Hindawi Publishing Corporation
Advances in Physical Chemistry
Volume 2016, Article ID 3240674, 10 pages
http://dx.doi.org/10.1155/2016/3240674

Research Article

Energy Landscape of Pentapeptides in a Higher-Order (ϕ, ψ) Conformational Subspace

Karim M. ElSawy[1,2]

[1] *York Centre for Complex Systems Analysis (YCCSA), University of York, York YO10 5GE, UK*
[2] *Department of Chemistry, College of Science, Qassim University, Buraydah 52571, Saudi Arabia*

Correspondence should be addressed to Karim M. ElSawy; km.elsawy@qu.edu.sa

Received 8 February 2016; Accepted 4 April 2016

Academic Editor: Dennis Salahub

The potential energy landscape of pentapeptides was mapped in a collective coordinate principal conformational subspace derived from principal component analysis of a nonredundant representative set of protein structures from the PDB. Three pentapeptide sequences that are known to be distinct in terms of their secondary structure characteristics, $(Ala)_5$, $(Gly)_5$, and Val.Asn.Thr.Phe.Val, were considered. Partitioning the landscapes into different energy valleys allowed for calculation of the relative propensities of the peptide secondary structures in a statistical mechanical framework. The distribution of the observed conformations of pentapeptide data showed good correspondence to the topology of the energy landscape of the $(Ala)_5$ sequence where, in accord with reported trends, the α-helix showed a predominant propensity at 298 K. The topography of the landscapes indicates that the stabilization of the α-helix in the $(Ala)_5$ sequence is enthalpic in nature while entropic factors are important for stabilization of the β-sheet in the Val.Asn.Thr.Phe.Val sequence. The results indicate that local interactions within small pentapeptide segments can lead to conformational preference of one secondary structure over the other where account of conformational entropy is important in order to reveal such preference. The method, therefore, can provide critical structural information for *ab initio* protein folding methods.

1. Introduction

Our understanding of sequence-structure relationships of proteins is increasing in both depth and breadth [1, 2]. Deeper insight is provided by detailed studies of protein folding and stability via sequence mutations and/or *de novo* protein design [3]. Breadth has been gained through a great tradition of serendipitous discovery [4, 5], supported in recent years by the systematised sampling strategies of the structural genomics programmes [6]. Our ability to predict protein structure from sequence is contingent on characteristics of the sequence: if a sequence possesses even weak homology to proteins of known structure, then a variety of methods can exploit this information to infer a structural model at a given level of detail with some (statistical) confidence [7]. If no homologies exist, our empirical knowledge of protein structure (and interactions) can yield good predictions for many (as yet) small proteins. However, pure ("physics-based") *ab initio* predictions of the native fold of a protein from sequence information alone are still formidably difficult. There are two principal challenges that must be met in order to faithfully model the thermodynamics and kinetics of protein folding: the ability to accurately compute the energetics of solvated proteins and the difficulty of generating structural models which bridge the unfolded and folded states. The latter difficulty arises from the very large number of degrees of freedom in proteins that makes a systematic search of conformational space computationally intractable and thus enforces a conformational sampling approach [8].

The conformational space of a single amino acid in a peptide can be effectively described by the backbone (ϕ, ψ) torsion angles and the sterically "allowed" space represented on the 2D "Ramachandran plot" [9]. Can information on single (ϕ, ψ) pairs be combined to represent the conformational space of polypeptides? Flory's isolated pair hypothesis [10] assumes that the conformational state of each pair of (ϕ, ψ) in a polypeptide is independent of the adjacent pairs and therefore the number of conformational states available to

a polypeptide segment is simply the product of possible conformational states of the (ϕ, ψ) pairs of individual residues. The validity of this model for polypeptides and proteins is debatable [11, 12] as it appears to overestimate the number of conformational states and thus implies that nonlocal effects—manifesting in terms of coupling between (ϕ, ψ) pairs—are important. However, for longer polypeptides, with increasing number of (ϕ, ψ) pairs, a Ramachandran-type conformational analysis is not generally possible due to the high dimensionality. As a consequence, the predominant focus of conformational analysis of polypeptides has been in terms of single (ϕ, ψ) pairs [13, 14].

A more complete understanding of sequence-structure relationships in polypeptides requires us to move beyond statistical/geometrical considerations to include physical interactions [15]. From the perspective of energetics, it is the interactions between different residues within the polypeptide chain which distinguish the behaviour of one sequence from another. In recent years the energy landscape concept has emerged as a unifying language for experimentalists and theorists [16–18]. The energy landscape describes how the energy changes within the conformational space. It is the topographical and topological features of the energy landscape which determine the thermodynamic and kinetic properties of the system. Knowledge of the protein energy landscape is therefore indispensable for studying a variety of important system properties such as stable states, dynamics, and folding/unfolding kinetics [19]. Several simulation studies have addressed the problem of evaluating the energy landscape for polypeptides [15, 20–22]. A common element in such studies is the detection of the energy minima and their connectivity based on extensive sampling of the energy landscape.

The relatively small number of folds adopted by experimentally determined protein structures suggests that native folds are confined to a low dimensional manifold (subspace) of the nominally high dimensional protein fold space [23, 24]. Also, it appears that the intrinsic conformational dynamics of a given protein fold may be confined to a relatively low dimensional space [25–27]. In support of this view, a multivariate analysis of the geometry of polypeptide segments from protein crystal structures has revealed that the conformational space available over multiple ("higher-order") (ϕ, ψ) segments is dramatically restricted over that expected from a consideration of individual (ϕ, ψ) pairs [28]. The origin of this low intrinsic dimensionality lies in the coupling of (ϕ, ψ) pairs within the segments as a consequence of steric packing constraints along the polypeptide chain.

In a previous work [29], we employed a similar multivariate approach to construct a low dimensional principal conformational subspace of single strand dinucleotide fragments from duplex DNA crystal structures. The low dimensionality was a consequence of the use of collective coordinates representing the coupled displacements of backbone torsion angles. The collective coordinates were used to map the underlying energy surface within the conformational subspace using an empirical interaction potential. A key aspect of this approach is that the energy landscape underlies a relevant region of conformational space representing the statistical distribution of the observed crystal structures. The method proved successful in predicting a new stable conformer of DNA and relative conformational propensities of different dinucleotide fragments and describing the dynamical behaviour of an oligomeric DNA sequence.

