their normalized λ_i eigenvalues. The larger the eigenvalue, the more efficient is the projection onto these axes (reflecting a large variance for the data in the 1D projection). When projecting onto the first m principal axes, the average deviation of the actual distance d_{ij} between data points and the distances $d_{ij}^{(m)}$ calculated in the m-dimensional subspace is given by

$$\text{average error}=1-\sum_{k=1}^{m}\lambda_i=\langle d_{ij}^2-d_{ij}^{(m)^2}\rangle_{ij}, \tag{1}$$

where $\langle\cdots\rangle_{ij}$ is the average over all possible ij distances in the ensemble. A detailed study of several peptide systems has shown that in many cases the first few principal axes represent the multidimensional data to accuracy greater than 70%.[31] It should be stressed that, even if the two- or three-dimensional subspaces represent the original distances only to 40% or 50% accuracy, the effective accuracy is higher than predicted by Eq. (1). This is because the average accuracy is skewed by a relatively small number of poorly represented points, while the majority is represented at accuracy levels greater than 40% or 50%, respectively.

A key element in the principal coordinate analysis is the choice of distance measure used to construct the distance matrix Δ. Studies have shown that the choice of distance measure, e.g., a Cartesian distance or a distance in dihedral angle space, has a significant effect on the resulting projection.[30–32] In the present study the distance between any two conformations is measured as the root mean square distance (rmsd) in Cartesian coordinates. The rms distance d_{ij} between conformations i and j of a given molecule is defined as the minimum of the functional

$$d_{ij}=\sqrt{\frac{1}{N}\sum_{k=1}^{N}|\mathbf{r}_k^{(i)}-\mathbf{r}_k^{(j)}|^2}, \tag{2}$$

where N is the number of atoms in the summation and $\mathbf{r}_k^{(i)},\mathbf{r}_k^{(j)}$ are the Cartesian coordinates of atom k in conformations i and j, respectively. The rms distances are typically calculated based either on the backbone atoms or on all the nonhydrogen atoms in the molecule.

When PcoorA is applied to a single molecule l, it needs to be supplied with the upper diagonal distance matrix Δ_l. However, when conformations of two analogous molecules are to be compared with each other, it is essential that they be projected together onto the *same* subspace. In order to project the conformation ensembles of two related molecules, l and m, onto the same subspace the ''cross''-distance matrix Δ_{lm} must be calculated (in addition to the two self-distance matrices). The elements of the rectangular Δ_{lm} ''cross'' matrix are the distances between all conformations of molecule l and all conformations of molecule m. Thus, to obtain a joint projection of molecules l and m PCoorA is applied to the combined distance matrix $\mathbf{D}$

$$\mathbf{D}=\begin{pmatrix}\Delta_l & \Delta_{lm}\\ 0 & \Delta_m\end{pmatrix}, \tag{3}$$

where Δ_l and Δ_m are the upper diagonal ''self''-distance matrices and Δ_{lm} is the rectangular ''cross'' distance matrix. The size of the joint $\mathbf{D}$ matrix is $(n+n')\times(n+n')$, with n conformations of molecule l and n' conformations of molecule m. Equation (3) is easily extended to any arbitrary number of analogous molecules.

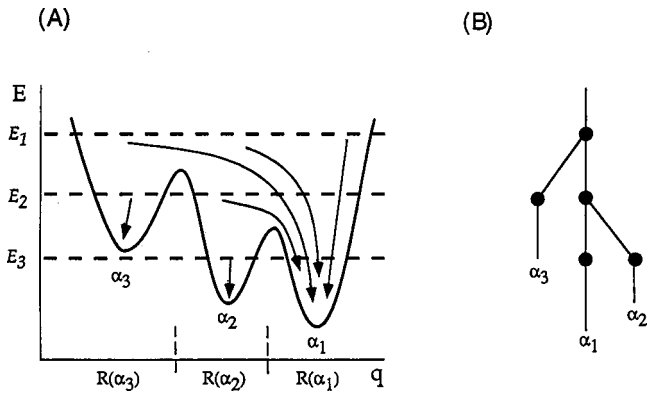

FIG. 3. Schematic representation of topological mapping of an energy landscape. (A) The energy landscape is studied at different energies E. Each energy region of connected conformations, denoted as a super-basin $R^E(\alpha)$, is mapped to its lowest minimum α. (B) The corresponding topological disconnectivity tree graph, $G^E(\Phi)$, reflects the way super-basins become disconnected as the energy decreases.

D. Topological mapping

Topological mapping, which was introduced by Becker and Karplus[15] and rests upon the connectivity pattern imposed by the energy barriers, characterizes the overall topography of complex energy landscapes. Using information about the barrier this method partitions a conformation space into energy basins, highlighting their interconnectivity. An elementary basin $R(\alpha)$ on the energy landscape is a connected set of molecular conformations that, when minimized, map to a common single local minimum. Topological mapping groups these elementary basins according to the barriers between them. At any energy level E (or temperature level T) this procedure partitions the multidimensional landscape into super-basins, $R^E(\alpha')$, defined as the union of elementary basins $R(\alpha)$ connected by barriers lower than energy E (or temperature T). Each such super-basin is then mapped to its lowest minimum α' in a way analogous to simulated annealing [Fig. 3(A)]. As a result minima connected by barriers lower than E are grouped together and separated from other minima to which they are connected by higher barriers. A topological disconnectivity graph is obtained by following the way these super-basins break up as the system's energy E decreases. Each node on this graph [Fig. 3(B)] reflects a conformational super-basin on the landscape, and the connecting edges reflect the basin connectivity. The topological mapping method resembles to the Lid method, which was independently developed by Sibani and Schön[52,53] in their study of the energy landscapes of crystals and glasses.

An important feature of topological mapping is that the resulting disconnectivity tree-graph $G^E(\Phi)$ [see, for example, Fig. 3(B)] reflects in a straightforward way the overall topography of the energy landscape Φ. As discussed and illustrated in the above-mentioned paper by Becker and Karplus,[15] a tree-graph reflecting a funnel topography will be characterized by a single main branch with many small side branches that do not show further branching. On the other hand, the tree-graph $G^E(\Phi)$ that corresponds to a landscape characterized by several large competing basins will exhibit several large branches, each exhibiting a complex branching pattern of their own. In the case of a completely rough landscape no significant branch will be detected in the $G^E(\Phi)$ graph.

In principle, a topological map is generated on the basis of information regarding direct barriers, i.e., barriers connecting neighboring minima. When working with conformation *samples* (versus a full enumeration of all local minima) the definition of direct barriers has to be extended. To compensate for missing data, indirect barriers along paths connecting sample points that are not strictly neighboring are taken into account, i.e., minima separated by other minima that are not included in the conformation sample. These indirect barriers are taken as the highest point along a multibarrier least energy path connecting two sample points. In this study the conjugated peak refinement algorithm of Fisher and Karplus[54] has been used to calculate these least energy paths. It is worth mentioning that with 500 conformations a staggering number of about 125 000 barriers have to be calculated. Fortunately, the number of barrier evaluations may be significantly reduced without affecting the results. First, the conformation pruning, as described above, already reduces the number of pathways to be calculated. Furthermore, it is reasonable to assume that indirect barriers between conformations that are very far apart on the energy landscape will hardly contribute to the connectivity pattern. Thus, a distance criterion can be imposed to refrain from calculating barriers between conformations that are too far away from each other. Re-evaluating the alanine–tetrapeptide topological map as computed by Becker and Karplus[15] it can be shown that the correct disconnectivity graph $G^E(\Phi)$ is reproduced even when only as little as the nearest 50% of the barriers are retained. Shorter cutoffs resulted in a disconnectivity graph with incorrect connectivity or missing features. Applying the combined procedure of pruning and cutoffs to the hexa-alanine analogs resulted in about 20 000 barrier evaluations when generating the topological map of Ala6, about 5500 barrier evaluations for the topological map of cyc-Ala6, and about 2000 barrier evaluations for chrg-Ala6.

IV. RESULTS AND DISCUSSION

A. Quality of sampling

A problem common to all sampling procedures is the difficulty to assess their thoroughness. Regardless of the sample size and the procedure used, there is always a question regarding how representative the resulting conformation sample is. In particular, it is important to know whether all accessible regions in the conformation space were sampled or whether some regions remained unvisited.

In this section a new way is proposed to address this fundamental question of sample quality, namely an evaluation of sampling overlaps. Let us assume that two conformation samples of the same system were generated by two different sampling protocols (e.g., different initial conditions or different methods). If it could be shown that two different samples overlap and occupy the same region in conformation space, this would indicate that the conformation search had indeed been exhaustive. On the other hand, a clear indication of incomplete sampling is if the conformation samples do not

