Conformational Analysis in a Multidimensional Energy Landscape: Study of an Arginylglutamate Repeat

Sara R. R. Campos and António M. Baptista*

Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Avenida da República, EAN, 2780-157 Oeiras, Portugal

Received: April 1, 2009; Revised Manuscript Received: August 14, 2009

The identification of the distinct conformation classes of a molecule is a common and often crucial step in establishing structure−function relationships. Many different methods have been suggested for that purpose which differ in their choice of a (dis)similarity measure and clustering algorithm. The present study discusses and analyzes these issues, proposing a method based on principal component analysis (PCA), which is applied to conformations obtained from molecular dynamics (MD) simulations of an arginylglutamate repeat. Simulations are done at different pH values, using both standard MD and constant-pH MD methods, with the peptide displaying a very high conformational variety. The conformational analysis starts with a comprehensive comparison of different sets of conformational coordinates and of their ability to preserve structural similarity between conformations. The selected set of conformational coordinates is then used to investigate the preservation of structural similarity after PCA transformation, concluding the need of using a multidimensional conformation space. This conformation space is then used to derive a multidimensional probability density and the corresponding energy landscape. The application of a simple cutoff algorithm to the resulting multidimensional landscape is then shown to produce a consistent set of distinct and homogeneous conformation classes. Overall, this methodology provides an efficient way to identify the major conformation classes of a molecule in a way that directly reflects the density of states in the multidimensional conformation space, contrasting with the more heuristic nature of standard clustering methods.

Introduction

Proteins and peptides constitute complex systems that can adopt a myriad of different and interconverting conformations. The characterization of the nature and relation of those conformations, particularly the identification of distinct classes, plays a fundamental role in rationalizing and understanding a diversity of functional aspects. Conformational analysis is thus an important tool in many subject areas of molecular modeling, ranging from the design of peptide drugs to the study of protein folding.

Except for very small molecules whose conformation spaces can be fully scanned, conformational analysis typically starts with the generation of a large set of conformations by molecular mechanics/molecular dynamics (MM/MD) simulations or other suitable methods.[1,2] A common approach is to proceed by choosing a measure of structural similarity between conformations (see below), compute it for all pairs in the generated set, and supply the resulting similarity matrix to a clustering algorithm that splits the set of conformations into distinct classes.[3,4] Although this approach can provide valuable insight into the structural diversity of conformations, it tends to neglect the relative agglomeration of states in conformation space. If the set of conformations was sampled from a proper ensemble (e.g., using MD simulations), that agglomeration or density of states in principle carries information about the energetics of the system because a high/low density of states means low/high energy in some kind of energy landscape.[5−11] By adopting a clustering method that neglects this density of states, we may end up with a collection of clusters poorly related to the actual energetics of the system.[12] An alternative approach is therefore to base the conformational analysis on an energy landscape.

The complete energy landscape of a molecule is a function of all conformational coordinates and, thus, contains all information required to build physically meaningful conformation classes. The "energy" mapped on this landscape is the potential energy for a molecule in vacuum, whereas for a solvated one, it is the conditional free energy (or potential of mean force) that already includes the average solvent effect. However, using the complete energy landscape is problematic. First, the complete specification of the conformation of a system of N atoms requires $3N - 6$ internal coordinates (in the absence of symmetry and constraints), which is a huge number of dimensions, even for relatively small systems. Thus, the complete characterization of the energy landscape defined over such $(3N - 6)$-dimensional space is feasible only for extremely simple systems.[13−15] Second, even if available, the complete energy landscape would give us much more information than we need in most actual studies, which often intend to identify just a few broad conformation classes (the active/inactive conformers of a peptide drug, the folded/misfolded/unfolded states of a protein, etc.). Even when the purpose is not merely to identify conformation classes but, rather, to understand in more detail the major conformational features of some process (e.g., finding the predominant folding pathways of a protein), we just want to characterize some kind of low-dimensional energy landscape that captures the relevant behavior of the system as a function of a small set of coordinates that represent the system in a more global and coarse-grained way.

The practical need to select a small number of degrees of freedom implies that the complete $(3N - 6)$-dimensional

* Corresponding author. E-mail: baptista@itqb.unl.pt.

10.1021/jp902991u CCC: $40.75
Published on Web 09/24/2009

conformation space must somehow be mapped onto a low-dimensional representation space retaining the most important features of the distribution of conformations. A common but somewhat subjective route is to choose a representation space whose coordinates are a set of properties that our physical intuition suggests as being appropriate. For example, some folding studies use as coordinates the number of native contacts and the radius of gyration (e.g., see refs 2 and 16), in which case the height on the landscape corresponds to the (conditional) free energy of the collection of all conformations that happen to have a given pair of values for those coordinates. A more objective route is to derive the coordinates of the representation space from the actual set of conformations, typically using collective coordinate methods[17] such as normal mode analysis, principal component analysis, and principal coordinate analysis. Normal mode analysis is restricted to the study of harmonic motions around an energy minimum, and it can give valuable insight on the global features of energy landscapes.[18] Principal component analysis (PCA) is a mathematical method, used in several scientific areas, whose main purpose is to reduce the dimensionality of a data set while maintaining most of its variation.[19,20] It was first applied to protein simulation as a tool to investigate biologically relevant motion in the beginning of the 1990s,[21−24] and since then, it has become a common approach to study not only protein motion but also energy landscapes.[25−32] Principal coordinate analysis (PCoorA) is a general method to map dissimilar objects onto a coordinate space,[33] being sometimes mistaken for PCA, and has also been applied to study protein conformation.[34−38] Unlike PCA, it does not require the objects (conformations) to be originally defined by other coordinates, but the PCoorA of a set of n objects requires the diagonalization of an $n \times n$ matrix, being unfeasible for large data sets. In the following, we restrict our discussion of collective coordinate methods to PCA.

The derivation of a representation space using PCA often starts by discarding some atoms, typically retaining only the peptide backbone or C_α atoms; this simply corresponds to removing some dimensions from the complete landscape. The Cartesian coordinates of the remaining atoms cannot be directly used as input for PCA because they include the uninteresting overall translational and rotational motion. Although translation can be easily removed by fixing the molecule's center of mass, rotation cannot be fully eliminated except for rigid bodies. Nonetheless, it is possible to reduce the overall rotation by a least-squares fitting of each structure on a reference one, and this approach has been used in many PCA studies.[23,29] Still, the arbitrariness behind the choice of the reference structure and of the subset of atoms used for fitting remains problematic, being specially critical for molecules with high conformational diversity, such as peptides. To avoid these problems, several strategies have been proposed. Prompers and Brüschweiler[39,40] developed a method that represents a given conformation as an ensemble with isotropic orientations. Another strategy is to use internal coordinates, which are independent of rotational and translational motion. The most common internal coordinates used with PCA are the distances between C_α atoms[41−43] and the backbone dihedrals.[29−31,44,45] In the latter case, the calculation problems associated with the periodicity of the dihedral space can be avoided by replacing each angle with its sine and cosine.[29,46]

Although the coordinates obtained from PCA, named principal components (PCs), are as many as the original coordinates, the rationale is to keep only the PCs needed to reasonably capture the distribution. Since the PCs are conventionally ordered by the amount of variation they encompass, a standard procedure is to select them sequentially until they account for a substantial percentage of the total variance, typically 70−90%.[19] The number of necessary PCs to achieve this must be determined in each case, since it depends on the system and the choice of original coordinates (e.g., see ref 23). Studies of protein motion using PCA often consider several PCs, trying to characterize the type of motion associated with each one.[23,47] In contrast, the characterization of the density of states in the PC space tends to follow a much more limited approach, typically displaying the density as a function of only two PCs (first PC versus second PC, first PC versus third PC, etc.)[25,27,29−32] and, more rarely, three PCs.[28,36,43,48] The very low-dimensional energy landscapes thus obtained are mainly intended as graphical aids, and a more detailed analysis often resorts to other techniques.[36,49,50] Therefore, most PCs are actually ignored when computing the density of states, meaning that the conformation classes supposedly revealed by those plots, which reflect the projection of many potentially discriminating dimensions, may end up mixing together disparate groups of conformations.

In the present work, we propose a PCA-based method to identify conformation classes in a multidimensional energy landscape, which is applied to MD trajectories of a small peptide. We start by comparing several types of conformational coordinates and evaluate their ability to preserve the relation between distance and structural dissimilarity. We then perform PCA using the selected coordinates, followed by the computation of the multidimensional density and energy landscape in the resulting PCs space, illustrating the problems of using just a couple of PCs. The identification of the minima and respective basins in the multidimensional landscape then leads to the determination of distinct conformation classes whose relation is derived from the transitions observed between. Our main objective is a conformational characterization of the peptide and not as much a topographical analysis of the landscape.

It is known that pH may significantly affect the conformation of peptides[51−54] and proteins,[55,56] and thus, it is interesting to know how well the proposed method deals with the sampling complications introduced by the consideration of protonation changes (e.g., each protonable group duplicates the total number of available states). With this in mind, we chose a peptide made of five arginylglutamic acid (RE) repeats, which will be referred to as RE5. There are several proteins containing repetitive motifs that alternate between basic and acidic residues in consecutive positions, some of which contain RE repeats.[57−62] The biological function of the RE repeats is not known, but experimental data indicates that it may be related to protein−protein interaction.[60−62] This peptide is essentially nonprotonable at pH 7 but highly protonable at both acidic and basic pH, thus being an interesting model to compare the performance of the method in both the nonprotonable and protonable cases. In the present work, we decided to study the neutral and acidic pH, performing the sampling of conformations using both standard MD and constant-pH MD simulations.[63]

Materials and Methods

Standard Molecular Dynamics. The standard MD simulations were performed with the GROMACS package,[64−66] version 3.2.1, using the GROMOS96 43a1 force field.[67] RE5 was solvated with 3999 single point charge (SPC) water molecules[68] in a rhombic dodecahedral box, with periodic boundary conditions. The minimum distance between the peptide and the box was 1.75 nm. The equations of motion were numerically integrated using a time step of 2 fs. The nonbonded interactions

were treated with a twin-range cutoff of 8/14 Å and updated every 10 fs. The reaction field method,[69] with a relative dielectric constant of 54.0,[70] was used for the long-range electrostatic interactions. Solvent and solute were separately coupled to temperature baths at 300 K, with Berendsen coupling[71] and relaxation time of 0.1 ps. A Berendsen pressure coupling[71] was used at 1 atm, with a relaxation time of 0.5 ps and isothermal compressibility of 4.5×10^{-5} bar^{-1}. All bonds were constrained using the LINCS algorithm.[72]

An initial energy minimization consisting of 2000 steps was performed using the steepest descent algorithm followed by another 2000 steps using the low-memory Broyden−Fletcher−Goldfarb−Shanno algorithm. The initiation procedure consisted of a 50 ps simulation with all atoms restrained, followed by a 50 ps simulation with only CA atoms restrained.

