A Method for Correlations Analysis of Coordinates: Applications for Molecular Conformations

Doron Chema* and Oren M. Becker‡

School of Chemistry, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel

Received December 20, 2001

We describe a new method to analyze multiple correlations between subsets of coordinates that represent a sample. The correlation is established only between specific regions of interest at the coordinates. First, the region(s) of interest are selected at each molecular coordinate. Next, a correlation matrix is constructed for the selected regions. The matrix is subject to further analysis, illuminating the multidimensional structural characteristics that exist in the conformational space. The method's abilities are demonstrated in several examples: it is used to analyze the conformational space of complex molecules, it is successfully applied to compare related conformational spaces, and it is used to analyze a diverse set of protein folding trajectories.

I. INTRODUCTION

Conformational analysis is an important computational tool for the investigation of the structure and flexibility of biomolecules, including peptides and proteins. It is used to identify the most stable structures, of small to medium sized molecules,[1,2] to analyze their overall flexibility,[3] and also to identify sets of biologically relevant conformations. Analysis of a protein's folding trajectories[4,5] as well as of protein dynamics near the crystal structure[6,7] contributes to our understanding of protein structure and function. In the context of rational drug design, conformational analysis is often applied to flexible molecules such as peptides, since it is well-known that conformational considerations play an important role in determining the specificity and the potency of peptide drugs.[8−10]

Molecular properties should in principle be computed on the basis of statistical distributions and investigated through ensembles of molecular conformations. Their analysis could be based on (1) characteristics of the whole molecule's structure, such as clustering molecular conformations into families,[1,2,11−14] that are physically separated in the conformational space,[15−17] and (2) analysis of subsets of molecular coordinates, for example, the assignment of rigid areas' growth in protein folding.[18,19] In this case, structural analysis may concentrate on separated rigid regions. Rigid regions, could contribute differently to the measured bioactivity than flexible regions,[20−22] and are worth separating from the rest of the molecule's coordinates. Another approach is the clustering of conformations that pertain only to known rigid parts of the molecule[21] or by applying different geometrical cutoffs to different parts of the molecule.[22] Nevertheless, the use of the latter is restricted to cases where the flexibility of the molecular regions is well established. In cases of a significant correlation between coordinates, Principal Component Analysis (PCA) may be used to simplify the dimensionality of the conformational space.[23−25] In very diverse conformational spaces, alternative methods may be required.[26,27]

Here, we introduce a method to establish the correlation between coordinates, by first selecting only "interesting" coordinates, followed by a selection of only "interesting" regions of each of these coordinates. This approach also neglects "noninteresting" coordinates of the set. It is especially suitable to analyze a conformational set, in which the definition of coordinate rigidity as "interesting" has a structural meaning, as we describe. The following steps are employed: (1) region(s) of interest are collected from several of the molecular coordinates, while completely omitting other coordinates from the set, following a predefined classification scheme. (2) The correlations between the chosen coordinates' regions are constructed. The advantages of the process are, first, the reduction of noise generated by "noninteresting" data, on each coordinate; second, the reduction of coordinates that contain almost no information; and third, contributing an easily interpretable structural context. One may thus gain structural information directly from the analysis of a matrix. In addition, as shown in some examples, complementary structural information may be gained from an analysis of the one-dimensional regions of interest, in addition to the cross-correlation analysis.

In the following, we present (i) a description of the analysis method and (ii) examples of three conformational spaces for a cyclic peptide [GSAGPV], for Substance P (SP), and for a Substance P analogue (SP-7). Finally, we apply this analysis to multiple (100) "early" folding trajectories of Crambin, a 46 amino acid protein.

II. METHODS

The starting point for the algorithm is a set of conformations, that represents the conformational space of a molecule, and a set of coordinates which best define these conformations. The algorithm consists of three steps:

(1) A set of representative clusters for each coordinate is constructed, separately for every coordinate. (2) One-

* Corresponding author phone: 972-52-222977; e-mail: doronc@emerald.tau.ac.il.