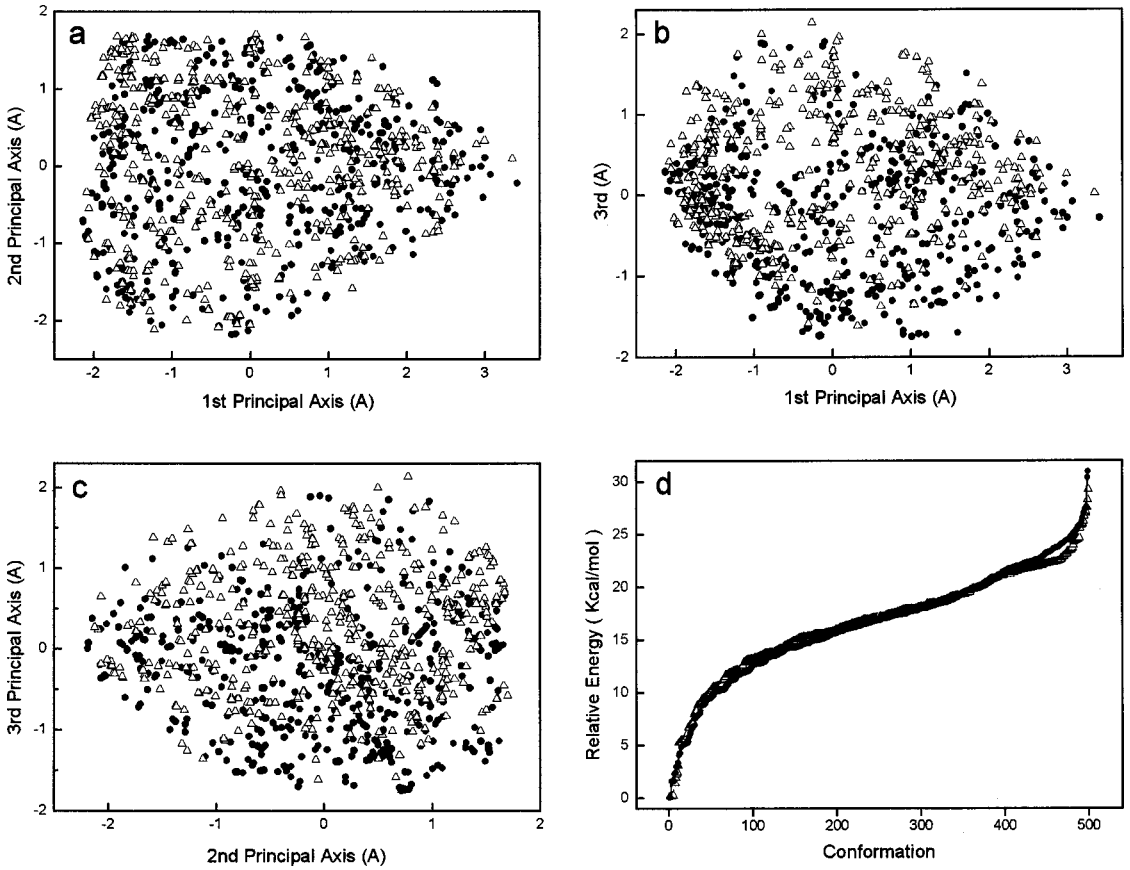

FIG. 4. Joint projection of two different Ala6 conformation samples (each consisting of 500 conformations) onto the best principal three-dimensional subspace, shown as three two-dimensional planes [(a), (b), and (c)]. Each point represents a single conformation, where circles denote conformations from one conformation sample and triangles from the other one. The energy profiles of the two conformation samples are also shown (d). The apparent overlap between the two conformation samples indicates that the conformation sample covers the whole available conformation space.

overlap. The evaluation of the overlap between the two conformation samples can be done with the aid of principal component projections through a joint projection onto the same low-dimensional principal subspace [Eq. (3)].[34] In the case of incomplete sampling the two conformation samples will have little overlap (if any) in the projected subspace.

The sample overlap approach was used to check whether the sampling procedure used in this study had indeed been exhaustive. Two conformation samples of linear Ala6, 500 structures each, were generated using the procedure described above, but starting from different initial structures. One sampling trajectory started from an extended peptide conformation, while the other trajectory started from an almost cyclic conformation of the linear peptide. A 1000 $\times$1000 joint distance matrix of the two conformation samples [Eq. (3)] was constructed using all-atom Cartesian rms distances, with PCoorA applied to it. The first three normalized eigenvalues of the joint projection were 21%, 12%, and 7%, indicating that the accuracy of the best joint 3D projection is about 40%. Figure 4 shows the three 2D cross sections through this best joint 3D projection of the two conformation samples. Each point in the projection is a single peptide conformation. As can be seen, the overlap between the two conformation samples along the first and second principal axes (which reflect the largest variance) is very high. A high level of overlap is also seen along the third principal axis, although some mismatch can be noted in this case. Figure 4(d) shows that the energy profiles of the two groups of sampled conformations are also very similar. Thus, it can be concluded that the two conformation samples exhibit a high level of overlap, indicating that the sampling procedure used was well suited to cover the peptide's conformation space (namely, those regions accessible at 1000 K). Since the linear Ala6 clearly has the largest conformation volume of the three hexapeptides studied here, it can be assumed that the same sampling method will also be appropriate for the other two conformationally more restricted molecules, cyc-Ala6 and chrg-Ala6. Consequently, a single 500-conformation sample seems to be sufficient to represent the conformation spaces of these two peptides as well.

B. Conformation samples reduction

The three conformation samples described above inevitably include groups of similar conformations, which are redundant as far as constructing the underlying energy landscape is concerned. As discussed above, by applying a distance criterion similar conformations are pruned from the conformation sample. Two conformations are defined as being similar if the all-atom Cartesian rms distance between the two is smaller than a given cutoff distance. This distance was set to 1.5 Å for Ala6 and to 0.9 Å for cyc-Ala6 and chrg-

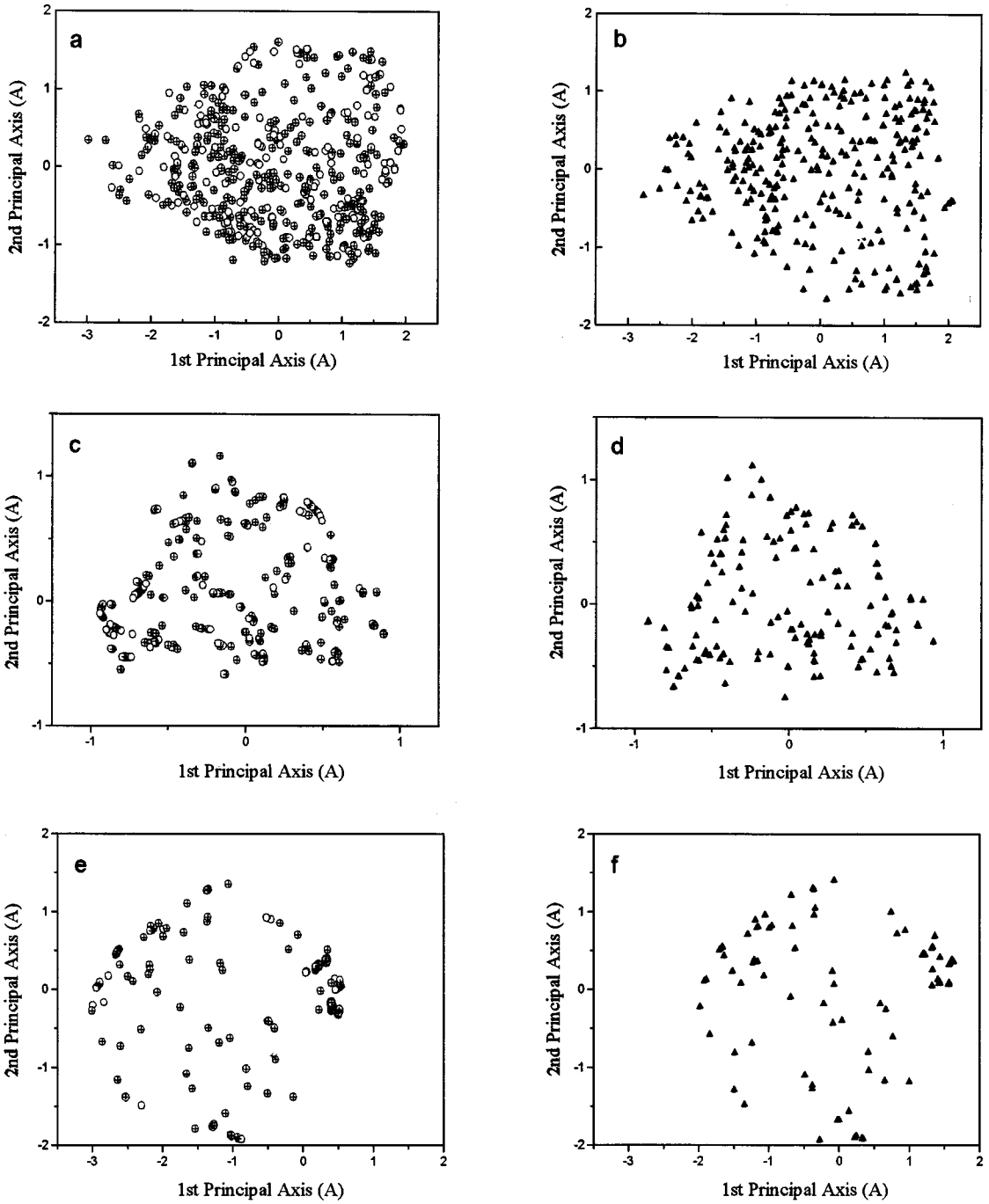

FIG. 5. Principal coordinate projections of full and reduced conformation samples of three alanine hexapeptide analogs. (a) The full sample of Ala6 500 conformations versus (b) the reduced conformation sample of Ala6 280 conformations. The sign (+) indicates those conformations in the full sample that were retained in the reduced sample. Plates (c) and (d) show the full and reduced conformation samples of cyc-Ala6 (500 vs 148 conformations). Plates (e) and (f) show the full and reduced conformation samples of chrg-Ala6 (500 vs 86 conformations).

Ala6. After removing the redundant conformations the diluted samples included 280 distinct Ala6 conformations, 148 distinct cyc-Ala6 conformations, and 86 distinct chrg-Ala6 conformations. In all cases the pruned conformations were very similar to other conformations, both in structure and in energy. The energy differences between similar conformations were below 0.8 kcal/mol for Ala6, below 0.2 kcal/mol for cyc-Ala6 and below 3.2 kcal/mol for chrg-Ala6.

To verify that the pruning had not affected the quality of the samples, PCoorA was used to compare the conformation coverage before and after the reduction. Figure 5 shows side by side the best two-dimensional PCoorA projections of the full and the reduced conformations samples of the three peptides. A separate projection was constructed for each full or diluted sample. Projections of full samples were based on the original 500×500 distance matrices, while projections of the diluted samples were based on the corresponding smaller distance matrices, depending on the size of the reduced sample. For the full samples, the 2D subspaces of the constrained peptides (cyc-Ala6 and chrg-Ala6) were accurate to 70% and 80%, respectively; with the corresponding 3D projections accurate to 80% and 85%, respectively (according to

backbone distances comparisons). Only the projection of the unconstrained peptide, Ala6, was of lower accuracy amounting to 40% and 45% for the 2D and 3D projections, respectively. The fact that the low-dimensional Ala6 projection was less accurate means that its effective conformation space is more isotropic than its constrained analog and that it requires additional dimensions for a more accurate description. The validity of the sample reduction procedure was confirmed by the fact that the accuracy of the reduced projections was very similar to that of the corresponding full projections. Similarity to the full samples of the constrained peptides, the two-dimensional projections of their reduced samples were very accurate (68% for cyc-Ala6, 79% for chrg-Ala6), while much less accurate for the reduced Ala6 sample (only 32% of the full variance is captured in this 2D projection). Figure 5 also clearly shows that for all three molecules the reduced conformation samples overlap very well with the original full samples, indicating that the reduction does not adversely affect the quality of the sample.