In this paper, we build on the approach developed in our previous work [29] in order to calculate the potential energy landscape of pentapeptides within the principal conformational space of known crystal structures [28]. Our aim is to gain a better physical understanding of the important sequence-structure relationships in polypeptides within a statistical mechanical framework. To achieve this, we generate a principal conformational subspace from a representative set of pentapeptide fragments extracted from structures from the protein databank. The resulting collective coordinates are then used to map the underlying potential energy surface for selected pentapeptide sequences. The topography and topology of the resulting energy surfaces are characterised and then used to derive the thermodynamic distribution of different bound conformational states. As a proof of principle we consider three pentapeptide sequences that are distinct in terms of their secondary structure characteristics: $(Ala)_5$, $(Gly)_5$, and Val.Asn.Thr.Phe.Val. The $(Ala)_5$ sequence represents the canonical sequence for sequence-structure studies; $(Gly)_5$ represents the extreme case of removing local steric hindrance along the polypeptide chain; and the Val.Asn.Thr.Phe.Val sequence is context dependent [30] as it appears once as part of an α-helix and once as part of a β-sheet. For the sequences considered, our model predicts relative conformational propensities that are comparable to those derived from the statistics of the sample crystal structure dataset. Thus, we present a framework for computing sequence-specific thermodynamic and kinetic properties of polypeptide fragments which we hope will have wider utility in studies concerned with the relationship of protein sequence, structure, and dynamics.

2. Methods

2.1. Data Acquisition and Preparation. Representative structures from the Protein Data Bank were created using PDBSELECT [31]. The dataset corresponds to structures of high resolution ($\leq$1.0 Å) with <25% sequence identity. The structures were disassembled into pentapeptide segments using a five residue sliding windows over contiguous chain segments with a lag of one residue. The descriptors of the backbone conformation for each segment were described by 10 variables representing the (ϕ, ψ) torsion angle pairs of each residue. A data matrix (D) was constructed from 9607 observations arranged row wise. In order to eliminate any potential bias towards conformers with high frequencies, matrix (D) was reduced to matrix (M) which comprised equally represented conformers. To achieve this, segments were categorized based on their secondary structure, in the full-length peptide chain, and the frequency of each distinct secondary structure category was determined. Secondary structure classification was carried out using the Sanders algorithm [32] as implemented in PROMOTIF [33]. Secondary structure categories with frequency

less than 10 were discarded. For each distinct secondary structure category, ten conformers were selected at random and the corresponding (ϕ, ψ) torsions pairs were fed into matrix M resulting in 1572 conformers.

2.2. Principal Component Analysis (PCA). In order to reduce the dimensionality of the conformational space of the pentapeptide segments, the data matrix (M) was subjected to principal component analysis (PCA) [34]. PCA and multidimensional scaling (MDS) are widely used as methods for reducing the dimensionality of multidimensional molecular structural data [35–37].

To avoid problems, such as periodicity and nonlinearity, in the calculation of variance for angular data [38], the data in the columns of the data matrix M representing individual torsion angles were adjusted in an iterative fashion such that differences from their mean values lie in the range −180 to +180° using the Bio3d [39] package in R 3.1 [40]. This transformed the data matrix M into a new data matrix N. Principal component analysis was then conducted on matrix N using the Bio3d package [39]. The principal components (or PCs) are mutually orthogonal collective variables that maximally describe the sample variance. The elements of the PCs (eigenvectors) are the coefficients of the linear combinations of the 10 torsion angles from which they are derived. Thus, the PCs (analogous to the normal coordinates derived from harmonic vibration analysis [41]) can be considered as collective coordinates for describing the pentapeptide conformational distribution via displacements from the mean torsion angles of the data matrix. The corresponding eigenvalues represent the variance of the conformational distribution along each of the PCs. The extent to which a given observation $\boldsymbol{\varphi}_i$ lies along the kth principal component $\mathbf{a}_k$ is given by the projection $z_{i,k} = \mathbf{a}_k \cdot (\boldsymbol{\varphi}_i - \overline{\boldsymbol{\varphi}})'$. These projections (or scores) may be used for visualising the distribution of observations in the principal conformational subspace (PCS).

Despite the duality of PCA and MDS in terms of the scores of the data matrix [42, 43], PCA offers advantages over MDS (as used by Sims et al. [28]) in view of the aims of this study. The principal components (eigenvectors) can be characterised directly as concerted atomic displacements [44] such that the magnitude and sign of the coefficients of the linear combinations reflect the correlations of the (ϕ, ψ) variables. Displacements along the principal components generate trajectories representing the structural displacements spanning the PCS. Also, PCA results in an analytical basis set (the principal components) which is used in the construction of structures within the PCS as required for the mapping of energy surface (see below).

2.3. Mapping the Potential Energy Surface (PES). The potential energy surface (PES) of a pentapeptide segment within the PCS was mapped via systematic energy evaluation on a grid defined by discrete points along the first three PCs. The coordinates of a grid point $\boldsymbol{\Omega}_i = (z_{i,1}, z_{i,2}, z_{i,3})$ correspond to a set of torsion angles, $\boldsymbol{\theta}_i = \overline{\boldsymbol{\varphi}} + \sum_{k=1}^{k=3} z_{i,k}\mathbf{a}_k$. Therefore, $\boldsymbol{\theta}_i$ represents the geometrical construction of structures within the {PC1, PC2, PC3} subspace (whose projections along higher principal components are initially zero). Cartesian coordinates for the full pentapeptide fragment were reconstructed from these torsion angles supplemented by other standard internal coordinates (bond lengths and angles). The potential energy of the structures was calculated using the extended-atom CHARMM27 force field [45, 46]. No nonbonded interaction truncation was performed. A distance-dependent dielectric constant ($\varepsilon_r = 4r_{ij}$, where r_{ij} is the interatomic distance in Å) [47] was used for the electrostatic terms. Use of distance-dependent dielectric constant is a computationally cheap approximation for the damping of the electrostatic interaction by the solvent. Incorporating the solvent effect into the energetics of the PES using more advanced methods such as Generalized Born [48] or APBS [49] is, however, computationally prohibitive due to the large number of structures used to construct the PES. The PES was mapped in 2D slices along PC3. A grid resolution of 20° in the subspace was used over the range −300° to 300° with respect to an origin corresponding to the mean structure. These search limits were chosen so as to encompass the range of scores spanned by the projected data matrix.