‡ Current address: BIO-IT (Bio-Information Technologies) LTD, 1 Betzalel St. Ramat-Gan 52521, Israel.

10.1021/ci0103471 CCC: $22.00
Published on Web 05/17/2002

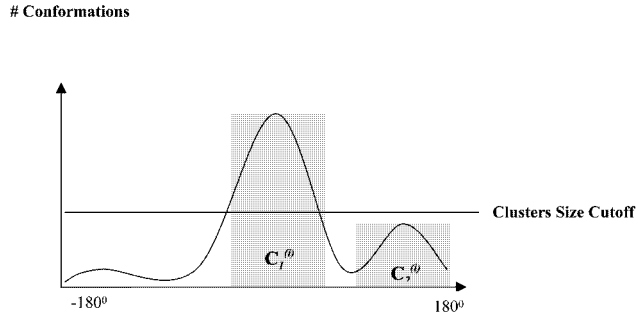


Figure 1. A scheme of the geometrical distribution of conformations along coordinate i. The largest cluster C_1^i is collected at the geometrical cutoff colored in gray. The next cluster collected C_2^i is below the size cutoff used.

dimensional clusters are selected, following certain rules. (3) The number of conformations that is common to each cluster pair is compiled into a correlation matrix. This matrix presents the correlation between specific rigid regions of the molecule. It may be demonstrated that the information gathered by these steps can be used to gain insight into the molecular conformational space.

1. Clustering the Coordinates of Interest. As a preliminary step, the coordinates which best describe the conformational properties of the molecule must be chosen. In peptides and proteins, these may primarily be the backbone dihedral angles, which are the most representative. The method uses the conformational matrix A ($m \times n$), where m and n are the number of conformations and coordinates, respectively.

After the relevant sets of coordinates have been chosen, each coordinate is assigned a set of one-dimensional clusters, $C_k^{(i)}$ (coordinate i and cluster serial number k). The algorithm was first introduced by Bravi et al.,[22] as a density clustering method applied to cluster peptide conformations. Here it has a different purpose, namely, as an efficient method to compile the distribution of conformations along each coordinate, into a discrete number of states. Briefly, the clustering algorithm works iteratively by eliminating, at each step, the densest cluster (i.e. one that contains the maximum number of conformations) from the total conformational space. A geometrical cutoff is used to define a similarity range between conformations, at the same coordinate. The process is completed when all conformations have been incorporated into clusters. Nevertheless, to enhance the calculation, one can stop the process once the size of the last cluster reaches a predefined value, as demonstrated in Figure 1. This cutoff value should be in principle less than the cutoff value used in the next step.

2. Choosing Informative Clusters. A given backbone torsion angle can be classified according to one of the following types: (1) restricted torsion angles, where a significant population from the conformations adopts similar values along the coordinate. One or two major clusters followed by smaller clusters would reflect such a coordinate. (2) A freely rotating torsion angle should result in several small sized clusters. The torsion angles around sp^3-sp^3 bond typically give rise to three clusters, each comprising about a third of the total conformations. To distinguish such coordinate types, we assume that the first type coordinates are reflected by one or two clusters of more than 40% of the conformations.

Applying these definitions, it is possible to filter out clusters that belong to very flexible torsion angles or small clusters associated with restricted torsion angles. The relationship between the restricted torsion regions, represented by the remaining clusters, is the essence of the final step.