C. Effect of constraints on conformation space volume

The volume in conformation space accessible to a molecule at a given temperature varies from one molecule to another, reflecting their flexibility. Therefore, conformation constraints that reduce the flexibility of the molecule also decrease the size of this volume. Since both the cyc-Ala6 (through cyclization with a covalent bond between the two termini) and the chrg-Ala6 (through a strong charge-charge interaction between the termini) are conformationally constrained analogs of linear Ala6, the accessible volumes in their conformation spaces are expected to be smaller than those available to the unconstrained peptide analog. However, a question that so far eluded quantification is to what degree their conformation volume is reduced. This study shows that a combination of sampling and joint principal coordinate projection offers a direct way to quantify the effect of conformation constraints on the conformation volume. In a separate study Becker, Levy, and Ravitz have shown that these conformation volumes can be used quantitatively in a QSAR formulation to predict the bioactivity of conformationally constrained analogs.[34]

A quantitative comparison of the conformation volumes accessible to the three alanine hexapeptides was obtained by jointly projecting the three conformation samples onto the best joint 3D subspace. Figure 6(a) shows the resulting joint 3D projection, which is based on backbone rms distances, and allows intermolecular comparisons since all three molecules share a common backbone. This joint 3D projection is quite accurate, representing all distances to an average accuracy of 60.7%. The ellipsoids shown in Fig. 6(b) are sketched to highlight the volumes occupied by each peptide analog. Two properties are clearly seen: (i) As expected, the two constrained analogs occupy significantly smaller conformation volumes than the conformation volume associated with the unconstrained analog Ala6, and (ii) the conformation volumes occupied by these two analogs are comprised within the conformation volume of Ala6. Furthermore, the fact that the conformation volumes accessible to the two constrained analogs are of similar size and exhibit significant overlap indicates that they share similar conformational properties. Specifically, both favor closed circular structures with short distance between amino-acids 1 and 6, which are brought about either by a covalent bond (cyc-Ala6) or by strong electrostatic attraction (chrg-Ala6).

To quantify the conformation volumes accessible to each hexapeptide analog a 3D ellipsoid is fitted around each conformation sample. These ellipsoids are based on 3D covariance matrices in the joint projected 3D subspace. Assuming that the contribution of the higher principal dimensions diminishes rapidly and that their contribution to all three conformation volumes is similar, comparing these 3D volumes should be quantitatively similar to comparing the full accessible conformation volumes. Furthermore, the logarithm of these volumes should be roughly proportional to the conformational entropy.[29] The 3D ellipsoid volumes thus calculated of cyc-Ala6, chrg-Ala6, and Ala6 are 0.68, 1.42, and 18.73 $Å^3$, respectively, which is equivalent to volume ratios of 4%, 8%, and 100% (it is however reiterated that the quality of the projection from which these volumes are derived is only 60%). This means that the conformation constraints imposed on cyc-Ala6 and chrg-Ala6 reduce the accessible conformation space to a fraction of its original size. Not surprisingly, cyclization reduces the conformation volume even more than introducing two terminal charges (for a system in a vacuum). The latter retains a certain (though small) amount of conformation freedom between the two terminal groups, which is completely lost by the introduction of a covalent bond.

An important property of the joint projection is that conformations (points) close to each other in the projection are conformationally similar, while conformations (points) that are mapped into different parts of the projected subspace are conformationally different. Figure 7 shows three conformations, one from each peptide analog, mapped close to each other in the region where the three conformation volumes overlap. The conformational similarity between the three is evident, indicating that the conformations of chrg-Ala6 and cyc-Ala6 are quite similar. Figure 8 shows the most stable Ala6 conformation that has an α-helical character. This conformation was mapped outside the overlap region of the three ellipsoids since it does not have counterparts in the conformations of the constrained molecules.

D. Minima-based view of the energy landscapes

Adding energy to the above principal coordinate projections allows for charting and visualizing the energy landscape of these peptides.[29,31] Figure 9 shows the principal two-dimensional projections of the conformation spaces of the three peptide analogs arranged according to the energies of their local minima. Each small rectangle, or slab, in this figure is a full principal 2D projection plane on which *only* those conformations that have energy within a given energy range are indicated. The first rectangle at the bottom includes the lowest energy conformations. The next rectangle depicts the same 2D plane on which only conformations in the second energy range are shown, and so forth. For Ala6 and cyc-Ala6 the energy scale [Figs. 9(a) and 9(b)] was divided into 2 kcal/mol slabs, while for chrg-Ala6 the slab width was

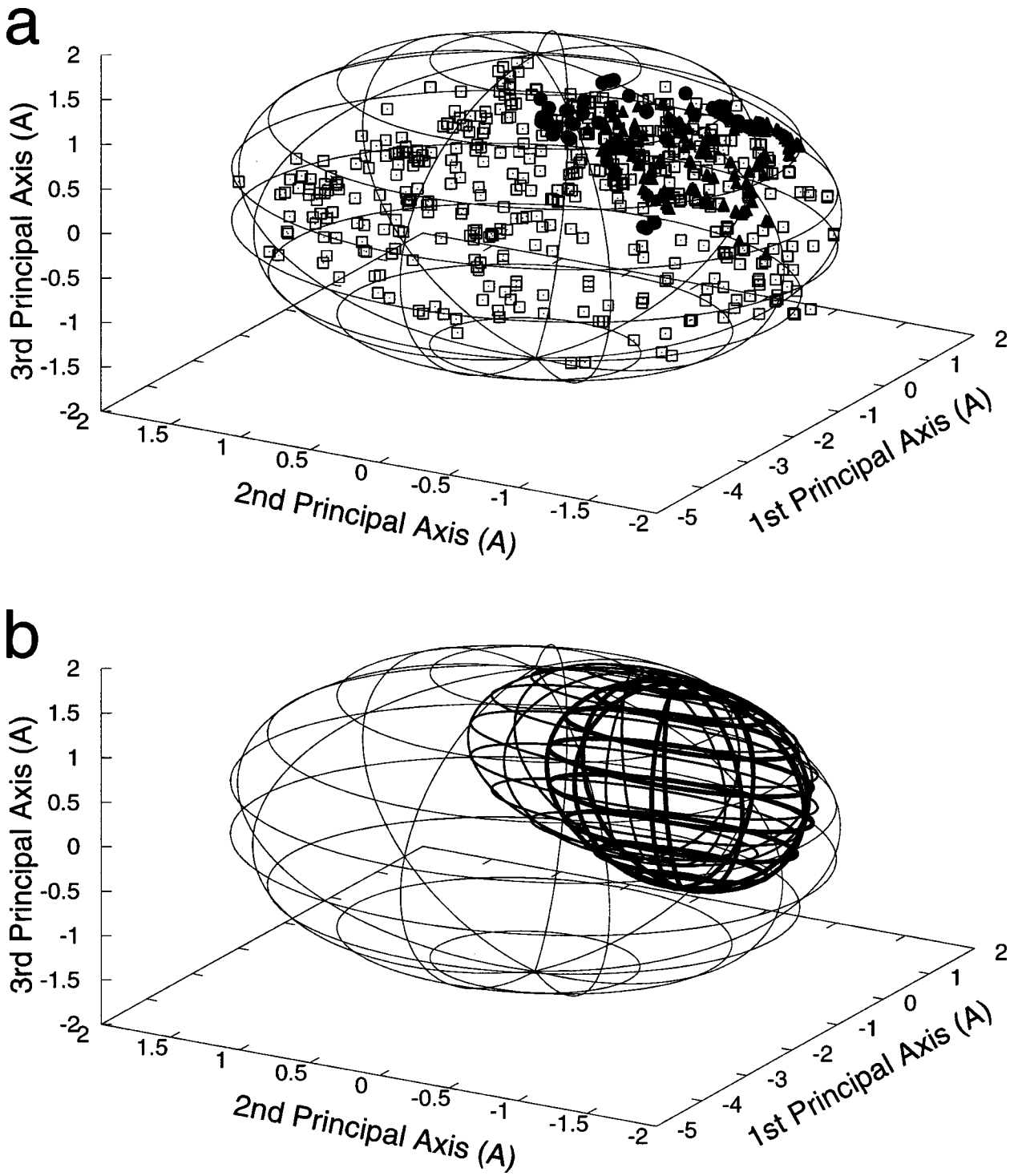

FIG. 6. (a) Joint 3D projection of the molecular conformation spaces of the three alanine hexapeptide analogs: Ala6 (squares), cyc-Ala6 (triangles), and chrg-Ala6 (circles). (b) Representation of the ellipsoids that enclose the conformation volumes of the three peptides in the projected 3D subspace (Ala6: thin line, cyc-Ala6: intermediate line, chrg-Ala6: heavy line). The reduction in conformation volume of the Ala6 peptide upon introduction of the constraints (cyc-Ala6, chrg-Ala6) is clearly evident.