At each grid point, the backbone torsion angles were restrained to their desired values using a force constant of 100 kcal mol^{-1} degree^{-2} and the system was energy minimised using 200 steps of steepest descent followed by 2000 steps of the ABNR method [45]. The side chain conformation was then generated using SCWRL 3.0 program [50] and the system was minimised using 200 steps of steepest descent.

As a consequence of the discretization, the structure and energy of any point in the PCS are considered to be the structure and energy corresponding to the closest grid point within the resolution limit of the grid. Characterisation of the features of the PES and calculation of the thermodynamic properties for the different conformational states were conducted as described previously [29]. For the sake of clarity, we describe it briefly in here: the PES within the PCS was partitioned into a set of distinct energy valleys as follows: starting from a local minimum, a set of 3D isoenergetic contours are generated at increasing discrete energy levels. An energy valley is defined by the isoenergetic surface contained within the contour level preceding the one which encapsulates another energy minimum. The energy valley, therefore, extends up to the (saddle) point which separates it from other minima. This process continues until no further minima are detected. The volume of each energy valley was calculated using the convex hull algorithm [51] as implemented in MATLAB 7.1. The fractional population of an energy valley ν can be estimated from the ratio $Q_\nu / \sum_{\nu=1}^{6} Q_\nu$ where Q_ν is the NVT partition function given by

$$Q_{\mathrm{NVT}} = \sum_i g_i \exp\left(\frac{-E_i}{k_B T}\right). \quad (1)$$

The energy levels, E_i, within each energy valley were delineated by a set of concentric isoenergetic closed surfaces in the 3D PCS corresponding to consecutive energies of 0.5 kcal mol^{-1}. The degeneracy g_i of the energy level i was estimated by the volume enclosed between the two surfaces corresponding to $E_i - 0.25$ and $E_i + 0.25$ kcal mol^{-1}. The (local)

partition function for each valley is then estimated via a summation over its discrete energy levels. Here, we use the term "energy level" in an *ad hoc* way since, in the classical regime, energy is continuous and not quantized. Enthalpic contribution to the fractional population of the PES valleys could be approximated by the relative energetics of their minima. The relative volumes of the valleys, however, are indicative of their relative entropic contribution [29, 52].

PCA, therefore, provides a basis set of linear collective coordinates describing the variance of observed structures within the PCS. The PCs allow for the construction of conformers within the PCS, interpolating and, optionally, extrapolating from the observed conformational distribution. Calculating the energy of conformers within the PCS results in a smooth continuous energy landscape. The collective nature of the basis set allows also for a straightforward definition of conformational reaction coordinates for complex biomolecules in contrast to the use of proxy coordinates ("order parameters"), which, are arbitrarily or heuristically chosen [53] and may not have a direct relationship to system geometry limiting interpretability and thus insight [54, 55]. Methods other than PCA could have also been used for reducing the dimensionality of the PES, such as principal coordinate analysis (PCoA) [56, 57] and locally nonlinear embedding (LLE) [58, 59]. In contrast to PCoA, PCA generates an orthogonal basis set that can be easily used as collective coordinates of the PES. Each of these collective coordinates is simply a linear combination of the original structural variables (e.g., ϕ, ψ torsion angles). Nonlinearity of the basis set generated by LLE, however, complicates the construction of the PES and makes interpretation of the PES topological features not straightforward.

3. Results and Discussion

3.1. A Principal Conformational Subspace for Pentapeptides. We generated a conformational space for pentapeptide segments extracted from protein crystal structures achieving a similar result to that of Sims et al. [28]. The collective coordinates (PCs) describing the conformational space naturally incorporate the coupling of (ϕ, ψ) pairs within the segment, which result in a reduction of dimensionality. The first three PCs were used as collective coordinates for defining a principal conformational subspace (PCS). Almost 60% of the total variance of the data matrix was captured by the first three PCs (34.5% along PC1, 18.4% along PC2, and 7.4% along PC3). Restricting the PCS to 3D is supported by the observation that the distribution of projections (scores) along higher components is virtually unimodal (data not shown). For geometrical characterisation of the collective coordinates refer to Supplementary Material available online at http://dx.doi.org/10.1155/2016/3240674. Projection of the pentapeptide data matrix (M) into the PCS reveals a clear separation of α and β secondary structures with intermediate conformational states lying in between (see Figure 1). The detailed description of the distribution of the projection of the data matrix within the PCS is discussed in relation to the features of the PES below.

3.2. The Potential Energy Surface. The PESs underlying the PCS were mapped for three selected pentapeptide sequences: $(Ala)_5$, $(Gly)_5$, and Val.Asn.Thr.Phe.Val. The $(Ala)_5$ sequence represents the canonical sequence for sequence-structure studies; $(Gly)_5$ represents the extreme case of removing local steric hindrance along the polypeptide chain, and the Val.Asn.Thr.Phe.Val sequence was chosen as a test case for the hypothesis that the secondary structure of Val.Asn.Thr.Phe.Val sequence is context dependent [30] as it appears once as part of an α-helix and once as part of a β-sheet.

The energy landscapes were assessed in terms of the correspondence of the distribution of observed conformers to the features of the underlying PES. The utility of the PES is demonstrated via computation of sequence-specific relative propensities of different secondary structure conformers. These propensities, computed on a statistical mechanical basis, are compared to the observed statistical distributions.

3.3. Relationship of Observed Conformational Distribution to the PES Features. Inspection of the projection of the observed conformational distribution (incorporating many different sequences) onto the PES of $(Ala)_5$ reveals that the data is confined to the low energy regions of the computed PES. Further, the positions of the predominant peaks of this distribution match well with the location of the energy valleys within that surface (see Figure 1(a)). Separation of the α-helix and the β-strand into different energy valleys is evident.