3. Constructing a Correlation Matrix. Correlation refers to different conformations that share similar values of interesting coordinates. Therefore, the number of conformations that would be found as joint ones for two one-dimensional clusters defines how much these two clusters are correlated. Based on this definition the correlation matrix, S, is constructed as follows. The matrix cell $S_{(i,j)}$ contains the number of conformations that are common to the one-dimensional clusters that are most populated, $C_1^{(i)}$ and $C_1^{(j)}$, normalized by the total number of the sampled conformations. Normalization has the advantage of leaving the matrix in a symmetric form, while its terms become the absolute fraction over the total sampled conformations.

$$s_{ij} \equiv \frac{C_m^{(i)} \cap C_n^{(j)}}{\Sigma \text{ conf}}$$

An alternative normalization is with the size of a single cluster. Such normalization is applied to the matrix named S^b. The matrix cell, $C_{(i,j)}$, presents the correlation between clusters i (row) and j (column) normalized by the size of the cluster, i. This matrix form can be employed to examine the correlation of one cluster with others, in terms of its own size. In the next sections, both kinds of matrix forms, **S** and S^b, would be used.

Only one cluster for each torsion angle, the most populated one, is introduced into the matrix. In case of flexible coordinates, the relevant matrix cells (row and column) remain empty. In case of two restricted geometries on a specific coordinate, the clusters' correlations with the other clusters are entered into different matrices or separated within the same matrix. Nevertheless, this complication can be overcome if there is a significant correlation between several of these alternative clusters, taken from different coordinates. Such a situation is described in our first example.

Examining the general form of the matrix, highly correlated rigid coordinates are translated into high matrix values. A high correlation at a continuous rigid region would be reflected in a "block diagonal" form (Figure 2(a)). Correlation between continuous regions, on the other hand, would be reflected by the off diagonal values (Figure 2(b,c)). Finally, the high correlation between separate regions, described in Figure 2(c), is transformed into a S^b type matrix (Figure 2(d)).

Computational Demands. The most time-consuming process in the analysis is the one-dimensional clustering step. The time needed for this analysis is a function of the number of conformations m and the number of torsion angles n, giving a time dependence of $n \times (m^2)$. The computational time also depends on the nature of the data, since flexible coordinates require more calculation cycles. It is recommended to use no more than 10^3 conformations for this analysis. Nevertheless, according to our experience, this process should take no more than several minutes for a set of 10−100 of coordinates and 500−1000 conformations on a 433 MHz Pentium CPU.

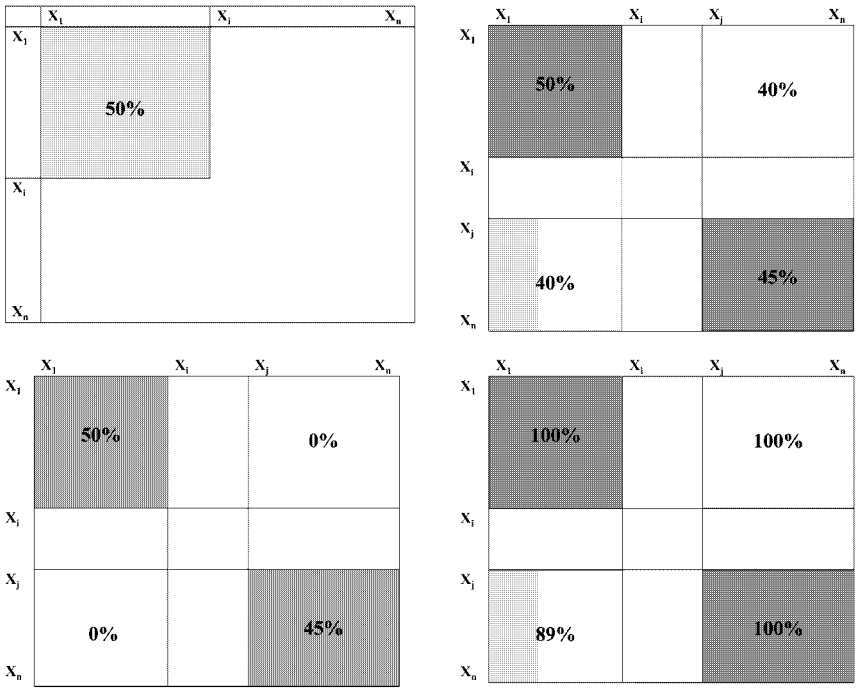


Figure 2. Illustrative correlation matrices of a molecule that has (a) one, completely correlated, rigid region that occupies 50% of the conformational space. (b) Two completely correlated rigid regions that have no correlation with each other. (c) The intercorrelation between the regions taken from (b) has changed to 40%. (d) The same situation as in (c), implemented into a correlation matrix of the type S^b.