3 kcal/mol (resulting in 13, 7, and 13 energy slabs, respectively). The mean energy in each slab is indicated in the right corner of each rectangle. Finalizing this issue, Fig. 9 in its entirely yields quantitative maps of the three energy landscapes (to the accuracy of the PCoorA projections). If all the rectangles in Fig. 9 were to be rotated by 90° and then stacked one on top of the other, a 3D view of the landscape would have resulted. The dashed curves in Fig. 9 highlight the underlying basin structure revealed in this way. According to these maps which, as already indicated, are based only on local minima, the energy landscape of Ala6 is dominated by a broad and deep basin, which looks like a broad funnel [Fig. 9(a)]. The landscape of chrg-Ala6 is also dominated by a single basin, which in this case is much deeper and narrower than the one characterizing Ala6 [Fig. 9(c)]. Finally, cyc-Ala6 shows a very different energy landscape that in-

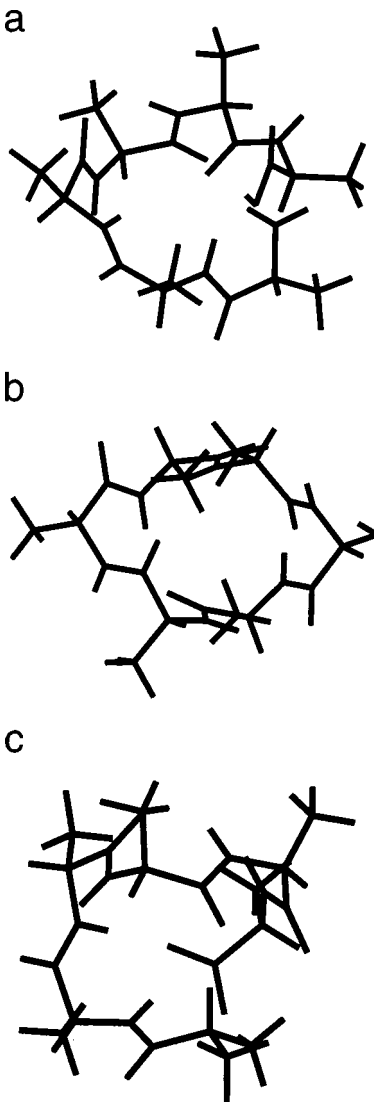

FIG. 7. Conformations of the three hexa-peptides situated closely to each other in the overlap region of the small ellipsoids (Fig. 6): cyc-Ala6 (a), chrg-Ala6 (b), and Ala6 (c). The three structures are closely packed, reflect the nearest points on the projections and characterize conformations with a similar geometry. Moreover, it is shown that the linear peptide, Ala6, has (a) in its conformation space region of closed structures even though its stable structure is α-helix.

cludes three competing basins, none of which dominates the landscape [Fig. 9(b)], with energy gaps between the deepest basin *A* and basins *B* and *C* being only 2 to 4 kcal/mol (the separation between basin *B* and *C* is along the second principal axis).

An alternative representation of the landscape, which is also based on local minima only, is the minimal energy envelope representation.[31] In this approach the energy envelope underlying the 2D principal projection is calculated and presented as a 3D topographical map (the minimal energy envelope is similar to the dashed curve in Fig. 9). Figure 10 shows the three energy landscapes obtained by the minimal energy envelope procedure. The resulting 3D surfaces highlight the topographical features observed in Fig. 9. Figure 10(a) shows the broad funnel-like basin of the Ala6 energy landscape. Figure 10(c) exhibits an extremely deep and narrow basin, which completely dominates the energy landscape

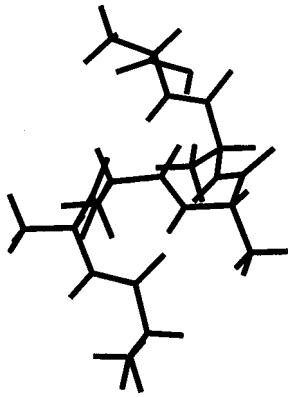

FIG. 8. The most stable conformation of the linear peptide, Ala6. This structure has an α-helical character and is located outside the region of overlap of the three ellipsoids.

of chrg-Ala6. It also shows a split at the bottom of this deep well and indicates that no single basin dominates the energy landscape of cyc-Ala6. On the other hand, the three basins pointed out in Fig. 9(b) are clearly observed in the 3D visualization. Figure 10 also shows that all three energy landscapes exhibit a significant amount of roughness. In particular, the bottom of the energy basins is characterized by many splits, indicating that the native structure of each peptide is an ensemble of conformations and not a single well-defined structure. Recalling the definition of foldability *F* as the ratio between the depth of the funnel and the roughness of the landscape,[36] it can be concluded that among the three peptides studied here chrg-Ala6 exhibits the highest degree of foldability (namely, it can get to its folded native structure very quickly), Ala6 has a smaller degree of foldability, while cyc-Ala6 is characterized by a low value of *F* indicating that no specific minimum has a significant advantage over the other.

E. Order parameter and the energy landscapes

Energy landscapes are inherently multidimensional and require high-dimensional representations in order to visualize them. However, in the absence of other tools, much of the current discussion is formulated in terms of a single order parameter, tacitly assuming that a one-dimensional reaction coordinate is sufficient to represent the kinetics of protein folding. The order parameter most commonly used in such studies is *Q*, which reflects the fraction of native contacts in any given conformation.[1,2] *Q* equals 1 for the native structure and is smaller for all other conformations. This order parameter has been used in many lattice folding studies as well as in some all-atom simulations,[6,7,55,56] although other order parameters have also been suggested.[8–11,13,57,58] As these order parameters are assumed to be proportional to the energy, it is interesting to explore to what extent such an order parameter really correlates with the real multidimensional energy landscape.

The order parameter *Q* is usually defined as the fraction of native contacts present in any conformation ($Q=1$ is the native conformation). This definition is clearly suitable for simplified lattice models where contacts are easily defined.

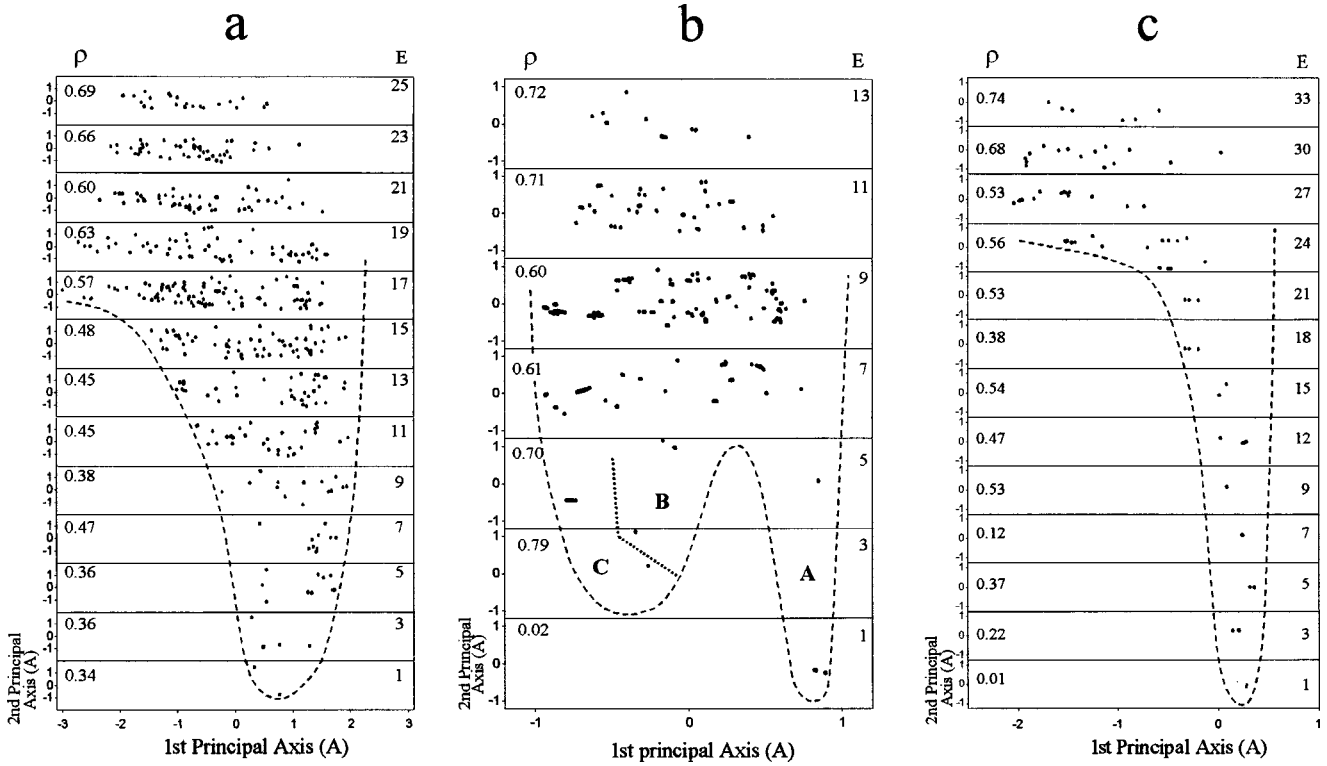

FIG. 9. Principal two-dimensional projections of the conformation spaces of the three peptide-analogs separated according to the energies of the local minima: (a) Ala6, (b) cyc-Ala6, and (c) chrg-Ala6. Each rectangle layer (slab) is the molecule's principal 2D projection plane on which only conformations with energy at a given energy range are shown. The first rectangle from the bottom includes the lowest lying conformations, the next rectangle depicts the same 2D plane on which only conformation in the second energy range are shown, and so forth. The mean energy of each slab is indicated at the top-right corner of each rectangle (grouped in ranges of 2 kcal/mol for Ala6 and cyc-Ala6, and 3 kcal/mol for chrg-Ala6). Also shown for each slab is the average value of the order parameter ρ (see text). The dashed lines highlight the underlying basin topography. The energy landscapes of Ala6 and of chrg-Ala6 show a single deep basin each, while the landscape of cyc-Ala6 includes three competing basins (basins *B* and *C* are separated along the second principal axis).

For off-lattice model studies, as well as for all-atom simulations, the definition of native contacts becomes imprecise. In their extensive analysis of all-atom simulations Sheinerman and Brooks[13,14] introduced a related but continuous order parameter ρ, which weights the fraction of native contacts according to the actual distance between a set of predefined contact pairs ($\rho=0$ for the native conformation). For the smaller peptides studied here even the Sheinerman and Brooks ρ order parameter is not directly applicable, because there are no real tertiary native contacts in hexapeptides. Instead, an analogous continuous order parameter is defined to characterize to what extent a given structure is nativelike. The new order parameter, denoted by ρ to indicate its similarity to the Sheinerman and Brooks ρ, measures nativeness based on the dihedral angles that characterize the native structure. The order parameter ρ for conformation i is calculated as

$$\rho(i)=\sum_{j=1}^{M} |\theta_j^{\text{native}}-\theta_j^{(i)}|/60, \tag{4}$$

where θ_j^{native} is the value of jth dihedral angle in the native conformation, $\theta_j^{(i)}$ is the value of the jth dihedral angle in the ith conformation and the summation is over all M dihedral angles. In the present application the summation is over the 10 backbone dihedral angles ϕ and ψ. In other applications side-chain χ angles may also be included in the summation of Eq. (4). The division by the factor of 60 qualitatively bins the dihedral angles into the $\pm 60°$ range around the native value. Using this measure the native conformation has the value of $\rho=0$ and ρ increases as the conformations becomes less nativelike.