Within the alpha helical regions, it is interesting to note that points whose secondary structure assignment is uniformly alpha helix (i.e., H.H.H.H.H) are found in the minimum of the α-helix energy valley, while those points whose secondary structure assignment corresponds to loops or turns at peripheries (e.g., X.H.H.H.H.X; X.X.H.X.X, where X is [h, N, E]) are found displaced towards the valley's rim (see Figure 1(b)). On the other hand, for beta structures, the other peak of the observed distribution is slightly shifted from the position of the corresponding energy minimum which is most likely due to the neglect of secondary interactions (i.e., the opposing strand of a β-sheet) in our pentapeptide model. Also, we cannot discount some discrepancies arising from the deficiencies in our collective coordinate representation of the conformational space (only 60% of the total variance is captured within the PCS).

3.4. Sequence-Specific Differences of the PESs. Analysis of the PESs of the selected three pentapeptide sequences reveals that their detailed topographical and topological features are strongly dependent on sequence composition (see Figure 2). This observation confirms the importance of our physical approach to characterising sequence-structure relationships, in being able to tackle sequence-specific differences which are difficult to capture from statistical approaches, due to sampling issues. The differences in PES relate to the number of minima and their relative energetics, energy barriers to transition between minima, and the volume of the associated energy valleys. The number of minima increases for sequences with rotatable side chains; however, their connectivity assumes a single funnel shape which is consistent with

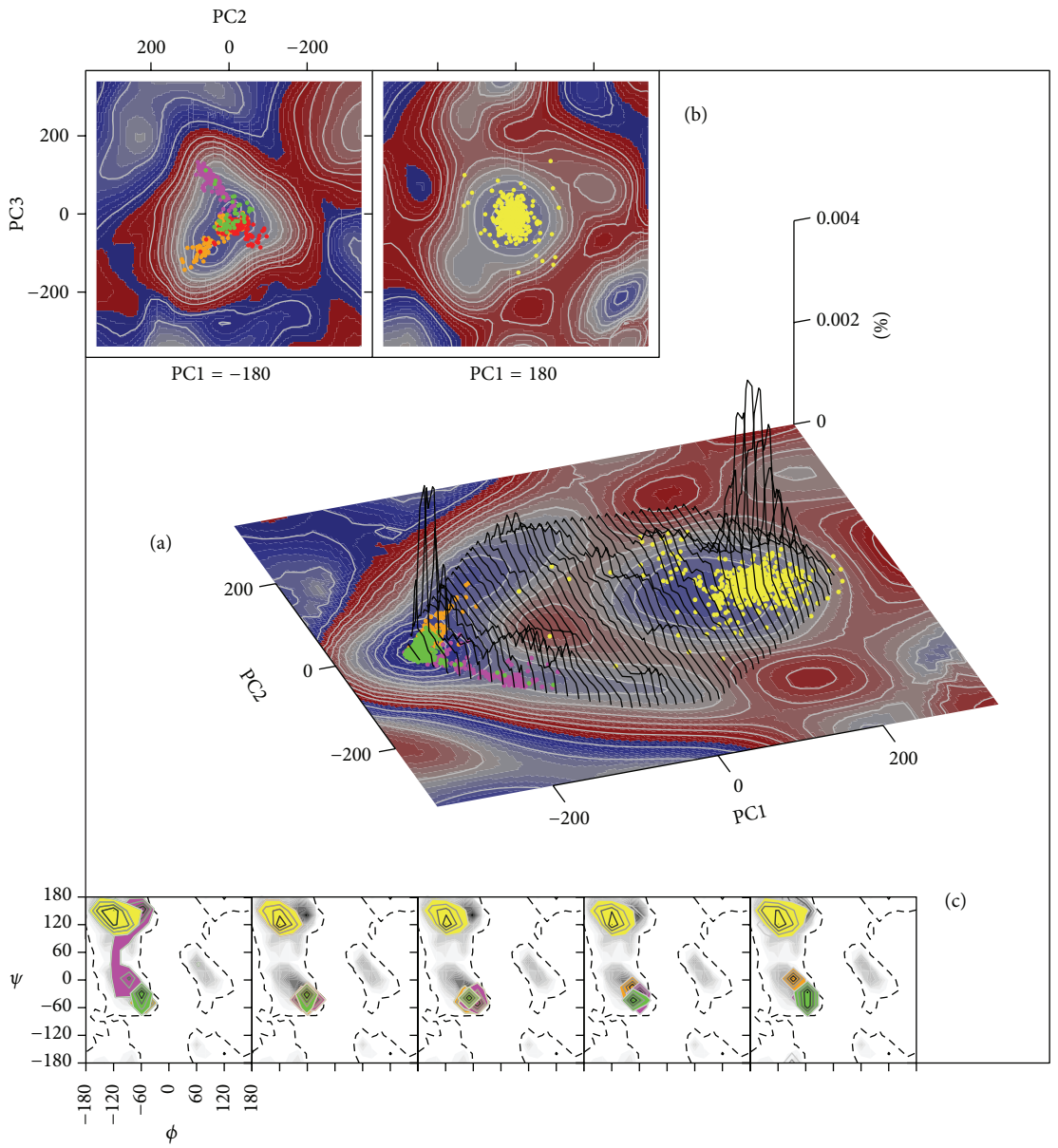

FIGURE 1: (a) The probability distribution (in percentage) of X-ray crystal structures data superposed on the potential energy slice in the PC1-PC2 plane. (b) Slices of the potential energy surface in PC1-PC3 and PC2-PC3 planes. In (a) and (b), energy is offset relative to the global minimum and the colour ramp changes linearly from dark blue (lowest energy; 0 kcal/mole) to red (30 kcal/mole) in 1 kcal/mole increment. (c) Ramachandran plots of selected secondary structure classes of the X-ray crystal structures dataset. Selected secondary structure classes are coloured in the three panels as follows; α-helix (H.H.H.H.H) in green, h.H.H.H.H in brown, H.H.H.H.h in yellow, and H.H.H.h.T in red while β-strand is in yellow (per residue: H indicates helix and h indicates beginning of a helix, while T indicates a turn).

the shape of the protein folding landscape [60, 61] (see Figure 2 top panel). On the other hand, the relative energetics of the minima and the volumes of the associated valleys play an important role in understanding the sequence-specific thermodynamic properties of the peptide system (see below).