The above methodology was applied to four molecules for which conformational studies were already at hand:

(a) ***Cyclic [GSAGPV]***. A sample of 1000 conformations was obtained from a 500 ps high-temperature (1000 K) molecular dynamics trajectory. Each high-temperature conformation was then gradually cooled to 300 K. At each step, the temperature was reduced by 100 K, and the conformations were simulated for 0.8 ps. Reaching 300 K, each conformation was minimized to the nearest local minimum, using 300 steps of Steepest Descent followed by 1000 steps of Adopted Basis Newton−Raphson (ABNR). The simulations were performed using the molecular dynamics program CHARMM[28] and the CHARMM all-atom force field[29] using 2 fs time steps. A 15 Å cutoff was applied to nonbonded interactions (VDW and electrostatics) and the SHAKE constraints limited bonds to hydrogen atoms.

(b) Substance P (SP) and SP-7 Analogue. The conformational data used for the analysis of these peptides was taken from a previous work of Becker et al.[30] The same protocol as in (a) was used to generate a sample of 500 conformations for each of these peptides.

(c) Early Folding Ensembles of Crambin. One hundred conformations of crambin were generated by heating a fully extended Crambin conformation to a high temperature (1000 K) while collecting conformations at 1 ps intervals. Each of the preheated conformers was gradually cooled (as described in (a)) to 300 K and was subject to further MD simulations, up to 3 ns, at this temperature. Conformations from each of these individual trajectories were collected and minimized using 300 steps of Steepest Descent followed by 1000 Steps of ABNR, to the closest local minima at 0.5, 1.0, 1.3, 1.6, 2.0, and 3.0 ns for further analysis. These time-dependent conformations' samples (100 conformations at each sample) were subject to further analysis. All simulations were performed using the molecular dynamic program CHARMM[28] and the CHARMM modified polar hydrogen force field[31] using a 1 fs time-step, and a distance-dependent dielectric constant. A cutoff of 15 Å was applied to nonbonded interaction (VDW and electrostatics).

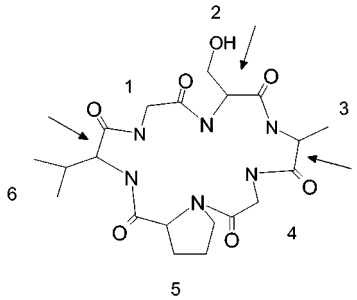


Figure 3. A diagram of the *Cyclic* [GSAGPV] peptide. The arrows point toward flexible regions, as defined in Figure 4(b). The amino acids numbers are given.

III. RESULTS

1. ***Cyclic [GSAGPV].*** Analyzing the backbone dihedral angles φ (phi) and ι (psi) (Figure 3) of the *Cyclic [GSAGPV]* peptide, one-dimensional clusters were collected to a minimal size of 20% or more. Each cluster is named by its coordinate, followed by a serial number. Coordinates may have more than a single cluster, as described in Methods. In most cases, no significant changes were found in cluster sizes and geometry while using different geometrical cutoffs of ±20°, ±25°, and ±30°. Dihedral angles $\varphi 3$ and $\varphi 5$-$\psi 5$-$\varphi 6$ are the most rigid, with a population of over 0.7 (70%), while other cluster populations are below 55% (Figure 4(a)). Clusters $\psi 3-3$ and $\varphi 6-2$ vanish as the cutoff distance is increased to ±30°. The reason for that is that increasing the cutoff may add smaller clusters to larger ones.