To see whether the order parameter ρ defined in Eq. (4) correlates with the energy landscapes, the average order parameter $\langle\rho\rangle$ over all the conformations at each energy slab in Fig. 8 was calculated by

$$\langle\rho\rangle_E=\frac{1}{N_E}\sum_{\{i\}_E}\rho(i), \tag{5}$$

where N_E is the number of states (conformations) in the energy slab $\{i\}_E$, defined as

$$\{i\}_E=\{i|E-\Delta E<E_i<E+\Delta E\} \tag{6}$$

and has a width of $2\Delta E$ around its median value E_i ($\Delta E=1$ kcal/mol for Ala6 and cyc-Ala6, $\Delta E=1.5$ kcal/mol for chrg-Ala6).

Figure 11 depicts the average $\langle\rho\rangle$ values calculated over all conformations within a given energy slab, according to Eqs. (4)–(6), versus the median potential energy E of each of these slabs (see Fig. 8). The native conformation has a value of $\rho=0$ and ρ increases as the conformations become less nativelike. Also indicated in Figs. 11 is the variability of individual ρ values within each energy slab (defined by the standard deviation). For linear Ala6 the average order parameter $\langle\rho\rangle$, as well as the individual ρ values, gradually decrease

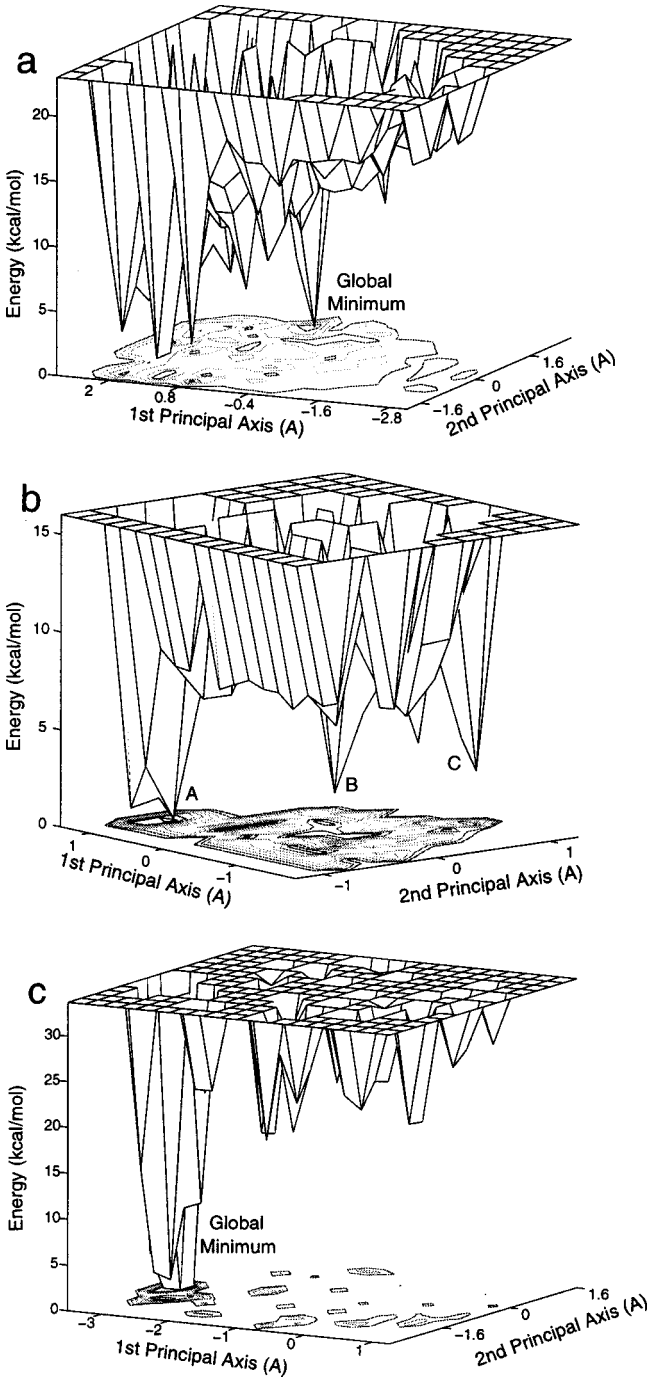

FIG. 10. Three energy landscapes obtained by the minimal energy envelope procedure (see text), for Ala6 (a), cyc-Ala6 (b), and chrg-Ala6 (c). The two principal axes indicate conformational similarity and the vertical-axis reflects the relative energy. The landscape is based on the energies of the local minima only (the connecting barriers are not shown).

as the energy decreases [Fig. 11(a)]. Namely, moving down into the native basin the internal structure of the conformations gradually becomes more nativelike. This quantitative observation supports the hypothesized correlation between order parameters such as Q or ρ and the potential energy. The relatively large value of $\langle\rho\rangle\approx 0.4$ present at the lowest energy slab indicates that even at the bottom of the funnel there is still a large degree of structural variability. This observation is in accord with previous observations regarding the energy landscape of alanine tetrapeptide.[29] Figure 11(a) also shows that the level of roughness, represented by the variability of individual ρ values within an energy slab, is almost constant throughout the funnel. For all but the lowest two energy slabs, this variability is about ± 0.12 around the average $\langle\rho\rangle$.

An even stronger correlation between the order parameter $\langle\rho\rangle$ and the energy landscape funnel is observed for chrg-Ala6 [Fig. 11(c)]. This molecule, which is characterized by a deeper and narrower funnel than Ala6, also exhibits a stronger decrease with energy of the average order parameter $\langle\rho\rangle$ than Ala6. For chrg-Ala6 the average value of the order parameter ρ decreases from 0.74 all the way to 0.01, indicating that all conformations at the bottom of the narrow funnel are highly nativelike. The kink at $E=7$ kcal/mol is due to the very small number of conformations in this slab. In addition to the decrease in $\langle\rho\rangle$, Fig. 11(c) shows also a decrease in the variability of individual ρ values with decreasing energy. This observation indicates that, in addition to being deep and narrow, the chrg-Ala6 funnel also becomes *smoother* as the energy decreases.

In agreement with the previously made observation, the correlation plot in Fig. 11(b) indicates that the energy landscape of cyc-Ala6 is significantly different from that of the other two peptides. In this case, with decreasing energy, there is practically no decrease, in the average order parameter, and $\langle\rho\rangle$ stays close to $\rho=0.7$ all the way down. A localization, first near $\rho=0.8$ (basins B and C) and then at $\rho\approx 0$ (basin A), is observed only at the lowest energies. The reason for the high $\langle\rho\rangle$ values at basins B and C is due, of course, to the definition of ρ as measuring the similarity to the lowest energy conformation in basin A. A qualitatively similar picture would result if ρ were to be defined relative to basin B, with basin A conformations yielding high ρ values.

Summing up, the results show that for the two hexapeptides that exhibit an overall funnel topography, the effective order-parameter ρ seems to be in good agreement with the multidimensional energy landscapes.

F. Folding pathways

Even though the conformation samples generated for the present study were not obtained via folding simulations, but rather through a sampling procedure, their distribution indicates accessible regions in conformation space and thus highlights effective folding pathways. To check for possible pathways connecting the unfolded manifold of states with the global minimum (representing here the native state), a scheme similar to that used by Sheinerman and Brooks[13,14] was adopted. The three conformation samples were projected onto a plane defined by two order parameters. One order parameter is ρ, reflecting the degree of similarity (in dihedral angles) to the native structure. The second order parameter is the backbone root mean square distance (rmsd, here measured in Å) between each conformation and the native state, reflecting primarily the degree of collapse (in this respect it has a similar role to the radius of gyration in globular proteins). If the conformation samples were to fully represent the Boltzmann weights in each region of the plane, the resulting plot would constitute a full description of the folding

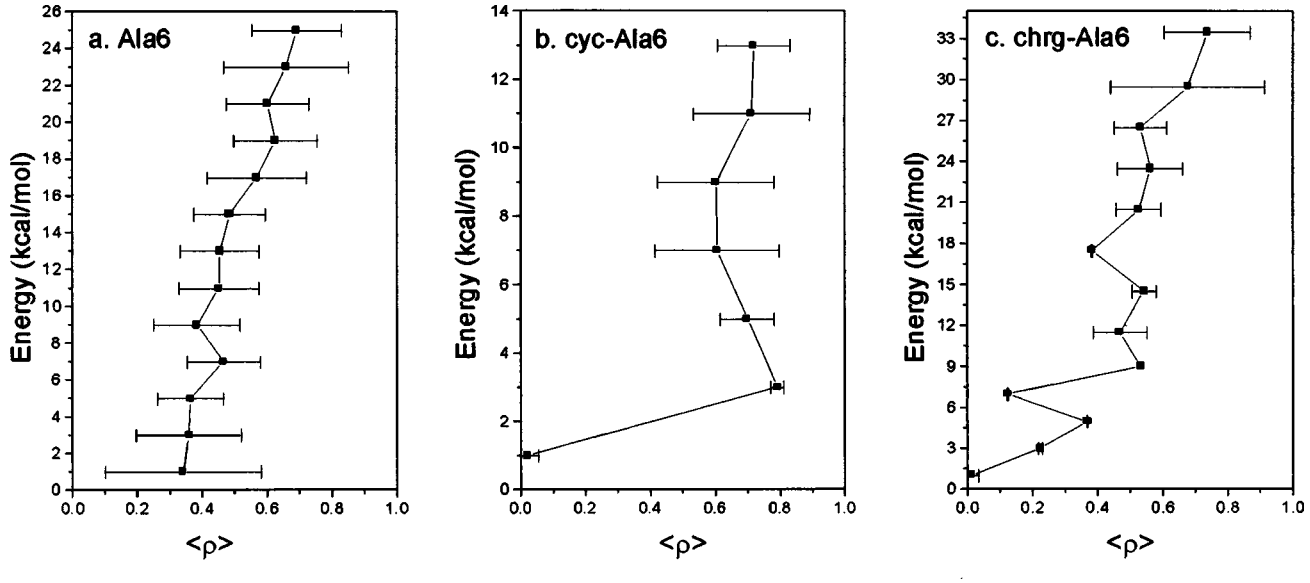

FIG. 11. Average $\langle\rho\rangle$ order parameter values versus potential energy values for the three hexapeptides: Ala6 (a), cyc-Ala6 (b), and chrg-Ala6 (c). Each point indicates the average $\langle\rho\rangle$ calculated over all conformations within a given slab (see Fig. 9). The native conformation has a value of $\rho=0$ and ρ increases as the conformations become less nativelike. The horizontal error bars indicate the variability (measured by the standard deviation) of individual ρ values within each energy range. The gradual decrease of the order parameter ρ with decreasing energy (observed for Ala6 and chrg-Ala6), indicates a funnel-like landscape, in agreement with Fig. 9. Cyc-Ala6 does not show this behavior.

pathways. However, since the sampling procedure used in this study has not been designed to characterize Boltzmann weights, the resulting pathways, while suggestive and indicative, are not necessarily complete.