The peptide was built in an α-helix conformation and capped at the N-terminus with an acetyl group and at the C-terminus with an amino group. All residues were charged in the standard MD simulations which represented pH 7 (pH7).

We performed five standard MD simulations; four of them were 100 ns long and the other was 120 ns. The system was equilibrated at different time lengths (1, 7, 8, 11, and 26 ns) and only the equilibrated trajectories were used. Snapshots were saved every 1 ps.

Constant-pH Molecular Dynamics. The constant-pH MD method used was the implementation for the GROMACS package of the stochastic titration method previously described.[63,73−75] It consists in a cycle of three sequential blocks:

1. A block of Poisson−Boltzmann/Monte Carlo (PB/MC) calculations assigns the protonation states of the titrable sites of the protein/peptide.

2. A short MD simulation of the system with frozen protein/peptide allows for solvent relaxation after the charge modifications. In this work, the simulation time was 0.2 ps.

3. A full MD simulation of the unconstrained system will produce a trajectory with the assigned set of protonation states. The last snapshot of the simulation will be used in the first block of the following cycle. In this work, this block was 2 ps long.

The theoretical rationale of the method can be briefly stated as follows:[63] MM/MD is assumed to correctly sample configurations for a fixed set of protonation states, whereas PB/MC is assumed to correctly sample protonation states for a fixed protein conformation. Using a standard theorem from the theory of Markov chains, the succession of MM/MD and PB/MC simulations is then shown to produce a Markov chain that asymptotically samples from the proper semigrand canonical ensemble. Since the MM/MD simulations use explicit solvent and the PB/MC simulations use an automatically relaxed solvent (a dielectric continuum), a short MM/MD solvent relaxation corresponding to point 2 is performed after selecting each new set of protonation states for approximate consistency of the conditional distributions (see ref 63 for further discussion of this approximation).

The MD settings used in the constant-pH MD procedure were the same as in the standard MD except for the use of the generalized reaction field[76] (instead of the reaction field) and the size of the box (containing 4087 water molecules).

The Poisson−Boltzmann (PB) calculations were performed with the MEAD package (version 2.2.0).[77] Proton isomerism was considered by including the tautomeric forms of each titrable site.[75,78] The atomic charges and radii were adapted from the GROMOS96 43a1 force field[67] as described elsewhere.[79] The molecular surface was defined with a solvent probe of radius 1.4 Å and a Stern layer of 2.0 Å. The PB calculations consisted of finite difference calculations using a two-step focusing procedure[80] with grid spacings of 1.0 and 0.25 Å. The temperature was 300 K, the ionic strength was 0.1 mol/L, and the dielectric constants were 2 for the peptide and 80 for the solvent.

The Monte Carlo (MC) runs were performed with the program PETIT (version 1.4).[78,81] Each new set of protonation states was obtained after 10^5 MC steps, with each step consisting of a cycle of random choices of state (including tautomeric forms) for all sites followed by double site changes for pairs of sites with coupling above 2.0 pH units.

The constant-pH MD simulations were performed at the pH values: 3.0 (pH3), 4.0 (pH4), and 5.0 (pH5). In the low-pH range, the glutamic acid residues were titrated while the arginines were charged.

We did five constant-pH MD simulations 100 ns long for each pH value. The first 10 ns in each simulation was discarded, assuring that the system was equilibrated. Snapshots were saved every 1 ps.

Conformational Coordinates and Dissimilarity. We tested the following conformational coordinates: ϕ and ψ dihedrals, cosine and sine of the ϕ and ψ dihedrals, several sets of intramolecular C_α distances, and Cartesian coordinates of backbone atoms after translational and rotational least-squares fitting on a reference structure. In the latter case, we tested different choices of reference structure (see Results and Discussion), including a *central structure* chosen among the n sampled ones as the structure i minimizing the dispersion measure

$$D_i^2 = \frac{1}{n-1}\sum_{j=1}^{n} \mathrm{rmsd}_{ij}^{\,2} \qquad (1)$$

where rmsd_{ij} is the backbone rmsd between structures i and j (see below). This definition of central structure follows directly from analogy with the Euclidean case, for which the mean square deviation has the least value (the variance) when taken relative to the mean.[82]

Distances in conformation space were computed using the usual Euclidean definition, except for the ϕ/ψ dihedral space, where periodicity was additionally taken into account (which corresponds to computing the geodesic distance on the surface of a hypertorus). Due to this periodicity, the ϕ/ψ dihedral space is not strictly Euclidean and, therefore, was not used for PCA. The space using the sine and cosine of these dihedral angles has no such problem.

The dissimilarity of a pair of conformations was measured as the root-mean-square deviation (rmsd) of their atomic Cartesian coordinates obtained after translational and rotational least-squares fitting;[83] only the backbone atoms were used for both the fitting and the rmsd calculation. The rmsd quantifies how well the atoms of two molecular structures can be superimposed and fulfills the requirements of a metric,[84,85] meaning that it is expected to preserve the relative proximity among all the conformations. The use of distances in conformation space as dissimilarity measures is discussed in the Results and Discussion.

Principal Component Analysis. As mentioned in the Introduction, PCA is a standard method that can simplify a data set by reducing its dimensionality while keeping most of the variation; a discussion of its mathematical and computational aspects can be found elsewhere.[19,20] In short, PCA consists of a linear transformation that changes a set of (possibly) correlated variables into a set of uncorrelated variables called principal

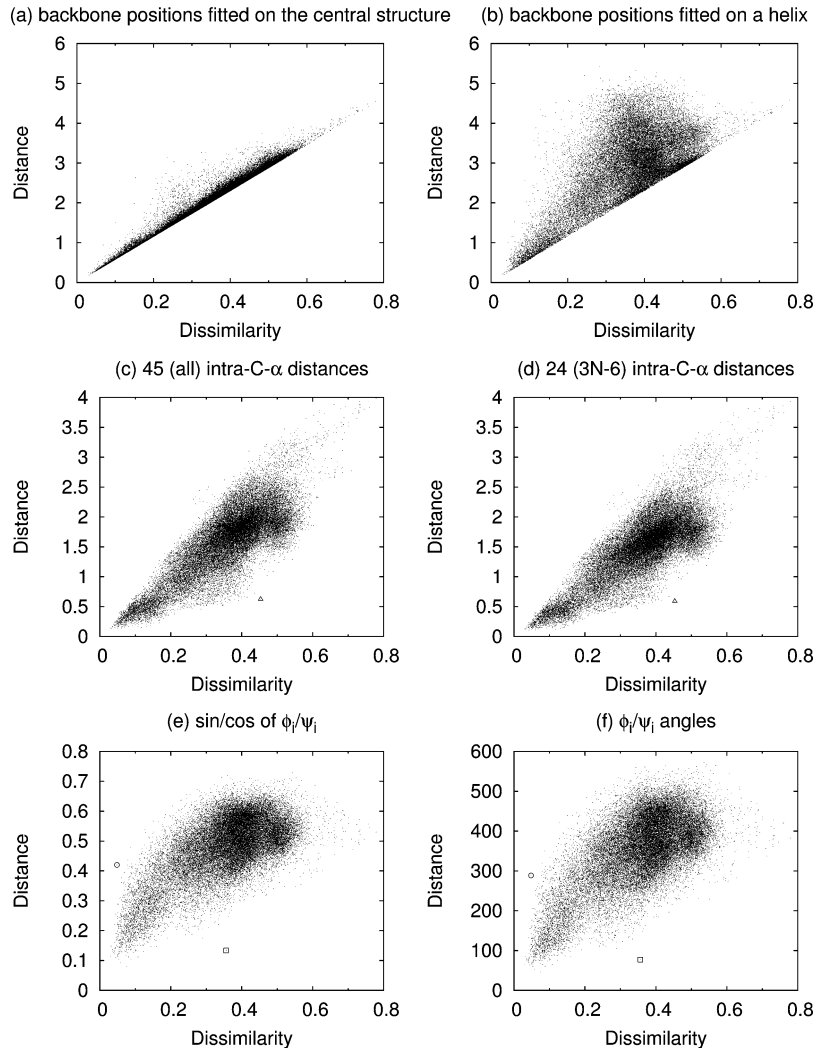

Figure 1. Comparison of the ability of different sets of coordinates in correlating distance and conformational dissimilarity for the peptide RE5 at pH 3 (data from one of the simulations). The degree of dissimilarity is assessed by the rmsd of backbone atomic positions, after fitting (see text). The points marked (triangle, square, and circle) correspond to examples of bad correlations whose pairs of structures are represented in Figures 2, 3, and 4.

components (PCs). The first PC accounts for as much of the variation in the data as possible, and each succeeding PC accounts for as much of the remaining variation as possible, while being uncorrelated with the previous ones. The strongest correlations in the data set tend to be captured by the first PCs. PCA can be used for reducing dimensionality by retaining only a few of the first PCs, hopefully without much loss of information. PCA is algebraically equivalent to the diagonalization of the covariance matrix: each PC is a coordinate whose axis is one of the eigenvectors, and its variance is given by the corresponding eigenvalue.