3.5. Distribution of the Conformational States within Potential Energy Funnels. Stratification of the disconnectivity trees of the three selected sequences into different energy intervals helps to highlight the sequence-specific distribution of conformational states over the energy funnel. The densities of the (ϕ, ψ) states down the potential energy surfaces are shown in the inset Ramachandran plots in Figure 2 (top row).

At the highest energy band of the energy funnels (Figure 2 top row; top inset Ramachandran plot), the $(Gly)_5$ sequence shows the most discontinuous distribution of states unlike the Val.Asn.Thr.Phe.Val sequence which shows a continuous distribution ranging from the β-sheet region to the α-helix region. The distribution of the conformational states of the $(Ala)_5$ sequence is quite striking (and unique in these few

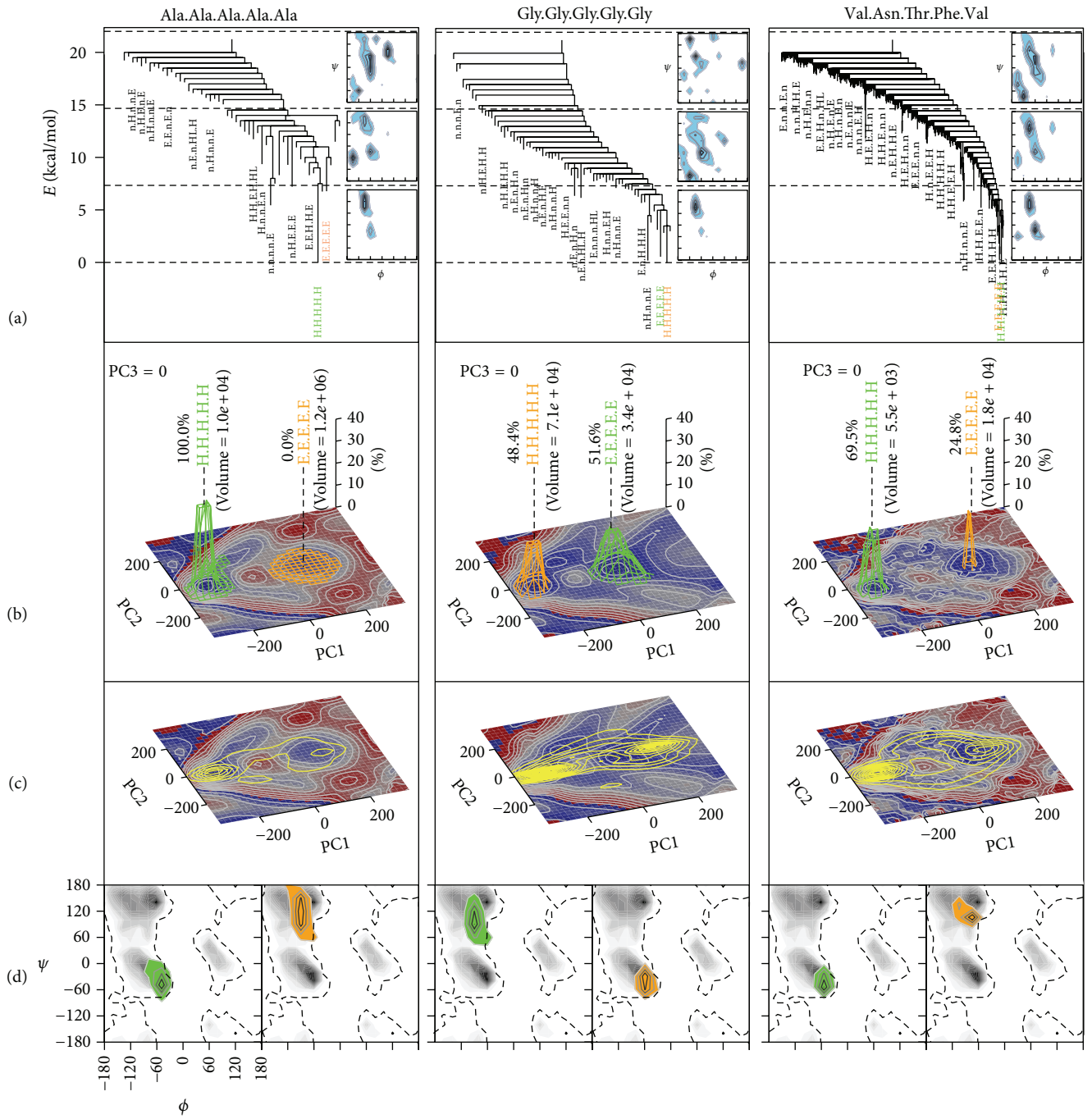

FIGURE 2: The disconnectivity graphs of the potential energy surfaces in the principal conformational space of pentapeptide segments (a), thermodynamic propensities (at 298 K) superposed on the potential energy slice corresponding to the dominant energy valley (b), superposition of respective sequences on these slices (c) (in green), and Ramachandran plots of dominant energy valleys (d) for the penta-alanine (1), pentaglycine (2), and Val.Asn.Thr.Phe.Val (3) sequences. The disconnectivity graphs (a) are offset relative to the respective global minima; minima are labelled with their secondary structure classification based on position on the Ramachandran plot (per residue: H indicates helix and E an extended sheet while n indicates nonavailable secondary structure classification). Minima corresponding to the two valleys with the highest propensities are coloured in green and orange. Inset: the ϕ, ψ distribution within energy intervals of the disconnectivity plots. The thermodynamic propensities (b) are superposed on the potential energy slice closest to the minimum of the highest propensity energy valley (shown in green). The predominant secondary structure classification of each energy valley is indicated. The density of the five (ϕ, ψ) pairs of the structures within the predominant energy valleys is superposed on the Ramachandran plots of the raw dataset (d) in respective colours while the density contours of raw data are shown in black.

selected sequences) in terms of the high density of the left handed α-helix conformational states (Figure 2(a1) top inset Ramachandran plot).