Following our classification, dihedral angles $\psi 2$, $\psi 3$, and $\psi 6$ were eliminated due to their flexibility (Figure 4(b)).

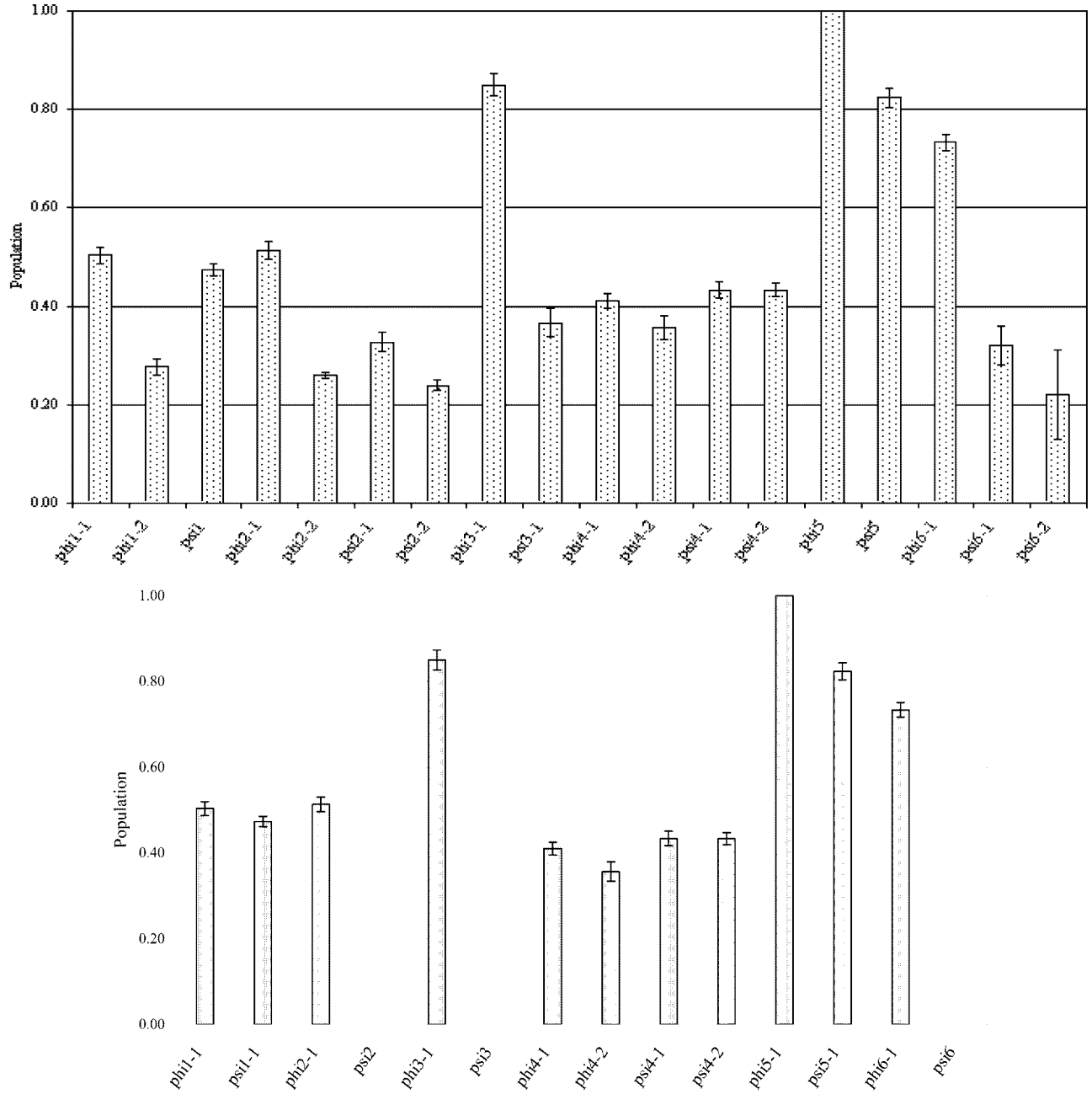
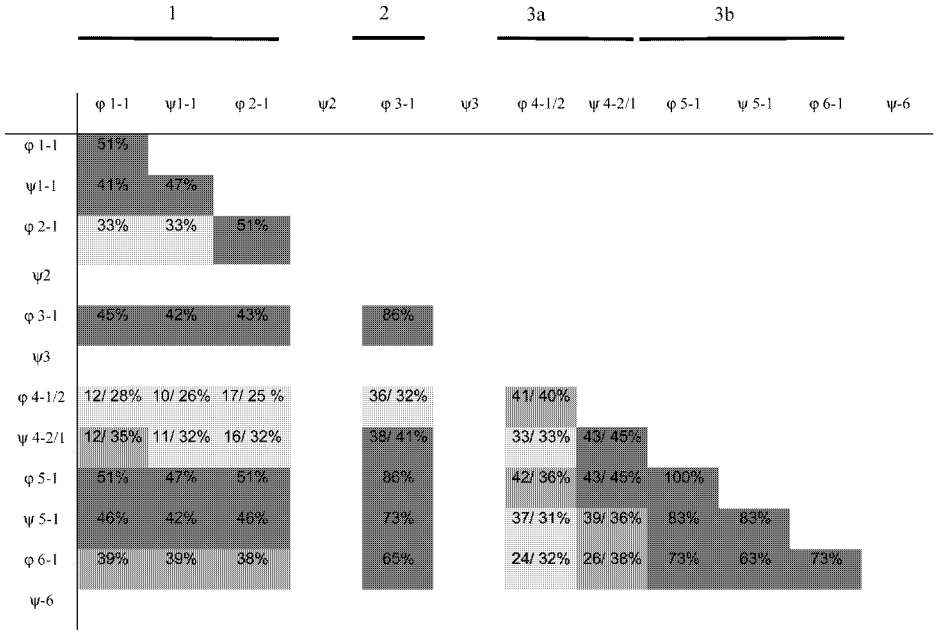