In the present work, PCA was applied to several sets of conformational coordinates (see previous subsection), using the g_covar and g_aneig programs of the GROMACS package,[64−66] version 3.2.1.

Density Estimation. The probability density function in the representation space was estimated from the simulation data using a Gaussian kernel estimator.[86] The kernel bandwidth (h) —in this case, the Gaussian's standard deviation —was chosen as[86]

$$h = \sigma\left(\frac{4}{n(2d+1)}\right)^{1/(d+4)} \quad (2)$$

where σ^2 is the average marginal variance, n is the number of data points, and d is the number of dimensions. This procedure formally defines a probability density function $P(\mathbf{r})$ at all points $\mathbf{r}$ in the d-dimensional space. In practice, values of $P(\mathbf{r})$ were stored for the position of each data point (i.e., conformation) and also for the nodes of a d-dimensional uniform grid.

The spread of points in each data set, whose number is around 450 000, imposes an upper limit of around 8 dimensions for the probability density, above which the usual problems of high-dimensional estimation are expected to occur.[86] The computer memory required for the multidimensional grids was also a major limitation, meaning that the total grid size had to be kept within a tractable size. In practice, the grid mesh size was chosen on the basis of the available computer memory, taking values of 0.5 Å for 2D and 3D estimations and between 2.5 and 3.5 Å for 7D and 8D (see Results and Discussion). Furthermore, some peripheral points were excluded for pH3 and pH7 to reduce the total grid size.

Energy Landscapes. The energy surface was calculated from $P(\mathbf{r})$ as

$$E(\mathbf{r}) = -RT \ln \frac{P(\mathbf{r})}{P_{\max}} \quad (3)$$

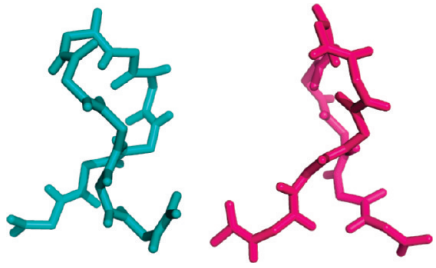

Figure 2. Pair of RE5 structures resembling each other's mirror image and corresponding to the triangle points depicted in Figure 1c and d.

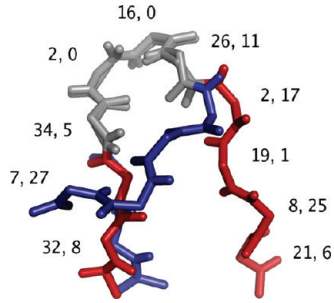

Figure 3. Pair of RE5 structures corresponding to the square points depicted in Figure 1e and f. For a better comparison, the structures were superimposed. Well-superimposed regions are shown in gray; red and blue represent the remaining regions of the different chains. The difference in degrees between corresponding dihedrals is represented as " $|\Delta\phi|,|\Delta\psi|$ " near the respective location.

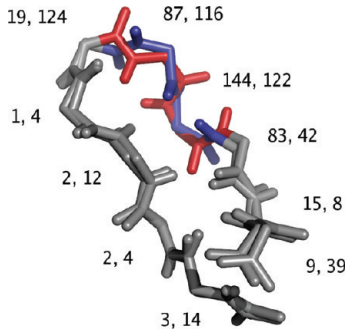

Figure 4. Pair of RE5 structures corresponding to the circle points depicted in Figure 1e and f. See the caption of Figure 3 for further details.

where P_{max} is the maximum of the probability density function, $P(\mathbf{r})$. This corresponds to (arbitrarily) assign a zero energy to the maximum of probability density. For simplicity, we will hereafter refer to $E(\mathbf{r})$ as "energy", although it is obviously a conditional free energy, as noted in the Introduction.

The energy landscapes were analyzed by determining the energy minima and respective basins. A *basin* is here defined as the set of all conformations whose steepest descent path along the energy hypersurface leads to a particular minimum, establishing a one-to-one relation between basins and minima;[87,88] in short, all conformations that "fall" to the same minimum belong to the same basin. In the present case, we computed the steepest descent paths on the grid used for the density estimation (see previous subsection), with each conformation inheriting the path of its corresponding grid cell.

Landscape regions with $E > 2.5\ RT$ were then discarded, and each of the resulting cutoff basins was defined as a conformation *class*. High-energy conformations structurally deviate from the respective minima and may even share greater similarity with other high-energy conformations from surrounding basins. Since our main objective is a conformational characterization of the peptide (see Results and Discussion), we excluded these high-energy conformations to obtain more homogeneous classes. Furthermore, these high-energy regions of the landscape are more likely to be affected by sampling problems and, thus, less reliable than the low-energy regions.

Transitions and Sampling. The temporal evolution of the system over the landscape basins was analyzed using concepts from the theory of Markov processes.[89] If that temporal evolution is Markovian, we should have

$$p_i(t + \Delta t) = \sum_j p_j(t)\, p_{ji} \tag{4}$$

where $p_i(t)$ is the probability that the system is in basin i at time t, p_{ij} is the time-independent probability that a system previously in basin i is found in basin j after a time Δt (here 1 ps, the time between snapshots), and the sum is taken over all basins j (including $j = i$). In the asymptotic limit, one should get $p_i(\infty) = p_i^{eq}$, with the latter being the equilibrium probability of finding the system in basin i.

After assigning the sampled conformation snapshots to their respective basins (previous subsection), the probabilities p_i^{eq} and p_{ij} were directly computed from the basin-labeled trajectories (as relative frequencies). Using these p_{ij} probabilities, eq 4 was then iterated until convergence, which was found to be always independent of the initial probabilities, $p_i(0)$. The resulting asymptotic $p_i(\infty)$ values were then compared with the previously computed p_i^{eq}. Since the computed $p_i(\infty)$ depend only on the p_{ij} values, this comparison is a consistency test between the observed p_i^{eq} and p_{ij} values. A good agreement would be a strong indication that the system shows no sampling problems (e.g., trapping in some basins) and is properly equilibrated over the visited basins. This test is more general than the common but unnecessarily restrictive check on detailed balance.[89] In any case, it should be noted that this simple consistency check does not imply that the system has a strictly Markovian behavior, which would require checking more specific conditions (e.g., the Chapman−Kolmogorov equation).[89]

Results and Discussion

Conformational Coordinates and Dissimilarity. Although the choice of conformational coordinates involves some arbitrariness, the distance between two conformations in the resulting representation space should reflect their structural dissimilarity. Therefore, the appropriateness of a set of conformational coordinates could in principle be inferred from the relation between its associated distance and structural dissimilarity. However, the concept of structural dissimilarity is intrinsically subjective, and not surprisingly, the same quantity is often regarded as both a dissimilarity measure and a distance in some conformation space. In practice, we have to select a measure that seems physically reasonable and use it as an operational definition of structural dissimilarity. As indicated in Materials and Methods, the present work adopts the backbone rmsd to measure structural dissimilarity. The eventual adequacy of other dissimilarity measures is also discussed below.

Figure 1 shows the relation between structural dissimilarity (rmsd) and distance in different conformation spaces of RE5. Clearly, the best correlation is observed in Figure 1a, where the conformational coordinates are the backbone positions after fitting on the central structure defined in Materials and Methods. Using instead an α-helix as reference drastically worsens the

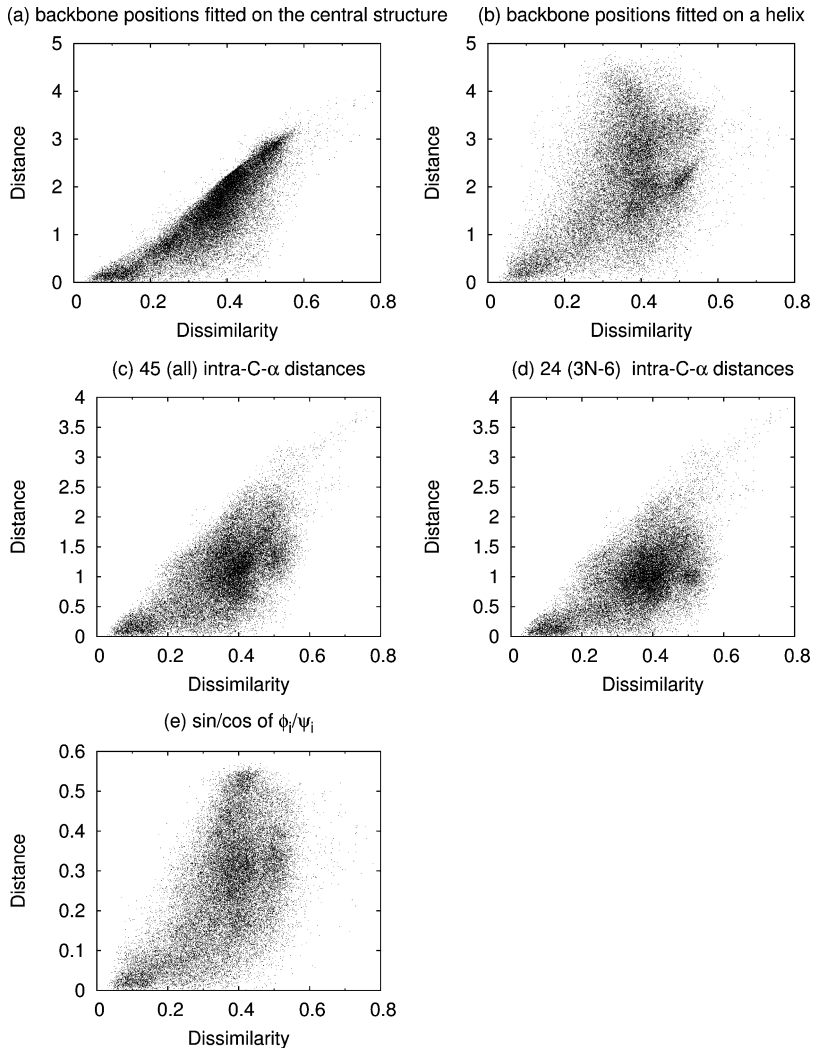

Figure 5. The ability of the first two PCs obtained from different sets of coordinates in correlating distance and conformational dissimilarity for RE5 at pH 3 (data from one of the simulations).

correlation (Figure 1b), illustrating very well the sensitivity to the choice of reference structure; the correlation is equally poor when using a coil structure and somewhat better with a β-like structure (results not shown). This sensitivity follows from the fact that the superposition of two very different structures can have several distinct solutions with very close rmsd values, making the choice of reference structure a crucial issue.[90] The central structure is clearly a very good choice of reference structure.