It is interesting to note that at the lowest band of the potential energy funnels (see Figure 2(a); row wise; bottom inset Ramachandran plot) the distributions of the conformational states appear very similar with the density of β-strand being the highest for all sequences. However, it is important to note that these conformational distributions are based on the consideration of the potential energy landscape alone and do not reflect any entropic contributions which we address below.

3.6. Sequence-Specific Thermodynamic Propensities. Quantitative analysis of the PES surfaces via partitioning into different energy valleys and calculation of the local partition functions (see Methods) reveals that the relative propensities of the conformational states at 298 K depend highly on the sequence composition of each pentapeptide segment (Figure 2(b), row wise).

Projection of pentapeptide fragments corresponding to the three selected sequences (X-ray crystal structures with resolution <2.0 Å from the protein databank) [62] onto their respective PESs reveals observed conformational preferences for the $(Ala)_5$ and Val.Asn.Thr.Phe.Val sequences which are largely in accord with their thermodynamic propensities derived from their PESs (compare Figures 2(b) and 2(c) row wise). The computed preference of the $(Ala)_5$ sequence to adopt an α-helical conformation (see Figure 2(b1)) is also in accord with the experimental observation of unusual high helix stability in short alanine-based peptides in water [63]. However, the Val.Asn.Thr.Phe.Val sequence is quite unique since a single secondary structure conformation does not predominate it thermodynamic distribution (at 298 K). The computed propensities of the α-helix and β-sheet for this sequence are comparable (almost 3 : 1 ratio). This indicates that the observation of the Val.Asn.Thr.Phe.Val sequence in either α-helix or β-sheet [30] may not be due to secondary structure interaction; it is rather due to the intrinsic nature of its energy landscape (see below).

The observed distribution of the $(Gly)_5$ sequence (Figure 2(c2)) suggests little preference for any conformation. This is in contrast with the thermodynamic propensities (Figure 2(b2)) which show equal preference for the α and β conformations with negligible contribution from other conformational states. This discrepancy could be due to the high flexibility of the $(Gly)_5$ sequence such that its conformational variance is not as well represented by our 3D PCS and would require either other dimensions to be considered, or alternatively to be considered in a separate analysis. Another consideration is that the conformation of poly-gly sequences is likely to be amongst the least precise in the crystallographic data.

3.7. Thermodynamic Bias towards Different Secondary Structure Classes. The conformational propensities for the three selected pentapeptide sequences can be ascribed to two contributions to the free energy bias: enthalpic and entropic. The enthalpic contribution stems from the relative energetics of the energy minima while the (conformational) entropic contribution is related to the relative volumes of the corresponding energy valleys. These two contributions can be intuitively deduced from inspection of the disconnectivity graphs and the topography of the surfaces respectively (see Figures 2(a) and 2(b)).

The large energetic difference of the minima of the $(Ala)_5$ α-helix and β-strand (see Figure 2(a1)) indicates that the free energy bias towards the helix formation is driven mainly by enthalpic stabilization of the α-helix relative to the β-strand. The lower entropic stabilization of α-helix versus β-strand in $(Ala)_5$ can be readily discerned from the relative volumes of their respective energy valleys; the volume of the α-helix valley is almost two orders of magnitude lower than that of the β-strand (see Figure 2(b1)). By contrast, the Val.Asn.Thr.Phe.Val sequence shows a nonnegligible propensity for the β-strand conformation (see Figure 2(b3)). Given the small energetic difference between the minima of the α-helix and β-strand conformations (0.27 kcal/mol) for this sequence (Figure 2(a3)); it is clear that the preference for the β-strand is provided by entropic factors. This is borne out by the volume of the β-strand energy valley which is almost three times greater than that of the α-helix (see Figure 2(b3)). On the other hand, the $(Gly)_5$ sequence experiences neither enthalpic nor entropic bias towards either of the α-helix and β-strand conformations resulting in them having very similar propensities.

4. Conclusion

The use of a collective coordinate basis set for mapping the potential energy surface of pentapeptide segments in a low dimensional principal conformational subspace allowed for a quantitative assessment of sequence-specific conformational preferences within a statistical thermodynamic framework. The calculated thermodynamic propensities (at 298 K) of $(Ala)_5$ and Val.Asn.Thr.Phe.Val sequences are in accord with their statistically derived secondary structure preferences based on structures in the protein databank. The $(Ala)_5$ sequence showed a predominant propensity for an α-helical secondary structure which is in accord with the reported high stability of the α-helical conformation in short alanine-based peptides [63]. By contrast, the Val.Asn.Thr.Phe.Val sequence showed a predominant preference for the β-strand conformers. For $(Gly)_5$, the observed conformational distribution showed little preference for specific secondary structures—as expected due to its greater intrinsic flexibility. However, our thermodynamic calculations suggest that $(Gly)_5$ has equal propensity for both α-helix and β-strand and further work is required to understand this discrepancy.

Analysis of the topography and topology of the energy landscapes reveals that preference for the α-helix in the $(Ala)_5$ sequence is mainly enthalpic in origin, whereas stabilization of the β-strand in the Val.Asn.Thr.Phe.Val sequence is mainly due to conformational entropic factors. The results indicate that local physical interactions within pentapeptide segments lead to significant sequence-specific conformational preferences resulting from an interplay of enthalpic and entropic

factors. Such decomposition is not possible for statistically derived conformational propensity scales [64].

The method therefore provides a powerful and general framework for investigating the sequence-specific thermodynamic and dynamic properties of polypeptide segments derived from underlying conformational energy landscapes. For example, our approach may also be of utility in the investigation of specific sequence dependent structural properties, such as the identification of protein folding initiation sites [65, 66] which focus on sequences associated with a limited range of conformational states. We are currently extending the methodology to longer segments and incorporation of secondary and tertiary interactions and to improving the representation of the environment.

Competing Interests

The author declares that there are no competing interests.

Acknowledgments

The author is very grateful to Dr. Leo Caves for insightful discussions and helpful remarks. The author gratefully acknowledges Qassim University, represented by the Deanship of Scientific Research, for the material support for this research under no. 3041 during the academic year 1436 AH/2015 AD.