Figure 12 depicts number-density contours resulting from projecting the three full conformation samples on the ''two order-parameter'' plane defined by ρ (indicating similarity to the native conformation) and on rmsd (indicating a collapse towards the native structure). The three frames of Fig. 12 show that a different type of folding pathway and folding kinetics characterizes each of the three peptides. For linear Ala6 Fig. 12(a) reveals a two-state $U\rightarrow N$ type folding pathway (where U stands for unfolded and N stands for native). The single diagonal arrow in Fig. 12(a) indicates that the two order-parameters are strongly correlated. This means that for this molecule the ρ order parameter would make a good effective reaction coordinate, capturing most of the systems kinetics. Of special interest is the gap seen between the bulk of unfolded states ($\rho>0.4$) and the native state ($\rho\approx 0$), which probably arises from insufficient sampling in that region. This insufficient sampling may indicate that the main barrier separating the folded basin N from the unfolded basin U is in the region characterized by ρ values between 0.1 and 0.3. It is interesting to note that in folding simulations of a 27-mer lattice model of a polypeptide the transition region was found to be close to the native state at Q values of 0.7 to 0.9, where Q measures the percentage of native contacts.[1,6] Since the discrete Q values are equivalent to the

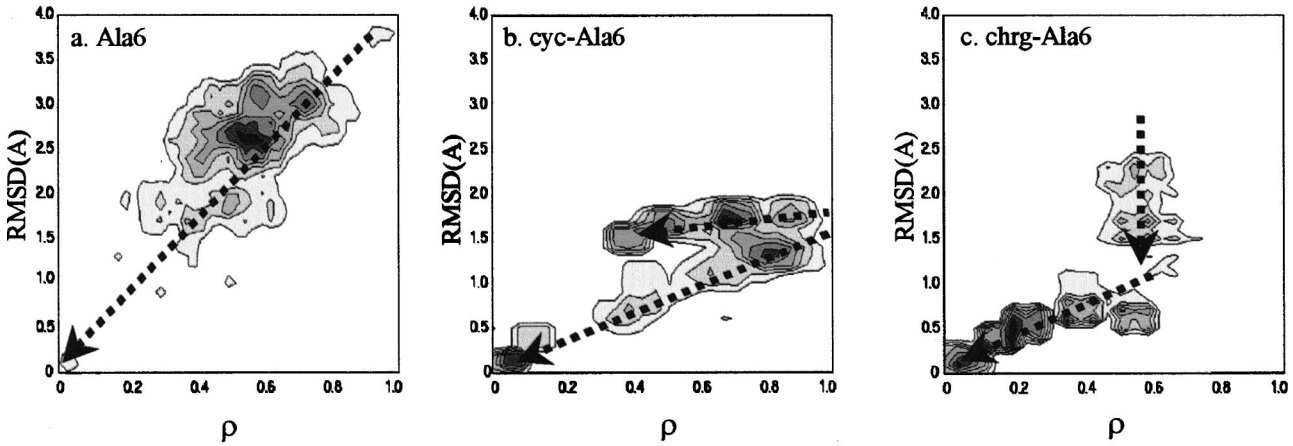

FIG. 12. Projection onto a ''two order-parameter'' plane $\{\rho$, rmsd$\}$ of the full conformation samples of the three hexapeptides: (a) Ala6, (b) cyc-Ala6, and (c) chrg-Ala6. The ρ axis indicates the degree of similarity (in dihedral angles) to the native conformation. The rmsd axis indicates a collapse towards the native structure (all atom rms distance in Cartesian space). The contours indicate the number density of sampled conformation at each point on this plane. Even though the projected conformations were not generated via folding simulations, their distribution highlights effective folding pathways, schematically illustrated by the arrows.

continuous $1-\rho$, the barrier in those lattice simulations is at $\rho \approx 0.1-0.3$, similar to the outcome for Ala6.

A more complicated folding pathway is observed for chrg-Ala6 [Fig. 12(c)]. For this molecule the ''two order-parameter'' plot indicates a three state folding pathway of the type $U \rightarrow I \rightarrow N$, where I is a nonnative collapsed intermediate. The first leg of the pathway, represented by the vertical arrow (rmsd values from 2.5 Å to 1.0 Å at a constant ρ value close to 0.6), is a collapse phase characterized by a rapid ring closure, brought about by the two opposite terminal charges that strongly attract each other. This intermediate state is non-native at $\rho \approx 0.6$. Following the collapse the backbone dihedral angles continue to rearrange until the native conformation is reached (indicated by the second arrow). The observed three-state process also indicates that neither ρ nor rmsd can be used as a single effective order parameter to describe the folding of chrg-Ala6. In this case the folding kinetics requires at least two order parameters for an adequate description. It should be stressed that the observed three-state behavior does not necessarily indicate a stable intermediate at the non-native collapsed state. It may well be that the collapsed state is a necessary on-pathway conformation but it does not constitute a stable intermediate, so that the overall kinetics may still obey the two-state $U \rightarrow N$ scheme. The present analysis was unable to resolve the two possibilities.

Finally, for the cyc-Ala6 system Fig. 12(b) points to a folding mechanism different from the previous two. Unlike the single pathways that seem to characterize both Ala6 (a two-state mechanism) and chrg-Ala6 (a three-state mechanism), for the cyclic analog cyc-Ala6 we observe three competing folding processes. The different pathways originate from the unfolded manifold U but proceed to different native states N_1, N_2, and N_3 respectively, the three processes being $U \rightarrow N_1$, $U \rightarrow N_2$, and $U \rightarrow N_3$. The first native state N_1 can be identified with basin A and is characterized by a $\rho = 0$ value. The second native state N_2 can be identified with basin C, which has an average $\langle \rho \rangle$ value of 0.67 ± 0.12 (where ρ is defined relative to the lowest minimum in basin A). The third folding process into the N_3 state, which corresponds to basin B, cannot be resolved in this plot. This is due to the fact that its average $\langle \rho \rangle$ value (relative to the lowest minimum in basin A) is 0.86 ± 0.07, similar to the order parameters characterizing the unfolded manifold U. The pathway into basin B would be easily observed if ρ was to be calculated relative to the structures characterizing basin C or B.

Figure 12(b) indicates that even the two observed folding pathways cannot be resolved along the one-dimensional ρ order parameter coordinate. This means that while each individual pathway is well described by the single ρ coordinate, it is not sufficient for representing the overall kinetics. At least one additional coordinate (and possibly more) is required for a full analysis of the kinetics. In fact, relying on a single order parameter in this case is likely to lead to incorrect conclusions.

G. Landscape topology and connectivity

The above analysis of energy landscape topography and folding pathways is based on the energies and spatial distribution of the local minima. The resulting description, while very informative, is still incomplete because it does not take into account the distribution of *barriers* which connect the local minima. These barriers define the actual kinetic connectivity of the landscape, determining which transition pathways are accessible and which are not. The method of topological mapping described above was developed to analyze that type of overall barrier connectivity, which is specific to each energy landscape.[15] In the context of the present study it is worthwhile to compare the picture that emerges from the barrier-based topological analysis with the views obtained from studying the distribution of local minima. In particular, the disconnectivity graphs characterizing the three hexapeptide landscapes should be compared with their principal coordinate projections (Figs. 9 and 10) and with their characterization using the effective order parameter ρ (Fig. 11).

To calculate the topological disconnectivity graphs of the three Ala6 analogs the two dilution processes described in Sec. III were used. First, conformations exhibiting high similarity were removed from the sample and then a distance constraint was imposed to refrain from calculating barriers between minima that are too distant on the energy landscape. For the molecules studied here the distance criteria were 3.5 Å for Ala6 (compared to a maximal distance of about 6.2 Å), 2.2 Å for cyc-Ala6 (maximal distance about 3.5 Å) and 2.7 Å for chrg-Ala6 (maximal distance about 3.9 Å). These criteria ensure, on an average, that for each local minimum, the barriers to most of its neighboring minima are calculated (the nearest 50% of all minima at least). Complying with the above criteria about 20 000 barrier evaluations were performed to generate the topological map of Ala6, about 5500 for cyc-Ala6 and about 2000 chrg-Ala6.