Figure 1c and d shows the results obtained using the distance in intra-C_α space, revealing a poor correlation with dissimilarity. As noted by Cohen and Sternberg,[83] this coordinate space has the disadvantage of being invariant under reflection. This invariance may be a minor problem for folded protein structures but can become much more serious for flexible peptides. This is illustrated in Figure 2, which shows a pair of structures that resemble each other's mirror reflection.

As expected, the use of intra-C_α distances approximates the structures of this type of pairs (triangle points in Figure 1c and d). Inspection of Figure 1 indicates that intra-C_α distances may approximate dissimilar structures due to invariance under reflection but do not seem to decrease the proximity among them. Overall, the present results suggest that intra-C_α coordinates are not a good choice for small flexible molecules. Likewise, the distances in intra-C_α space are not adequate measures of structural dissimilarity.

Figure 1e and f shows the results obtained using the distance in two different dihedral spaces. The use of a dihedral space avoids one of the problems of using atomic positions after fitting, where high local similarity can be hidden by overall dissimilarity.[36,90] On the other hand, a single drastic torsion about one bond may significantly alter the overall conformation while a second torsion about another bond can largely cancel the effect of the first one.[36,90−92] Therefore, there is no simple relationship between dihedral distances and dissimilarity in terms of atomic positions (as measured by the rmsd), as clearly seen in the plots. Figure 3 illustrates a case in which two considerably different structures are considered to be close when using dihedrals as coordinates.

Although in this example the differences between corresponding dihedrals are quite low (less than 35°), these small differences accumulate, orientating the chains in different directions and originating rather different structures. Conversely, dihedral space may also distance similar conformations, as illustrated in Figure 4. Despite the large differences observed in this example between some of the corresponding dihedrals, these differences cancel each other, bringing the chains to similar structures.

These results do not necessarily imply that dihedral spaces are inappropriate to represent molecular conformations, but simply that, as noted above, they do not correlate well with the dissimilarity of the relative atomic positions. Since the confor-

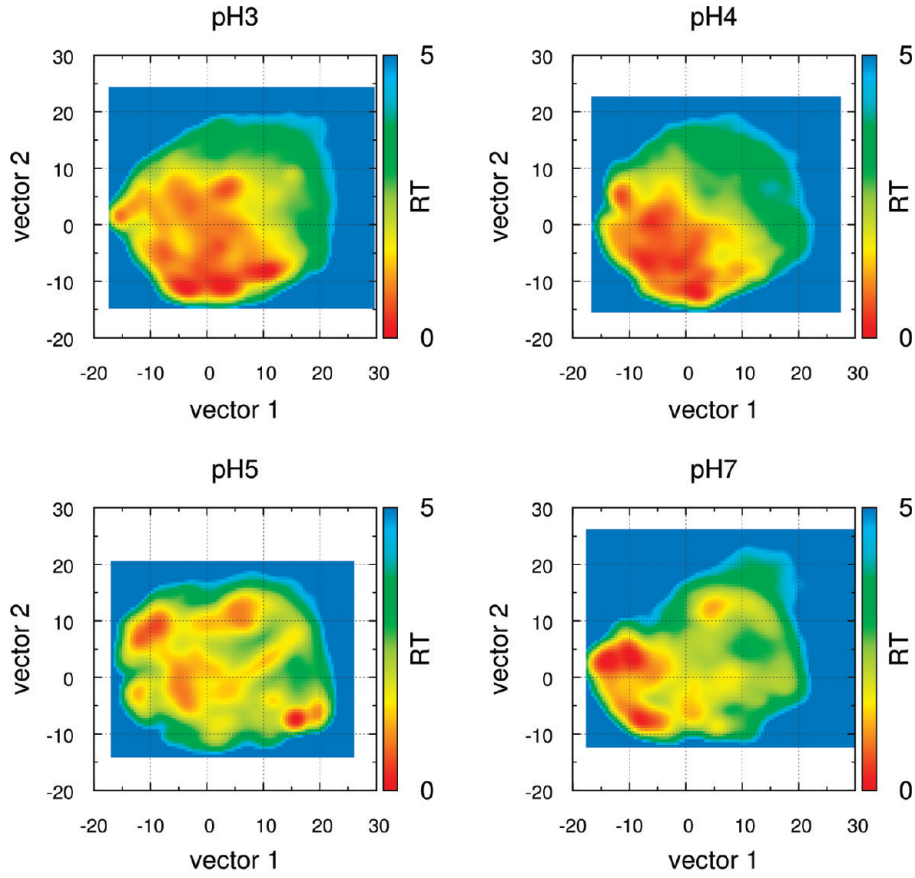

Figure 6. Energy landscapes at different pH values using the first two PCs.

mational changes effectively taking place in a protein occur mainly by means of torsional rotations, a dihedral space is expected to reflect the topography of the energy landscape in a much more direct way than an atomic-positions space, as is, indeed, observed for a tetrapeptide.[12] On the other hand, biological function is often dependent on the relative positions of the atoms involved in some kind of molecular interaction, regardless of the landscape details. For example, if a structure of the type seen in Figure 4 is required for binding to another molecule, the two depicted structures will contribute to the overall amount of bound form, regardless of the energy barrier between them. In contrast, even though the two structures in Figure 3 may be in the same landscape valley, they are very unlikely to be involved in the same type of interaction. This suggests that dihedral distances may be good dissimilarity measures when studying the kinetics of conformational change (e.g., folding/unfolding/misfolding), whereas the rmsd may be better suited to study the thermodynamics of biological interactions. Therefore, our choice of the rsmd as a dissimilarity measure corresponds to adopting the latter view.

Since we want to subject our data to PCA and retain only some PCs as coordinates, it is important to examine also how the different conformation spaces are affected by this procedure. Since PCA corresponds to a translation and rotation of the coordinate axes,[19] distances are preserved after PCA, meaning that the ordinate values in Figure 1 represent also the post-PCA distances when retaining all PCs (except for plot f, since the periodic dihedral space cannot be directly subjected to PCA, as already noted). The effect of retaining only two PCs can be seen in the corresponding Figure 5. Clearly, the loss of dimensions worsens the correlation between distance and dissimilarity, emphasizing the importance of the choice of coordinates that represent the data, before applying PCA. The best correlation is again obtained when using the backbone positions after fitting on the central structure.

Given the above results and considerations, the conformational coordinates selected to be used for PCA were the Cartesian coordinates of the backbone atoms obtained after fitting on the central structure defined in Materials and Methods.

Number of Principal Components. Using the conformational coordinates selected in the previous subsection, we performed PCA on the sets of structures sampled from the MD simulations. We started by investigating the number of PCs required to adequately discriminate different RE5 structures.

Figure 6 shows the energy landscapes computed from the probability density projected on the plane spanned by the first two eigenvectors, for different pH values.

To allow for a direct comparison of the landscapes, we performed the PCA using the trajectories for all pH values, after which they were treated separately; this ensures that all landscapes are represented in the same space. This type of two-dimensional (2D) representation is quite common and is often expected to give a first, even if rough, insight into the major conformation classes of the system. However, this is far from true in the case of RE5, as illustrated in Figure 7. Instead, one finds considerably (or even strikingly) different structures at essentially the same positions in the resulting representation space, showing that the discrimination of structures is actually quite poor. The problem is that, although the first two PCs are the pair capturing most of the variation, this does not necessarily imply that the major groups of structures are arranged in a nonoverlapping manner in this 2D subspace. As a consequence, the projection of the whole landscape on this 2D plane will in general mix different groups of conformations in an unpredict-

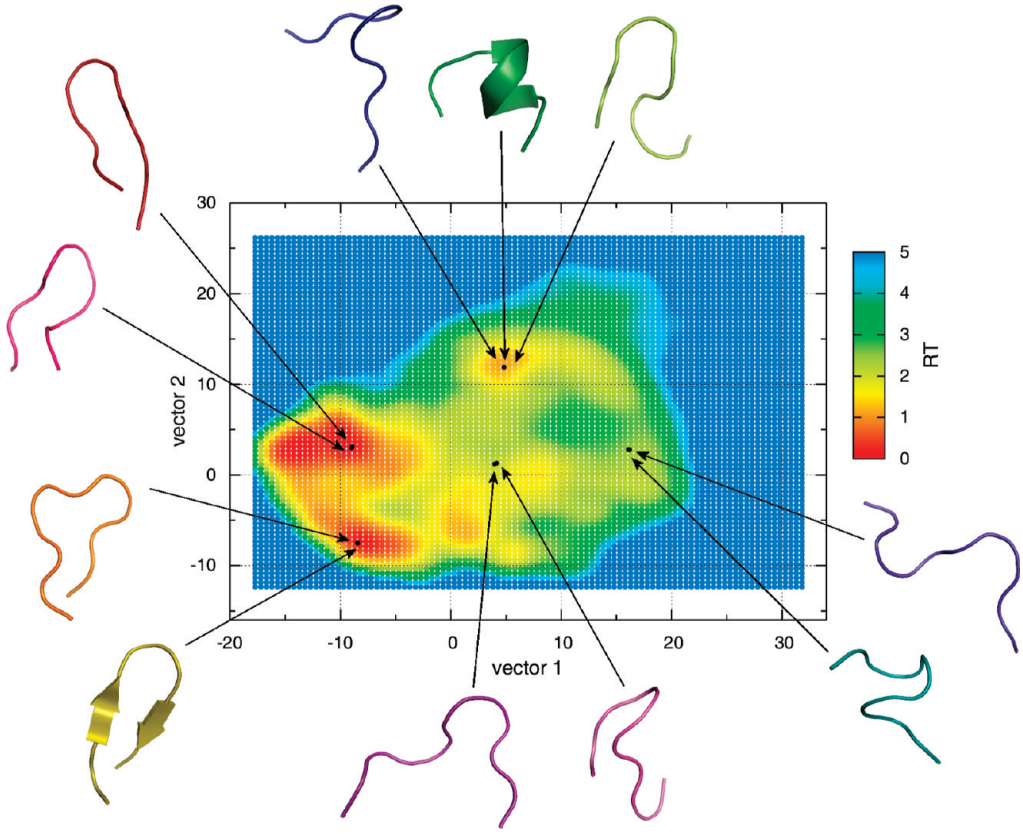

Figure 7. Energy landscape representation as a function of the first two PCs for pH7. Some structures are represented pointing to the respective positions in the 2D space.