References

[1] K. A. Dill and J. L. MacCallum, "The protein-folding problem, 50 years on," *Science*, vol. 338, no. 6110, pp. 1042–1046, 2012.

[2] S. W. Englander and L. Mayne, "The nature of protein folding pathways," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, pp. 15873–15880, 2014.

[3] R. J. Pantazes, M. J. Grisewood, and C. D. Maranas, "Recent advances in computational protein design," *Current Opinion in Structural Biology*, vol. 21, no. 4, pp. 467–472, 2011.

[4] N.-W. Hsiao, T.-S. Tseng, Y.-C. Lee et al., "Serendipitous discovery of short peptides from natural products as tyrosinase inhibitors," *Journal of Chemical Information and Modeling*, vol. 54, no. 11, pp. 3099–3111, 2014.

[5] A. Eniade, M. Purushotham, R. N. Ben, J. B. Wang, and K. Horwath, "A serendipitous discovery of antifreeze protein-specific activity in C-linked antifreeze glycoprotein analogs," *Cell Biochemistry and Biophysics*, vol. 38, no. 2, pp. 115–124, 2003.

[6] O. Pible and J. Armengaud, "Improving the quality of genome, protein sequence, and taxonomy databases: a prerequisite for microbiome meta-omics 2.0," *Proteomics*, vol. 15, no. 20, pp. 3418–3423, 2015.

[7] A. Nayeem, D. Sitkoff, and S. Krystek Jr., "A comparative study of available software for high-accuracy homology modeling: from sequence alignments to structural models," *Protein Science*, vol. 15, no. 4, pp. 808–824, 2006.

[8] D. Baker and A. Sali, "Protein structure prediction and structural genomics," *Science*, vol. 294, no. 5540, pp. 93–96, 2001.

[9] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, "Stereochemistry of polypeptide chain configurations," *Journal of Molecular Biology*, vol. 7, pp. 95–99, 1963.

[10] P. J. Flory, *Statistical Mechanics of Chain Molecules*, John Wiley & Sons, New York, NY, USA, 1969.

[11] R. V. Pappu, R. Srinivasan, and G. D. Rose, "The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 23, pp. 12565–12570, 2000.

[12] Y. Z. Ohkubo and C. L. Brooks III, "Exploring Flory's isolated-pair hypothesis: statistical mechanics of helix-coil transitions in polyalanine and the C-peptide from RNase A," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 2, pp. 13916–13921, 2003.

[13] S. Hovmoller, T. Zhou, and T. Ohlson, "Conformations of amino acids in proteins," *Acta Crystallographica D: Biological Crystallography*, vol. 58, pp. 768–776, 2002.

[14] R. J. Anderson, Z. Weng, R. K. Campbell, and X. Jiang, "Main-chain conformational tendencies of amino acids," *Proteins: Structure, Function and Genetics*, vol. 60, no. 4, pp. 679–689, 2005.

[15] S. V. Krivov and M. Karplus, "Hidden complexity of free energy surfaces for peptide (protein) folding," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 41, pp. 14766–14770, 2004.

[16] C. L. Brooks III, J. N. Onuchic, and D. J. Wales, "Statistical thermodynamics: taking a walk on a landscape," *Science*, vol. 293, no. 5530, pp. 612–613, 2001.

[17] J. N. Onuchic, H. Nymeyer, A. E. García, J. Chahine, and N. D. Socci, "The energy landscape theory of protein folding: insights into folding mechanisms and scenarios," *Advances in Protein Chemistry*, vol. 53, pp. 87–152, 2000.

[18] M. S. Cheung, L. L. Chavez, and J. N. Onuchic, "The energy landscape for protein folding and possible connections to function," *Polymer*, vol. 45, no. 2, pp. 547–555, 2004.

[19] P. G. Wolynes, "Energy landscapes and solved protein-folding problems," *Philosophical Transactions of the Royal Society A Mathematical Physical and Engineering Sciences*, vol. 363, no. 1827, pp. 453–464, 2005.

[20] S. V. Krivov and M. Karplus, "Free energy disconnectivity graphs: application to peptide models," *The Journal of Chemical Physics*, vol. 117, no. 23, pp. 10894–10903, 2002.

[21] O. M. Becker and M. Karplus, "The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics," *Journal of Chemical Physics*, vol. 106, no. 4, pp. 1495–1517, 1997.

[22] R. Czerminski and R. Elber, "Reaction path study of conformational transitions in flexible systems: applications to peptides," *The Journal of Chemical Physics*, vol. 92, no. 9, pp. 5580–5601, 1990.

[23] A. Magner, W. Szpankowski, A. Magner, and D. Kihara, "On the origin of protein superfamilies and superfolds," *Scientific Reports*, vol. 5, pp. 2045–2322, 2015.

[24] S. Govindarajan, R. Recabarren, and R. A. Goldstein, "Estimating the total number of protein folds," *Proteins: Structure, Function and Genetics*, vol. 35, no. 4, pp. 408–414, 1999.

[25] H. Frauenfelder, F. Parak, and R. D. Young, "Conformational substates in proteins," *Annual Review of Biophysics and Biophysical Chemistry*, vol. 17, pp. 451–479, 1988.

[26] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, "Essential dynamics of proteins," *Proteins: Structure, Function and Genetics*, vol. 17, no. 4, pp. 412–425, 1993.

[27] M. Duan, M. Li, L. Han, and S. Huo, "Euclidean sections of protein conformation space and their implications in dimensionality reduction," *Proteins: Structure, Function and Bioinformatics*, vol. 82, no. 10, pp. 2585–2596, 2014.

[28] G. E. Sims, I.-G. Choi, and S.-H. Kim, "Protein conformational space in higher order ϕ-ψ maps," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 3, pp. 618–621, 2005.

[29] K. M. Elsawy, M. K. Hodgson, and L. S. D. Caves, "The physical determinants of the DNA conformational landscape: an analysis of the potential energy surface of single-strand dinucleotides in the conformational space of duplex DNA," *Nucleic Acids Research*, vol. 33, no. 18, pp. 5749–5762, 2005.

[30] W. Kabsch and C. Sander, "On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 81, no. 4, pp. 1075–1078, 1984.