Figure 4. (a) The size of the clusters collected for each of the *Cyclic* [GSAGPV] dihedral angles after the first step of the analysis algorithm. The size of the clusters presented is an average over three cutoffs of ±20°, ±25°, and ±30°. (b) The clusters left after filtering flexible coordinates and minor clusters. The filtration step was applied at a cutoff of ±25°.

Dihedral angles $\varphi 4$ and $\psi 4$ contain two major clusters with sizes of 40 ± 5% that differ by no more than 5%. Following this "filtering step", the peptide structure may be divided into three continuous regions, separated by flexible coordinates. Region-1 that includes the dihedral angles $\varphi 1$-$\varphi 2$, region-2 that includes only $\varphi 3$, and region-3 that includes the dihedral angles $\varphi 4$-$\varphi 6$ (Figure 4(b)). Region-3 includes two subregions: (a) coordinates $\varphi 4$-$\psi 4$, with similar small rigid clusters, and (b) coordinates $\varphi 5$-$\varphi 6$.

Given the number of clusters per coordinate at $\varphi 4$-$\psi 4$, two for each, 2^2 correlation matrices are needed for a full correlation description. Collecting clusters that are highly correlated into a single set can overcome this complication. Of the two pairing possibilities between $\varphi 4$ and $\psi 4$, the clusters combinations $\varphi 4-1/\psi 4-2$ and $\varphi 4-2/\psi 4-1$ were better correlated by 73−83%, respectively. Following that, the contribution of the alternative combinations could be neglected. The correlation matrix containing the correlation between clusters is given in Figure 5(a). Empty rows and columns belong to the missing flexible dihedral angles. The correlation values of $\varphi 4-1$ and $\psi 4-2$ appear at the left of each relevant matrix cell, while those of $\varphi 4-2$ and $\psi 4-1$ appear at the right.