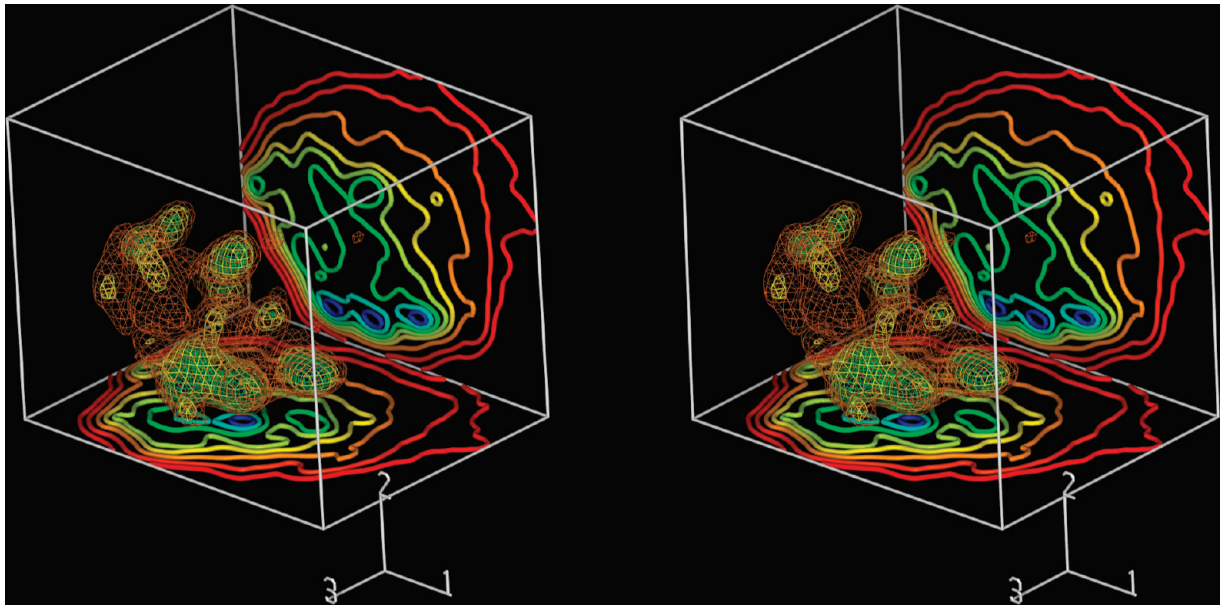

Figure 8. Tridimensional and bidimensional energy landscapes obtained for pH3, in cross-view stereo. The contours correspond to the energy levels 0.2, 0.5, 1, 1.5, and 2 *RT* (2D and 3D contours); and 3, 5, and 8 *RT* (2D contours). The color gradient goes from blue (lower energies) to red (higher energies).

able way. In some cases, different groups may contribute to create a valley where none of them had a particularly high density. In other cases, a group may be split[31,88] due to the cumulative effect of other groups upon different regions. Figure 8 illustrates the problems associated with the projection from 3D to 2D, showing the fusion of some of the 3D valleys.

It should also be noted that the common approach of using several 2D landscapes combining the first PCs (as seen in Figure 8) cannot replace the 3D landscape because a joint probability density cannot be obtained from the lower-dimensional marginal densities.[82] Thus, a given set of 2D landscapes corresponds, in general, to many different 3D landscapes, meaning that a 3D "interpretation" of the set is not possible.

The problems just discussed are obviously not restricted to the mapping from 3D to 2D, affecting any reduction in the number of dimensions: the 3D landscape loses some features

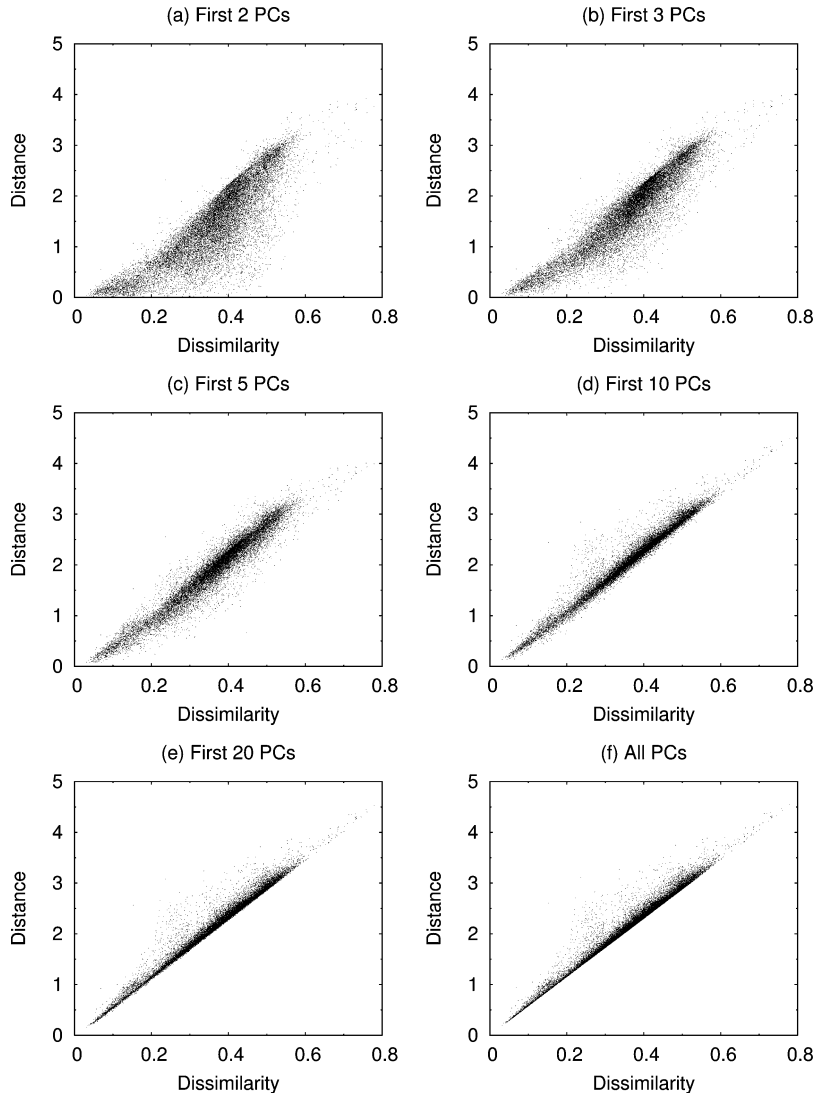

Figure 9. Number of PCs needed to maintain the proximity between RE5 conformations, for pH3.

of the 4D landscape, which in turns loses some features of the 5D landscape, and so on. For example, the presence of dissimilar conformations in the same landscape region, observed at 2D (Figure 7), is also observed at higher dimensions (results not shown). Therefore, the representation space for RE5 needs more dimensions to properly discriminate its alternative conformations. This need for further dimensions can also be seen by again comparing structural dissimilarity and distance in representation space. As illustrated in Figure 9, this correlation improves with the number of PCs, as expected, until reaching the best correlation using all PCs (note that Figure 9f is identical to Figure 1a). The plots corresponding to fewer PCs display an agglomeration of points below the diagonal, corresponding to the approximation of dissimilar conformations. This agglomeration has essentially vanished when using the 10D distances, which are already very similar to those in the all-PCs space.

These results show that low-dimensional energy landscapes of RE5 correlate very poorly with conformational dissimilarity, being useless to identify conformation classes. In the next subsection, we select a minimum number of dimensions using a simple heuristic approach balancing dissimilarity preservation and computational tractability.