[31] U. Hobohm and C. Sander, "Enlarged representative set of protein structures," *Protein Science*, vol. 3, no. 3, pp. 522–524, 1994.

[32] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[33] E. G. Hutchinson and J. M. Thornton, "PROMOTIF—a program to identify and analyze structural motifs in proteins," *Protein Science*, vol. 5, no. 2, pp. 212–220, 1996.

[34] J. E. Jackson, *A User's Guide to Principal Components*, John Wiley & Sons, New York, NY, USA, 1991.

[35] D. Bharanidharan and N. Gautham, "Principal component analysis of DNA oligonucleotide structural data," *Biochemical and Biophysical Research Communications*, vol. 340, no. 4, pp. 1229–1237, 2006.

[36] G. E. Sims and S.-H. Kim, "Global mapping of nucleic acid conformational space: dinucleoside monophosphate conformations and transition pathways among conformational classes," *Nucleic Acids Research*, vol. 31, no. 19, pp. 5607–5616, 2003.

[37] S. Mesentean, S. Fischer, and J. C. Smith, "Analyzing large-scale structural change in proteins: comparison of principal component projection and Sammon mapping," *Proteins: Structure, Function and Genetics*, vol. 64, no. 1, pp. 210–218, 2006.

[38] T. H. Reijmers, R. Wehrens, and L. M. C. Buydens, "Circular effects in representations of an RNA nucleotides data set in relation with principal components analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 56, no. 2, pp. 61–71, 2001.

[39] B. J. Grant, A. P. C. Rodrigues, K. M. ElSawy, J. A. McCammon, and L. S. D. Caves, "Bio3d: an R package for the comparative analysis of protein structures," *Bioinformatics*, vol. 22, no. 21, pp. 2695–2696, 2006.

[40] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, http://www.R-project.org.

[41] A. Kitao and N. Go, "Investigating protein dynamics in collective coordinate space," *Current Opinion in Structural Biology*, vol. 9, no. 2, pp. 164–169, 1999.

[42] J. C. Gower, "Some distance properties of latent root and vector methods used in multivariate analysis," *Biometrika*, vol. 53, pp. 325–338, 1966.

[43] J. C. Gower, "Adding a point to vector diagrams in multivariate analysis," *Biometrika*, vol. 55, no. 3, pp. 582–585, 1968.

[44] L. S. D. Caves, J. D. Evanseck, and M. Karplus, "Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin," *Protein Science*, vol. 7, no. 3, pp. 649–666, 1998.

[45] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: a program for macromolecular energy, minimization, and dynamics calculations," *Journal of Computational Chemistry*, vol. 4, no. 2, pp. 187–217, 1983.

[46] N. Foloppe and A. D. MacKerell Jr., "All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data," *Journal of Computational Chemistry*, vol. 21, no. 2, pp. 86–104, 2000.

[47] B. R. Gelin and M. Karplus, "Sidechain torsional potentials and motion of amino acids in porteins: bovine pancreatic trypsin inhibitor," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 72, no. 6, pp. 2002–2006, 1975.

[48] D. Bashford and D. A. Case, "Generalized born models of macromolecular solvation effects," *Annual Review of Physical Chemistry*, vol. 51, pp. 129–152, 2000.

[49] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, "Electrostatics of nanosystems: application to microtubules and the ribosome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 18, pp. 10037–10041, 2001.

[50] A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack Jr., "A graph-theory algorithm for rapid protein side-chain prediction," *Protein Science*, vol. 12, no. 9, pp. 2001–2014, 2003.

[51] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Transactions on Mathematical Software*, vol. 22, no. 4, pp. 469–483, 1996.

[52] J. P. K. Doye and D. J. Wales, "On potential energy surfaces and relaxation to the global minimum," *The Journal of Chemical Physics*, vol. 105, no. 18, pp. 8428–8445, 1996.

[53] G. M. Verkhivker, P. A. Rejto, D. Bouzida et al., "Navigating ligand-protein binding free energy landscapes: universality and diversity of protein folding and molecular recognition mechanisms," *Chemical Physics Letters*, vol. 336, no. 5-6, pp. 495–503, 2001.

[54] A. Baumketner, J.-E. Shea, and Y. Hiwatari, "Improved theoretical description of protein folding kinetics from rotations in the phase space of relevant order parameters," *Journal of Chemical Physics*, vol. 121, no. 2, pp. 1114–1120, 2004.

[55] D. Idiyatullin, I. Nesmelova, V. A. Daragan, and K. H. Mayo, "Heat capacities and a snapshot of the energy landscape in protein GB1 from the pre-denaturation temperature dependence of backbone NH nanosecond fluctuations," *Journal of Molecular Biology*, vol. 325, no. 1, pp. 149–162, 2003.

[56] O. M. Becker, "Quantitative visualization of a macromolecular potential energy "funnel"," *Journal of Molecular Structure*, vol. 398-399, pp. 507–516, 1997.

[57] O. M. Becker, "Principal coordinate maps of molecular potential energy surfaces," *Journal of Computational Chemistry*, vol. 19, no. 11, pp. 1255–1267, 1998.

[58] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[59] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, "Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 26, pp. 9885–9890, 2006.

[60] J. N. Onuchic and P. G. Wolynes, "Theory of protein folding," *Current Opinion in Structural Biology*, vol. 14, no. 1, pp. 70–75, 2004.

[61] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, "Navigating the folding routes," *Science*, vol. 267, no. 5204, pp. 1619–1620, 1995.

[62] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[63] S. Marqusee, V. H. Robbins, R. L. Baldwin, S. Marqusee, and R. L. Baldwin, "Unusually stable helix formation in short alanine-based peptides," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 14, pp. 5286–5290, 1989.

[64] Y. Zhang and J. Skolnick, "The protein structure prediction problem could be solved using the current PDB library," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 4, pp. 1029–1034, 2005.

[65] F. Avbelj and J. Moult, "Determination of the conformation of folding initiation sites in proteins by computer simulation," *Proteins: Structure, Function and Genetics*, vol. 23, no. 2, pp. 129–141, 1995.

[66] K. F. Han, C. Bystroff, and D. Baker, "Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns," *Protein Science*, vol. 6, no. 7, pp. 1587–1590, 1997.