In a previous study it has been shown that the disconnectivity graphs of Ala6 and cyc-Ala6 differ significantly from each other, demonstrating the dramatic effect of conformation constraints on the topography of the molecular energy landscape.[16] As seen in Fig. 13(a) the disconnectivity graph of linear Ala6 depicts a single dominant branch reflecting a simple funnel topography. Each node reflects barriers that split a basin into its disconnected sub-basin components. Following the graph from the top down, only simple splitting is encountered from the main branch to disconnected, mainly high-energy, conformations (although a single low energy conformation also becomes disconnected fairly high up in the tree). Still, near the bottom of the graph the splitting pattern becomes more complicated, indicating a richer structure of the landscape near the bottom of the funnel. Furthermore, Fig. 13(a) also indicates that the energy landscape of Ala6 has a fairly constant roughness over a very broad range of energy levels. This property is reflected in the disconnectivity graphs by the similar number of minima that branched off from the main branch at the different energy levels.[16]

Thus, both the topological analysis and the minima-based analysis give rise to a picture of Ala6's energy landscape as being dominated by a single funnel. A different question is whether the effective order parameter ρ, which seemed to be correlated with the minima-based picture, can

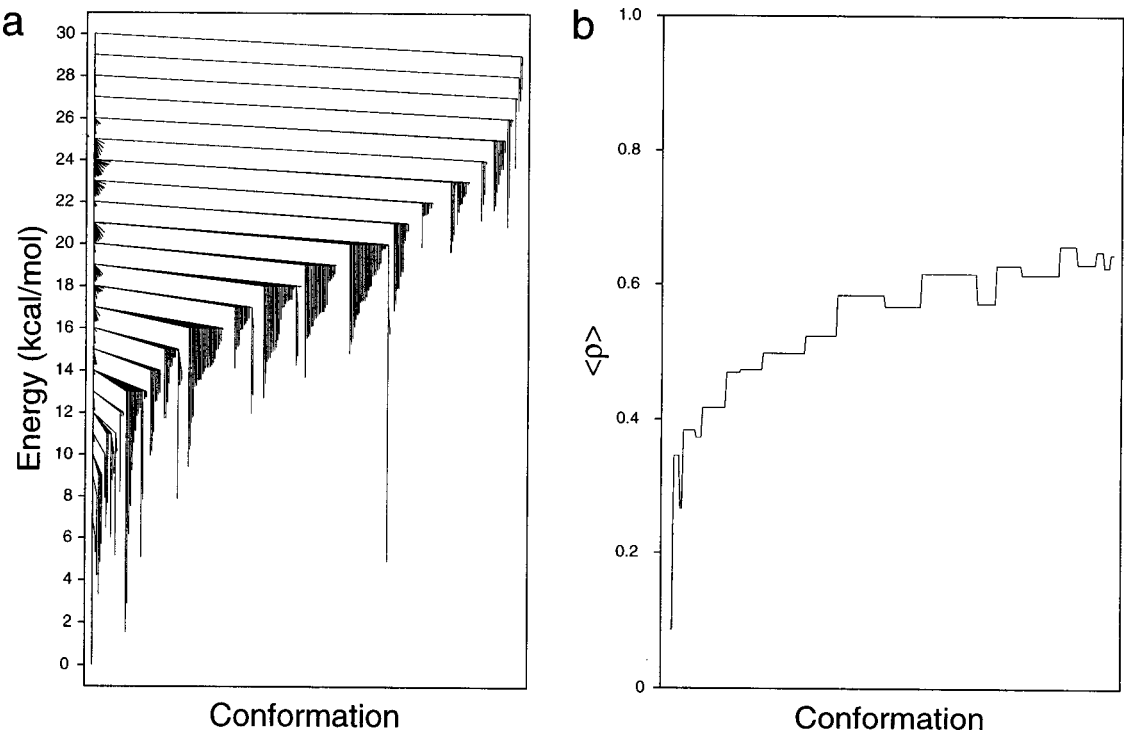

FIG. 13. (a) The topological disconnectivity graph for Ala6 (see text) and (b) the average $\langle\rho\rangle$ values for each branch of the disconnectivity graph, displayed on an identical horizontal axis. Both graphs indicate a single dominant funnel on the energy landscape of Ala6.

capture the descent into the main funnel (reflected by the disconnectivity graph). To answer this question the average ρ value was calculated for each branch of the graph. Figure 13(b) shows the average $\langle\rho\rangle$ values for each branch of the Ala6 disconnectivity graph, displayed on a horizontal axis identical to the one used for the disconnectivity graph [Fig. 13(a)]. It becomes clear that the average $\langle\rho\rangle$ values of the conformations that part away from the main funnel gradually decrease as one proceeds down into the funnel, indicating that the conformations gradually become more similar to the native state. The strong correlation between the disconnectivity pattern and the average $\langle\rho\rangle$ values indicates that for Ala6 the energy landscape is indeed governed by a single large funnel and supports the effectiveness of the order parameter as a sufficient reaction coordinate.

Figure 14 depicts the disconnectivity graph and average $\langle\rho\rangle$ value per branch for chrg-Ala6. As with Ala6 the graphs are highly correlated and show a single dominant funnel structure. This funnel is deeper than that of Ala6, in agreement with the previous observations. It is interesting to note that the disconnectivity graph of chrg-Ala6 shows a gap between the unfolded high energy states (nodes above 20 kcal/mol) and the nativelike low energy conformations (nodes below 14 kcal/mol). This gap, which represents a steep localization on the energy landscape, is reflected by a plateau in the average $\langle\rho\rangle$ values and then followed by a sudden drop towards the native state. This structure is probably associated with the two-step folding pathway that was observed on the ''two order-parameter'' projection of Fig. 12.

Finally, the disconnectivity graph clearly indicates (as is indicated by other approaches too) that the energy landscape of cyc-Ala6 is very different from the other two hexapeptides [Fig. 15(a)]. The disconnectivity graph of cyc-Ala6 splits into two additional basins at a barriers height of 12 kcal/mol. The main branch, denoted as basin *A*, includes a group of minima connected by 5 kcal/mol barriers and is separated from the rest of the system by barriers in the range of 10–12 kcal/mol. Basins *B* and *C* appear as two branches that, unlike most other branches, continue to split after being disconnected from the main basin. Compared to basin *A*, these two basins show more gradual internal splitting patterns with barriers on the order of 7–10 kcal/mol. It is interesting to note that on this energy landscape the roughness is restricted mainly to a narrow energy range in the vicinity of the barriers that split the landscape into three competing basins. The nonfunnel multiple basin character of the cyc-Ala6 landscape is also seen in the corresponding average $\langle\rho\rangle$ values depicted in Fig. 15(b). The average $\langle\rho\rangle$ values decrease only once the peptide reaches the edge of basin *A*, while the other two basins have much higher $\langle\rho\rangle$ values.

V. CONCLUSIONS

In this paper an analysis was made of the topography of three alanine hexapeptide analogs: linear alanine hexapeptide with neutral terminals (Ala6), linear alanine hexapeptide with charged terminals (chrg-Ala6) and cyclic alanine hexapeptide (cyc-Ala6). These analogs differ in the amount of constraints imposed on their conformation flexibility. While the motion of the linear peptide with uncharged terminals is essentially unconstrained, adding opposite charges at the two terminals forces the molecule into a cyclic conformation (in vacuum), thus significantly reducing its range of motion. The constraint on the motion becomes even stricter

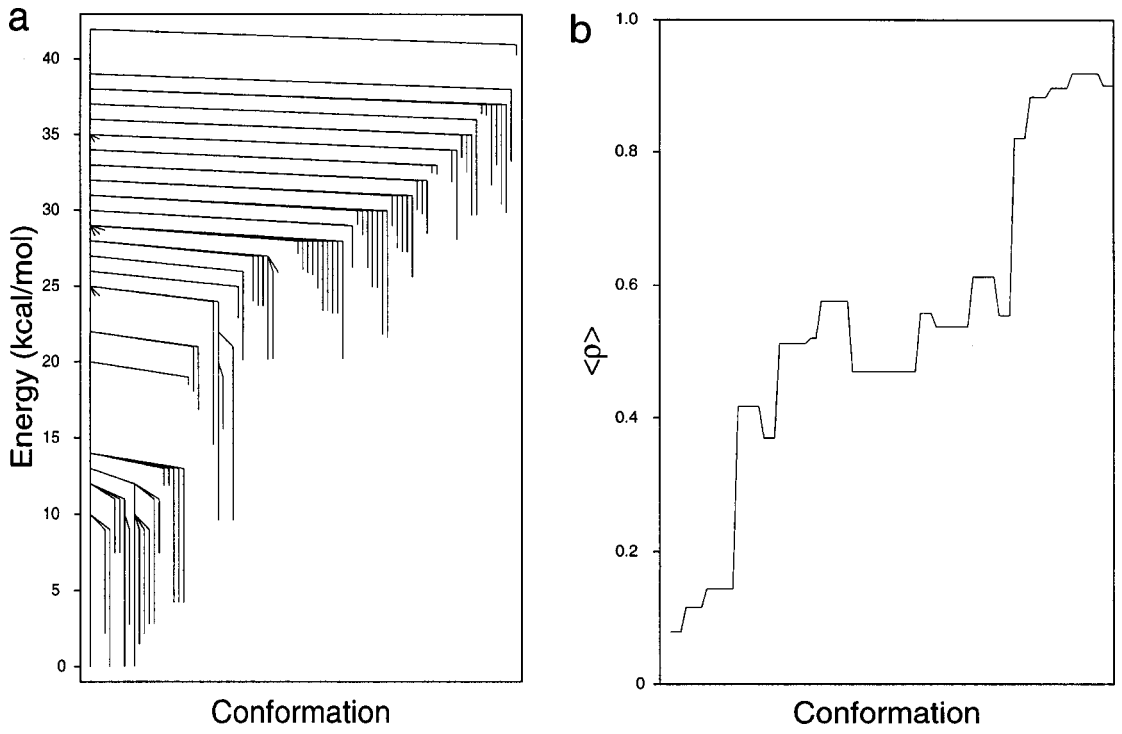

FIG. 14. Similar to Fig. 13, for chrg-Ala6. Both graphs indicate a single dominant funnel on the energy landscape of chrg-Ala6.

when a covalent bond is introduced between the two terminals of the peptide, forming the conformationally restricted cyclic analog. The present study has shown that the effect of these conformation constraints on the energy landscape that underlies the chemo-physical properties of these peptides, can be quantified and compared. In particular, it has been found that the three analogous peptides are characterized by significantly different energy landscapes. While the energy landscape of Ala6 is that of a broad and rough funnel, and the energy landscape topography of chrg-Ala6 that of a deep and narrow funnel, the energy landscape of cyc-Ala6 is characterized by three competing basins. The differences in en-