Conformation Classes from Energy Landscapes. Given the above considerations, energy landscapes were rebuilt using a different approach. Instead of performing PCA on the collection

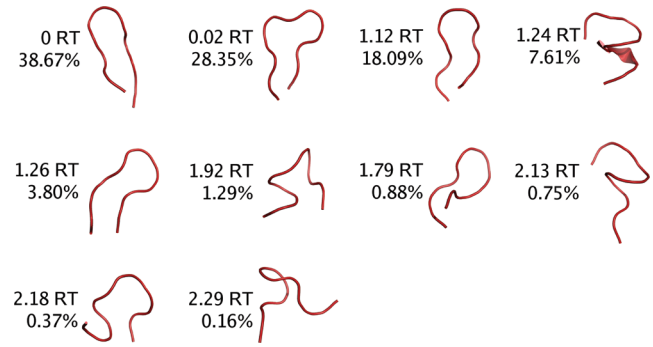

Figure 10. Conformation classes obtained from the standard MD simulations (pH7). The depicted conformation corresponds to the energy minimum in that class, whose energy value (in *RT* units) is shown on the left. The percentage of structures grouped in each class is also indicated.

of all sampled conformations, each pH was treated separately. The conformations sampled at each pH value were subjected to a separate PCA using as coordinates the backbone atomic positions obtained after fitting on the central structure (see Materials and Methods) determined for that pH value. Although this precludes the direct comparison of the landscapes (whose coordinates become different), it leads to a higher concentration of variation in the lowest PCs for each pH value, because each

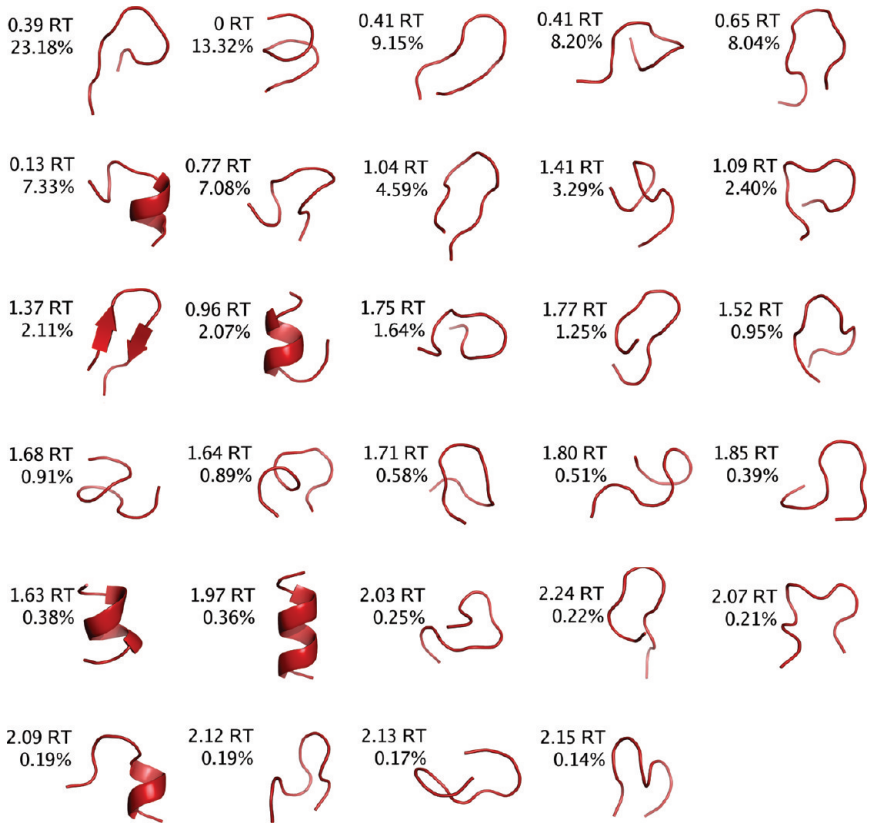

Figure 11. Conformation classes obtained from the pH3 simulations. See the caption of Figure 10 for further details.

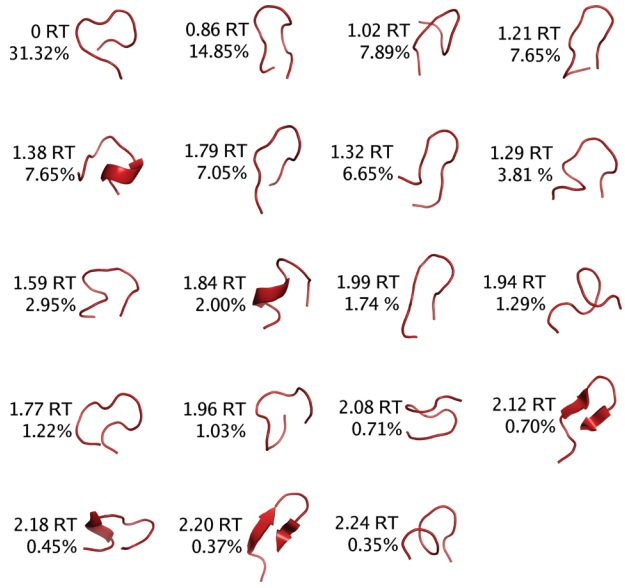

Figure 12. Conformation classes obtained from the pH4 simulations. See the caption of Figure 10 for further details.

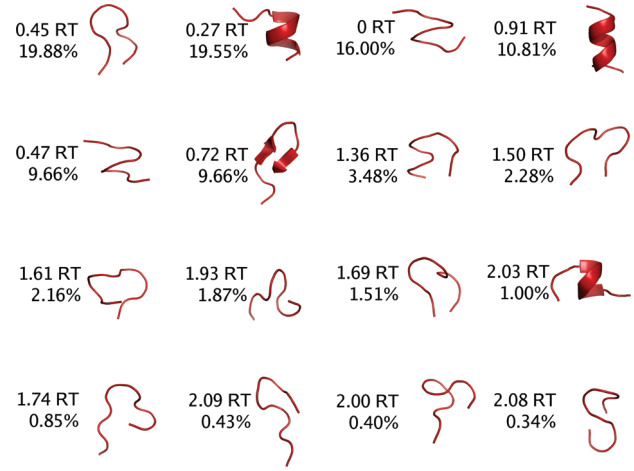

Figure 13. Conformation classes obtained from the pH5 simulations. See the caption of Figure 10 for further details.

consecutive PC maximizes the remaining variation (within the orthogonality constraint).[19] The probability density function and the corresponding landscape were then calculated using the first seven (pH3 and pH4) or eight (pH5 and pH7) PCs. This number of dimensions results from a compromise between preserving structure dissimilarity and the possibility of running the necessary computations (see Materials and Methods); in all cases, it captures at least 80% of the total variance. After the conformation classes were determined from the landscape as already described, the conformations in each class were visually inspected to detect eventual conformational heterogeneities such as those in Figure 7. The classes presented below correspond to homogeneous or almost homogeneous groups of structures. Classes with less than 100 elements were discarded.

Analysis of the data from the standard MD simulations (pH7) resulted in 10 conformation classes whose minima are shown in Figure 10. It is clear that the peptide displays a high conformational diversity, adopting some β-like and helix-like structures, as well as other less canonical conformations. The predominance of β-like structures at pH 7 is in agreement with experimental results from circular dichroism studies of arginy-laspartic acid repeats.[58]

The results obtained from the constant-pH MD simulations are shown in Figures 11, 12, and 13. The peptide shows a very high conformational diversity at low pH values, displaying again structures resembling β-hairpins and helices, although with

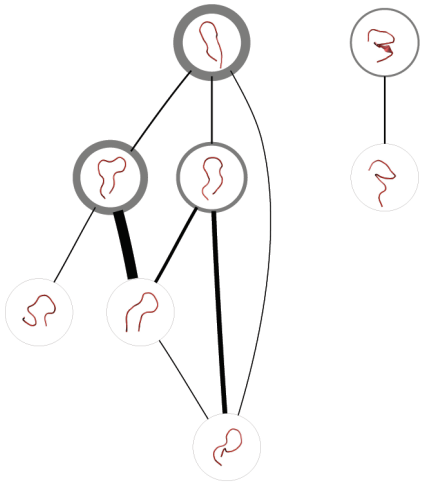

Figure 14. Observed transitions between classes for pH7. Each class is represented by its minimum structure enclosed in a circle, whose line width is a linear function of the number of conformations in the class. The line width of the connections between classes is a linear function of the total (forward + backward) number of transitions. Graph representations were done with Graphviz (version 2.18, http://www.graphviz.org).

different ratios (e.g., helical forms predominate for pH5). It is noteworthy that the global minimum does not correspond to the more populated class in the cases of pH3 and pH5, showing that the "width" of the basin is a relevant factor.

Interestingly, more conformation classes are identified for the constant-pH MD data than for the standard MD data. When protonation takes place, each conformation region may be differently stabilized by alternative charge configurations, which may originate a larger number of distinct classes in that region (if the stabilizing charge configurations are populated at the considered pH, of course). In other words, the charge alternatives may eventually lead to more ways of stabilizing a structure and, thus, lead to more minima/classes.

The relation between the conformation classes can be partly inferred from the direct transitions observed between them during the simulations. Those transitions are depicted in the graphs of Figures 14 and 15. Two groups of interconverting classes are seen for pH5 and pH7, corresponding broadly to β-like and helix-like structures, although some less canonical structures are also observed. A single group is observed for pH4, although the more β-like structures can interconvert more directly. For pH3, a separated group of less canonical structures is seen in addition to the groups of β-like and helix-like structures. We note that some significantly populated classes are absent from the graphs (e.g., the helical class at the top-right corner of Figure 13), indicating that they cannot reach other classes within the time step between snapshots (1 ps); such transitions need additional time to overcome energy barriers, eventually passing through intermediate basins above the selected cutoff (2.5 *RT*). Naturally, this could be analyzed by examining all transitions between the complete (noncutoff) basins, with groups of interconnected basins being regarded as kinetic "super basins".[88] However, as discussed before, our

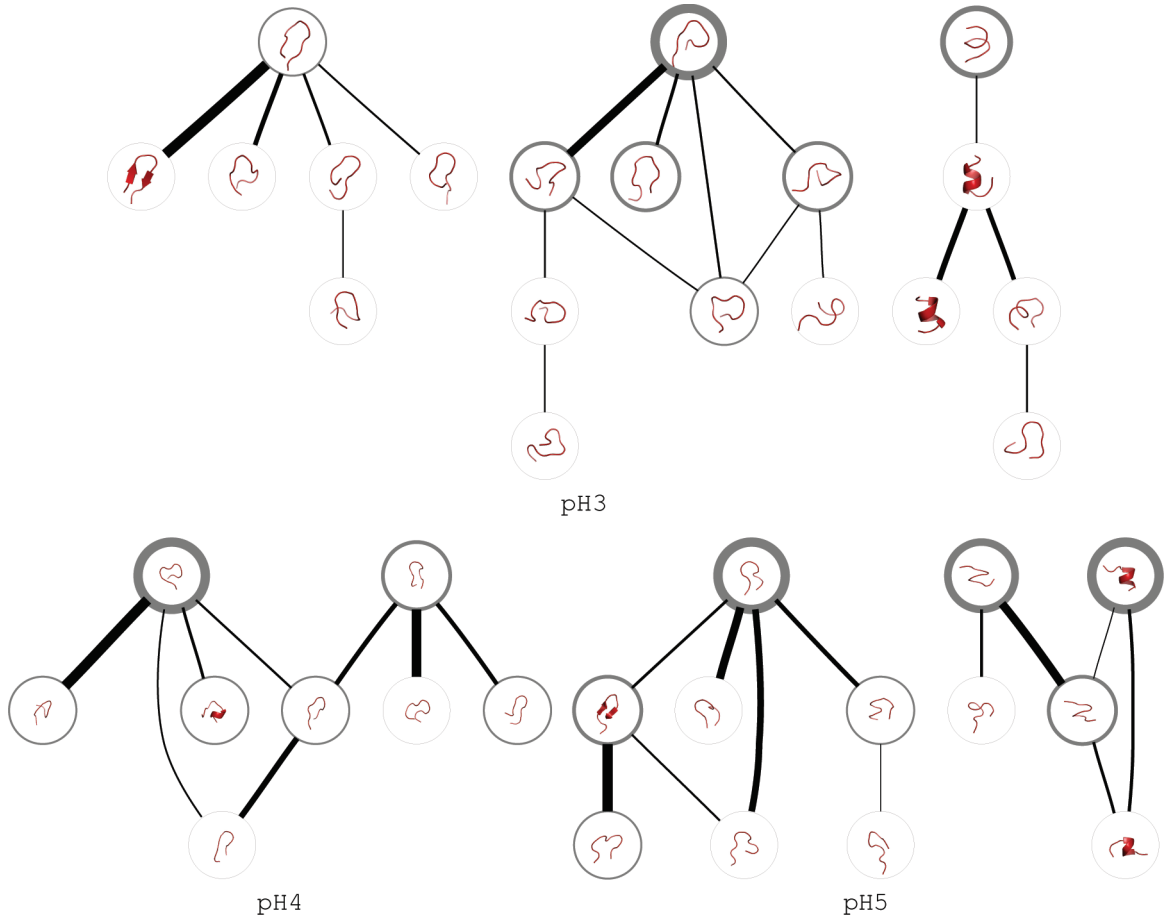

Figure 15. Observed transitions between classes for pH3, pH4, and pH5. To avoid cluttering, connections corresponding to rare transitions (below the 0.05 quantile) are not included for pH3 and pH4. See the caption of Figure 14 for further details.

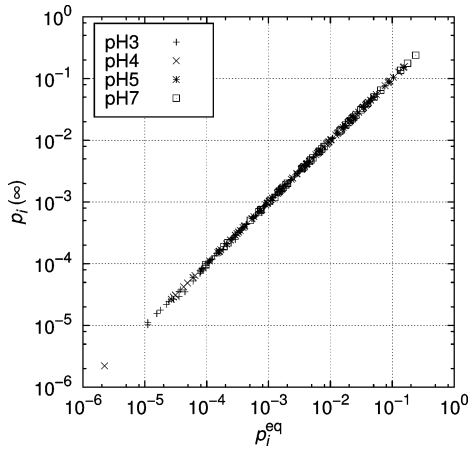

Figure 16. Comparison between the basin equilibrium probabilities computed directly from the trajectories, p_i^{eq}, and the asymptotic ones obtained from eq 4, $p_i(\infty)$.

purpose in this work is not to study the kinetics of conformational change but, rather, to make a structural characterization in terms of homogeneous conformation classes. Hence, Figures 14 and 15 simply are intended to show that the conformation classes produced by our multidimensional method are related in a consistent way.

As a test on the quality of sampling, we checked for the consistency of the populations of the basins and the transitions between them, as described in Materials and Methods; since this intends to check the overall sampling, no cutoff was considered. As shown in Figure 16, the populations directly computed from the trajectories, p_i^{eq}, were always in excellent agreement with the asymptotic ones derived from eq 4, even for very low values. This is a very strong indication that the systems show no sampling problems (e.g., trapping in some basins) and are well-equilibrated over the visited basins.

Overall, these results show that PCA-derived multidimensional energy landscapes can be used to successfully identify homogeneous conformation classes, producing a consistent conformational characterization of RE5. It is also interesting to note that the large conformational diversity exhibited by this peptide depends on pH in a nontrivial way, involving the (dis)appearance of different classes rather than a simple shift of their populations. This suggests that we should avoid thinking in terms of "intrinsically stable" classes whose populations are affected by the environment conditions (e.g., pH) because things are clearly more complex.

Concluding Remarks

Several interesting conclusions follow from the present study in terms of the comparison of molecular conformations, their mapping on a conformation space, the use of PCA to derive a lower dimensional representation space, and the identification of distinct conformation classes reflecting the density in that representation space.

The study shows that different conformation spaces/coordinates map the level of resemblance of molecular conformations in a significantly (and sometimes drastically) different way (Figure 1). The finding of this kind of differences is obviously not new,[12,29,83,93] but the direct comparison between distances in conformation space and a pairwise dissimilarity measure is particularly effective in displaying the lack of agreement. This also points to the fundamental conceptual role played by the choice of a particular dissimilarity measure.

The properties of the Cartesian space obtained after least-squares fitting are found to depend markedly on the choice of the reference structure. This space preserves very well the structural dissimilarity when the fit is done on the *central structure* that minimizes the dispersion measure of eq 1, whereas the choice of an arbitrary reference structure can give rather poor agreement. This points to the need to carefully select a reference structure before fitting, an issue overlooked by many studies.

Also very significant are the properties of the conformation spaces obtained after PCA by retaining only the first PCs. It is found that the low-dimensional spaces are unable to preserve structural dissimilarity and local conformational homogeneity. Only higher-dimensional spaces (7D or 8D in the present study) are able to satisfy these requirements. This indicates that the 2D plots displaying the distribution on the first PCs (first PC versus second PC, first PC versus third PC, etc.), commonly found in many conformational studies, should be interpreted with great care because even obvious agglomerations in 2D space may correspond to highly heterogeneous regions (as in Figure 7).

After transformation of the 7D or 8D probability densities into the corresponding multidimensional energy landscapes, the use of a simple energy cutoff is shown to produce a consistent set of major conformation classes. These classes are distinct among themselves, display internal conformational homogeneity, and interconvert in a manner that is coherent with their conformational characteristics.

Overall, the methodology proposed in the present study provides an efficient approach to identify the major conformation classes of a molecule in a way that directly reflects its energetics in a multidimensional landscape or, equivalently, its density of states in a multidimensional conformation space. Despite some remaining arbitrariness in the initial choice of a dissimilarity measure, we think that this approach retains more physical content than the more heuristic and widespread use of standard clustering methods.

Acknowledgment. We thank Cláudio M. Soares and Miguel Machuqueiro for fruitful discussions and Paulo J. Martel for providing us with a modified version of the g_rms tool of GROMACS. We acknowledge financial support from Fundação para a Ciência e a Tecnologia, Portugal, through grants POCTI/BME/45810/2002 and SFRH/BD/23506/2005.

References and Notes

(1) Becker, O. M. Conformational analysis. In *Computational Biochemistry and Biophysics*; Becker, O. M., MacKerell, A. D., Roux, B., Watanabe, M., Eds.; Marcel Dekker: New York, 2001.
(2) Shea, J. E.; Brooks, C. L. *Annu. Rev. Phys. Chem.* **2001**, *52*, 499–535.
(3) Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644–649.
(4) Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. *J. Chem. Theory Comput.* **2007**, *3*, 2312–2334.
(5) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins* **1995**, *21*, 167–195.
(6) Dill, K. A.; Bromberg, S.; Yue, K.; Fiebig, K. M.; Yee, D. P.; Thomas, P. D.; Chan, H. S. *Protein Sci.* **1995**, *4*, 561–602.
(7) Karplus, M.; Šali, A. *Curr. Opin. Struct. Biol.* **1995**, *5*, 58–73.
(8) Thirumalai, D.; Woodson, S. A. *Acc. Chem. Res.* **1996**, *29*, 433–439.
(9) Shakhnovich, E. *Curr. Opin. Struct. Biol.* **1997**, *7*, 29–40.
(10) Finkelstein, A. V. *Curr. Opin. Struct. Biol.* **1997**, *7*, 60–71.
(11) Frauenfelder, H.; Leeson, D. T. *Nat. Struct. Mol. Biol.* **1998**, *5*, 757–759.
(12) Becker, O. M. *Proteins* **1997**, *27*, 213–226.
(13) Czerminski, R.; Elber, R. *J. Chem. Phys.* **1990**, *92*, 5580–5601.
(14) Pettitt, B. M.; Karplus, M. *J. Phys. Chem.* **1988**, *92*, 3994–3997.