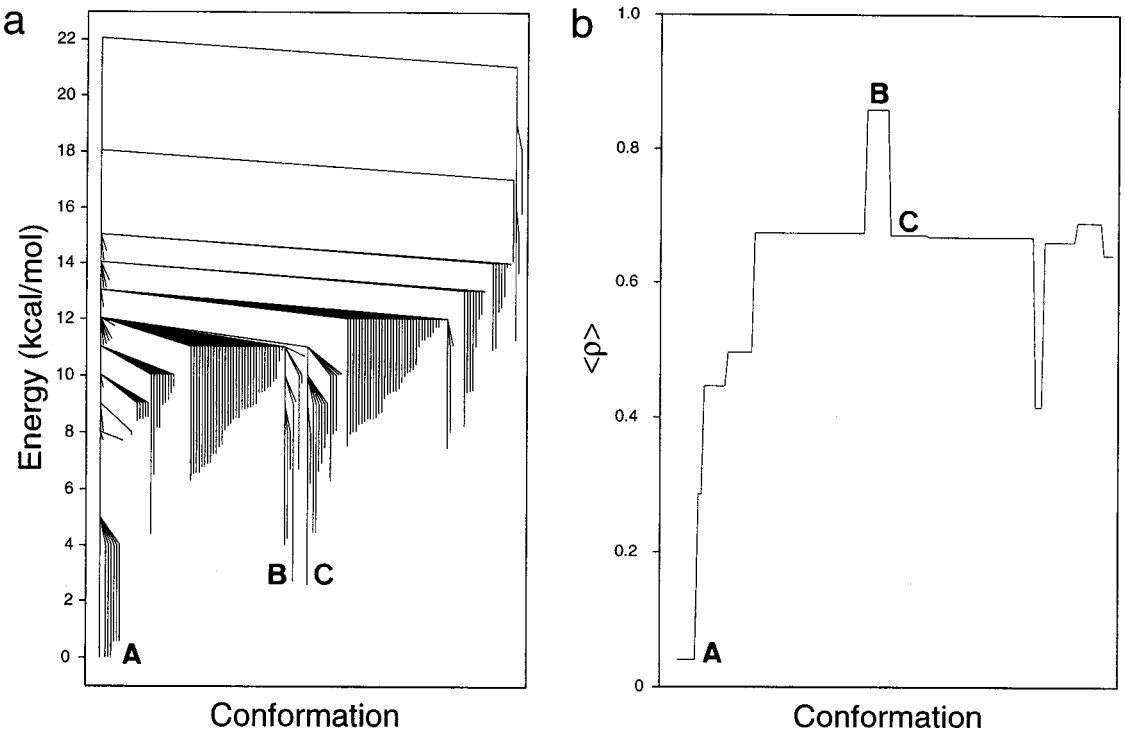

FIG. 15. Similar to Fig. 13, for cyc-Ala6. The graphs indicate that the energy landscape of chrg-Ala6 does not exhibit a funnel structure, but rather incorporates three competing basins.

ergy landscapes are reflected in different folding mechanisms, with Ala6 exhibiting a direct one-step folding, chrg-Ala6 a two-step folding, while cyc-Ala6 exhibits competing pathways leading from one basin to another. Although not studied here, the differences in landscapes would clearly yield differences in thermodynamic properties as well. The results obtained are of importance to protein folding in general, since the three peptides studied here would, in principle, be encoded by the same gene. This means that the differences observed between their energy landscapes arise from environmental or post-translational modifications and are not encoded in the gene. Namely, these results suggest that non-genomic post-translational modifications may play an important role in determining the properties of proteins and their folding patterns.

In addition, the present study indicates that, at least for two of the peptides studied, there is a strong correlation between the different views of the landscapes obtained by principal component analysis, topological mapping and order parameter analysis. In particular it has been demonstrated that one or two order parameters, such as Q, are able to capture much of the information regarding the overall topography of the landscape even in these realistic all-atom models. However, this seems to be true only in the presence of a dominant funnel on the landscape (as in the case of Ala6 and chrg-Ala6). In cases of several competing basins on the same energy surface the simple order parameters were inadequate to resolve the complexity of the surface topography. Nevertheless, in all three cases, both the principal component analysis (which was based only on local minima) and the topological mapping analysis (based on minima and barrier information) were able to resolve the complex topographies, yielding similar results.

ACKNOWLEDGMENTS

This study was funded in part by a grant from the Israel Ministry of Science and Technology. The authors are very grateful to Yaacov Vardi and Jacqueline Gorsky for critical reading of this paper.

[1] C. M. Dobson, A. Sali, and M. Karplus, Angew. Chem. Int. Ed. Engl. **37**, 868 (1998).
[2] M. Karplus, J. Phys. Chem. B **104**, 11 (2000).
[3] A. Ansari, J. Berendzen, S. F. Bowne, H. Frauenfelder, I. E. T. Iben, T. B. Sauke, E. Shyamsunder, and R. D. Young, Proc. Natl. Acad. Sci. U.S.A. **82**, 5000 (1985).
[4] I. E. T. Iben, D. Braunstein, W. Dister, H. Frauenfelder, M. K. Hong, J. B. Johnson, S. Luck, P. Ormos, A. Schulte, P. J. Steinbach, A. H. Xie, and R. D. Young, Phys. Rev. Lett. **62**, 1916 (1989).
[5] H. Frauenfelder, S. G. Sliger, and P. G. Wolynes, Science **254**, 1598 (1991).
[6] A. Sali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).
[7] N. D. Socci, J. N. Onuchic, and P. G. Wolynes, J. Chem. Phys. **104**, 5860 (1996).
[8] H. S. Chan and K. A. Dill, Proteins **30**, 2 (1998).
[9] A. R. Dinner and M. Karplus, J. Phys. Chem. B **103**, 7969 (1999).
[10] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).
[11] H. S. Chan and K. A. Dill, J. Chem. Phys. **100**, 9238 (1994).
[12] R. Elber and M. Karplus, Science **235**, 318 (1987).
[13] F. B. Sheinerman and C. L. Brooks III, Proc. Natl. Acad. Sci. U.S.A. **95**, 1562 (1998).
[14] F. B. Sheinerman and C. L. Brooks III, Proteins **29**, 193 (1997).
[15] O. M. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).
[16] Y. Levy and O. M. Becker, Phys. Rev. Lett. **81**, 1126 (1998).
[17] D. J. Wales, M. A. Miller, and W. TR, Nature (London) **394**, 758 (1998).
[18] J. P. K. Doye, M. A. Miller, and D. J. Wales, J. Chem. Phys. **110**, 6896 (1999).
[19] R. S. Berry and R. E. Kunz, Phys. Rev. Lett. **74**, 3951 (1995).
[20] R. E. Kunz and R. S. Berry, J. Chem. Phys. **103**, 1904 (1995).
[21] K. D. Ball, R. S. Berry, R. E. Kunz, F.-Y. Li, A. Proykova, and D. J. Wales, Science **271**, 963 (1996).
[22] K. D. Ball and R. S. Berry, J. Chem. Phys. **109**, 8557 (1998).
[23] K. D. Ball and R. S. Berry, J. Chem. Phys. **111**, 2060 (1999).
[24] A. E. Garcia, Phys. Rev. Lett. **68**, 2696 (1992).
[25] A. E. Garcia, in *Nonlinear Excitations in Biomolecules*, edited by M. Peyrard (Springer, Berlin, 1994), pp. 191–207.
[26] R. Abagyan and P. Argos, J. Mol. Biol. **225**, 519 (1992).
[27] M. A. Balsera, W. Wriggers, Y. Oono, and K. Schulten, J. Phys. Chem. **100**, 2567 (1996).
[28] L. S. D. Caves, J. D. Evanseck, and M. Karplus, Protein Sci. **7**, 649 (1998).
[29] O. M. Becker, J. Mol. Struct.: THEOCHEM **398–399**, 507 (1997).
[30] O. M. Becker, Proteins **27**, 213 (1997).
[31] O. M. Becker, J. Comput. Chem. **19**, 1255 (1998).
[32] N. Elmaci and R. S. Berry, J. Chem. Phys. **110**, 606 (1999).
[33] S. B. Prusiner, Cell **93**, 337 (1998).
[34] O. M. Becker, Y. Levy, and O. Ravitz, J. Phys. Chem. B **104**, 2123 (2000).
[35] R. D. Levine and R. B. Bernstein, *Molecular Reaction Dynamics and Chemical Reactivity* (Oxford University Press, New York, 1987).
[36] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 539 (1997).
[37] F. H. Stillinger and T. A. Weber, Science **225**, 983 (1984).
[38] T. Noguti and N. Go, Proteins **5**, 97 (1989).
[39] R. S. Berry, N. Elmaci, J. P. Rose, and B. Vekhter, Proc. Natl. Acad. Sci. U.S.A. **94**, 9520 (1997).
[40] A. Fernandez, K. S. Kostov, and R. S. Berry, J. Chem. Phys. **112**, 5223 (2000).
[41] R. Czerminski and R. Elber, J. Chem. Phys. **92**, 5580 (1990).
[42] M. A. Miller and D. J. Wales, J. Chem. Phys. **111**, 6610 (1999).
[43] R. E. Bruccoleri and M. Karplus, Biopolymers **29**, 1847 (1990).
[44] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[45] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, J. Phys. Chem. B **102**, 3586 (1998).
[46] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Proteins **17**, 412 (1993).
[47] A. Amadei, A. B. M. Linssen, B. L. de Groot, D. M. F. van Alten, and H. J. C. Berendsen, J. Biomol. Struct. Dyn. **13**, 615 (1996).
[48] B. L. de Groot, A. Amedi, D. M. F. van Alten, and H. J. C. Berendsen, J. Biomol. Struct. Dyn. **13**, 741 (1996).
[49] J. M. Troyer and F. E. Cohen, Proteins **23**, 97 (1995).
[50] A. E. Garcia and G. Hummer, Proteins **36**, 175 (1999).
[51] J. C. Gower, Biometrika **53**, 325 (1966).
[52] P. Sibani, J. C. Schön, P. Salamon, and J. O. Andersson, Europhys. Lett. **22**, 479 (1993).
[53] J. C. Schön, Ber. Bunsenges. Phys. Chem. **100**, 1388 (1996).
[54] S. Fischer and M. Karplus, Chem. Phys. Lett. **194**, 252 (1992).
[55] J. M. Shin and W. S. Oh, J. Phys. Chem. B **102**, 6405 (1998).
[56] A. Sali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[57] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).
[58] J. N. Onuchic, P. G. Wolynes, Z. Luthey-Schulten, and N. D. Socci, Proc. Natl. Acad. Sci. U.S.A. **92**, 3626 (1995).