(15) Hermans, J.; Anderson, A. G. Microfolding: Use of simulations to study peptide/protein conformational equilibria. In *Theoretical Biochemistry & Molecular Biophysics*; Beveridge, D. L., Lavery, R., Eds.; Adenine Press: Schenectady, NY, 1990.
(16) Takada, S.; Luthey-Schulten, Z.; Wolynes, P. G. *J. Chem. Phys.* **1999**, *110*, 11616–11629.
(17) Kitao, A.; Go, N. *Curr. Opin. Struct. Biol.* **1999**, *9*, 164–169.
(18) Hayward, S.; Go, N. *Annu. Rev. Phys. Chem.* **1995**, *46*, 223–250.
(19) Jolliffe, I. T. *Principal Component Analysis*, 2nd ed.; Springer-Verlag: New York, 2002.
(20) Rencher, A. C. *Methods of Multivariate Analysis*; 2nd ed.; Wiley-Interscience: New York, 2002.
(21) Kitao, A.; Hirata, F.; Gō, N. *Chem. Phys.* **1991**, *158*, 447–472.
(22) García, A. E. *Phys. Rev. Lett.* **1992**, *68*, 2696–2699.
(23) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. *Proteins* **1993**, *17*, 412–425.
(24) Hayward, S.; Kitao, A.; Hirata, F.; Gō, N. *J. Mol. Biol.* **1993**, *234*, 1207–1217.
(25) García, A. E.; Sanbonmatsu, K. Y. *Proteins* **2001**, *42*, 345–354.
(26) Higo, J.; Ito, N.; Kuroda, M.; Ono, S.; Nakajima, N.; Nakamura, H. *Protein Sci.* **2001**, *10*, 1160–1171.
(27) Sanbonmatsu, K. Y.; García, A. E. *Proteins* **2002**, *46*, 225–234.
(28) Ikeda, K.; Galzitskaya, O. V.; Nakamura, H.; Higo, J. *J. Comput. Chem.* **2003**, *24*, 310–318.
(29) Mu, Y.; Nguyen, P. H.; Stock, G. *Proteins* **2005**, *58*, 45–52.
(30) Nguyen, P. H.; Stock, G.; Mittag, E.; Hu, C.-K.; Li, M. S. *Proteins* **2005**, *61*, 795–808.
(31) Maisuradze, G. G.; Leitner, D. M. *Proteins* **2007**, *67*, 569–578.
(32) Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. *J. Mol. Biol.* **2009**, *385*, 312–329.
(33) Cox, T. F.; Cox, M. A. A. *Multidimensional Scaling*, 2nd ed.; Chapman & Hall/CRC: Boca Raton, 2001.
(34) Becker, O. M. *J. Comput. Chem.* **1998**, *19*, 1255–1267.
(35) Elmaci, N.; Berry, R. S. *J. Chem. Phys.* **1999**, *110*, 10606–10622.
(36) Hamprecht, F. A.; Peter, C.; Daura, X.; Thiel, W.; van Gunsteren, W. F. *J. Chem. Phys.* **2001**, *114*, 2079–2089.
(37) Levy, Y.; Becker, O. M. *J. Chem. Phys.* **2001**, *114*, 993–1009.
(38) Levy, Y.; Jortner, J.; Becker, O. M. *J. Chem. Phys.* **2001**, *115*, 10533–10547.
(39) Prompers, J. J.; Brüschweiler, R. *Proteins* **2002**, *46*, 177–189.
(40) Prompers, J. J.; Brüschweiler, R. *J. Am. Chem. Soc.* **2002**, *124*, 4522–4534.
(41) Abseher, R.; Nilges, M. *J. Mol. Biol.* **1998**, *279*, 911–920.
(42) Gallicchio, E.; Andrec, M.; Felts, A. K.; Levy, R. M. *J. Phys. Chem. B* **2005**, *109*, 6722–6731.
(43) Mitomo, D.; Nakamura, H. K.; Ikeda, K.; Yamagishi, A.; Higo, J. *Proteins* **2006**, *64*, 883–894.
(44) van Aalten, D. M. F.; de Groot, B. L.; Findlay, J. B. C.; Berendsen, H. J. C.; Amadei, A. *J. Comput. Chem.* **1997**, *18*, 169–181.
(45) Horstmann, M.; Ehses, P.; Schweimer, K.; Steinert, M.; Kamphausen, T.; Fischer, G.; Hacker, J.; Rösch, P.; Faber, C. *Biochemistry* **2006**, *45*, 12303–12311.
(46) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. *J. Chem. Phys.* **2007**, *126*, 244111+.
(47) Caves, L. S. D.; Evanseck, J. D.; Karplus, M. *Protein Sci.* **1998**, *7*, 649–666.
(48) García, A. E.; Blumenfeld, R.; Hummer, G.; Krumhansl, J. A. *Physica D* **1997**, *107*, 225–239.
(49) Komatsuzaki, T.; Hoshino, K.; Matsunaga, Y.; Rylance, G. J.; Johnston, R. L.; Wales, D. J. *J. Chem. Phys.* **2005**, *122*, 084714+.
(50) Altis, A.; Otten, M.; Nguyen, P. H.; Hegger, R.; Stock, G. *J. Chem. Phys.* **2008**, *128*, 245102+.
(51) Ma, K.; Clancy, E. L.; Zhang, Y.; Ray, D. G.; Wollenberg, K.; Zagorski, M. G. *J. Am. Chem. Soc.* **1999**, *121*, 8698–8706.
(52) Dluhy, R. A.; Shanmukh, S.; Leapard, J. B.; Krüger, P.; Baatz, J. E. *Biophys. J.* **2003**, *85*, 2417–2429.
(53) Abedini, A.; Raleigh, D. P. *Biochemistry* **2005**, *44*, 16284–16291.
(54) Khandogin, J.; Chen, J.; Brooks, C. L. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 18546–18550.
(55) Swietnicki, W.; Petersen, R.; Gambetti, P.; Surewicz, W. K. *J. Biol. Chem.* **1997**, *272*, 27517–27520.
(56) Uversky, V. N.; Li, J.; Fink, A. L. *J. Biol. Chem.* **2001**, *276*, 10737–10744.
(57) Levi-Strauss, M.; Carroll, M. C.; Steinmetz, M.; Meo, T. *Science* **1988**, *240*, 201–204.
(58) Pelsue, S.; Agris, P. F. *J. Protein Chem.* **1994**, *13*, 401–408.
(59) Assier, E.; Bouzinba-Segard, H.; Stolzeenberg, M.-C.; Stephens, R.; Bardos, J.; Freemont, P.; Charron, D.; Trowsdale, J.; Rich, T. *Gene* **1999**, *230*, 145–154.
(60) Hartmann, A. M.; Nayler, O.; Schwaiger, F. W.; Obermeier, A.; Stamm, S. *Mol. Biol. Cell* **1999**, *10*, 3909–3926.
(61) Yanagisawa, H.; Bundo, M.; Miyashita, T.; Okamura-Oho, Y.; Tadokoro, K.; Tokunaga, K.; Yamada, M. *Hum. Mol. Genet.* **2000**, *9*, 1433–1442.
(62) Trembley, J. H.; Hu, D.; Slaughter, C. A.; Lahti, J. M.; Kidd, V. J. *J. Biol. Chem.* **2003**, *278*, 2265–2270.
(63) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. *J. Chem. Phys.* **2002**, *117*, 4184–4200.
(64) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43–56.
(65) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306–317.
(66) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
(67) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: the GROMOS96 Manual and User Guide*; vdf Hochschulverlag AG an der ETH Zürich: Zürich, 1996.
(68) Hermans, J.; Berendsen, H. J. C.; van Gunsteren, W. F.; Postma, J. P. M. *Biopolymers* **1984**, *23*, 1513–1518.
(69) Barker, J. A.; Watts, R. O. *Mol. Phys.* **1973**, *26*, 789–792.
(70) Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1994**, *100*, 3169–3174.
(71) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
(72) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463–1472.
(73) Machuqueiro, M.; Baptista, A. M. *J. Phys. Chem. B* **2006**, *110*, 2927–2933.
(74) Machuqueiro, M.; Baptista, A. M. *Biophys. J.* **2007**, *92*, 1836–1845.
(75) Machuqueiro, M.; Baptista, A. M. *Proteins* **2008**, *72*, 289–298.
(76) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451–5459.
(77) Bashford, D.; Gerwert, K. *J. Mol. Biol.* **1992**, *224*, 473–486.
(78) Baptista, A. M.; Soares, C. M. *J. Phys. Chem. B* **2001**, *105*, 293–309.
(79) Teixeira, V. H.; Cunha, C. A.; Machuqueiro, M.; Oliveira, A. S. F.; Victor, B. L.; Soares, C. M.; Baptista, A. M. *J. Phys. Chem. B* **2005**, *109*, 14691–14706.
(80) Gilson, M. K.; Sharp, K. A.; Honig, B. H. *J. Comput. Chem.* **1987**, *9*, 327–335.
(81) Baptista, A. M.; Martel, P. J.; Soares, C. M. *Biophys. J.* **1999**, *76*, 2978–2998.
(82) Mood, A. M.; Graybill, F. A.; Boes, D. C. *Introduction to the Theory of Statistics*, 3rd ed.; McGraw-Hill: New York, 1974.
(83) Cohen, F. E.; Sternberg, M. J. E. *J. Mol. Biol.* **1980**, *138*, 321–333.
(84) Kaindl, K.; Steipe, B. *Acta Crystallogr.* **1997**, *A53*, 809.
(85) Steipe, B. *Acta Crystallogr.* **2002**, *A58*, 506.
(86) Silverman, B. W. *Density Estimation for Statistcs and Data Analysis*; Chapman and Hall: London, 1986.
(87) Stillinger, F. H.; Weber, T. A. *Science* **1984**, *225*, 983–989.
(88) Becker, O. M.; Karplus, M. *J. Chem. Phys.* **1997**, *106*, 1495–1517.
(89) van Kampen, N. G. *Stochastic Processes in Physics and Chemistry*; North-Holland: Amsterdam, 2007.
(90) Murray-Rust, P.; Raftery, J. *J. Mol. Graph.* **1985**, *3*, 50–59.
(91) Karpen, M. E.; Tobias, D. J.; Brooks, C. L. *Biochemistry* **1993**, *32*, 412–420.
(92) Torda, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **1994**, *15*, 1331–1340.
(93) Wallin, S.; Farwer, J.; Bastolla, U. *Proteins* **2003**, *50*, 144–157.

JP902991U