Analysis of the correlation within each region reveals the following picture. Region-1 is characterized by an average correlation value of 36 ± 5% and average cluster sizes of 50 ± 2%. The difference between the numbers is mainly caused by $\varphi 2-1$, which is significantly less correlated. The average correlation in region-3 is of 71 ± 11% and the average cluster size is 85 ± 10%, while taking for simplicity, the contributions of the clusters in regions 3(a) together. A summary matrix presents the average values (Figure 5(b)).



Region	1	2	3
1	36±8%		
2	43±5%	86%	
3	44±5%	74±8%	71±11%

Region	1	2	3
1	74±10%	50±1%	50±4%
2	57±10%	100%	89±5%
3	58±10%	86±10%	90±5%

Figure 5. The correlation matrices constructed for the clusters in Figure 4. (a) The basic correlation matrix, in which shaded colors refer to the following values: above 40% (dark gray), 40−35%, 35−30%, and below 20% (light gray). The major regions are indexed from 1 to 3. The clusters size (described in Figure 4) at the matrix diagonal were entered into the matrix only for clarification reasons. (b) A simplified form of the correlation matrix in (a) included only average correlations within and between regions. (c) A correlation matrix of the type S^b included the average correlations with and between regions.

In this case, the correlations between regions (off-diagonal values) 2−1, 3−1, and 3−2 are slightly greater than the correlations within regions (diagonal values) 1−1 and 3−3, respectively, suggesting a better inter-regional correlation than the intra-regional correlation.

The S^b type correlation matrix defines the average correlation within the regions (diagonal values) in terms of the region clusters (Figure 5(c)). The information gained from the two last matrices (Figure 5b,c), can be projected into a schematic Venn Diagram,[36] presented in Figure 6. In this figure, the different regions are represented as circles, each represent the average internal correlation within a region. The overlap between the circles reflects the average correlation between the regions.

The relative contribution of each of the cluster pairs (region 3(a)) to the inter-correlation with region-1 was extracted. Clusters $\varphi 4-2$ and $\psi 4-1$ are correlated by 29 ± 4% with region-1, while clusters $\varphi 4-1$ and $\psi 4-2$ are correlated by 13 ± 2% only. While being almost of equal size, their relative contributions differ significantly.

2. Substance P. Substance P (SP) is an 11 amino acid neuropeptide with the sequence H−Arg[1]-Pro[2]-Lys[3]-Pro[4]-

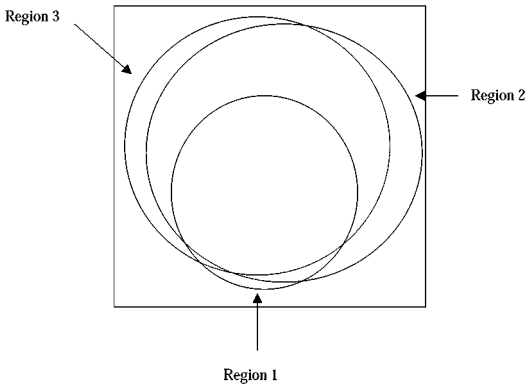


Figure 6. A Venn diagram for the size and the conformational space overlap between the three regions in the *Cyclic* [GSAGPV] peptide. The box symbolizes the complete conformational space.

Gln^5-Gln^6-Phe^7-Phe^8-Gly^9-Leu^{10}-Met^{11}-NH_2. It belongs to the tachykinin family and functions as a neurotransmitter with a variety of biological activities. Extensive studies[32−35] have shown that the C-terminal portion of the molecule, starting at Gln^6, is responsible for the binding of Substance P to the NK1 receptor. SP-7 is a single-mutation analogue of Substance P (SP), in which L-Gln^6 was replaced with D-Gln^6. Its binding affinity to the NK1 receptor was determined to be 2.4 kcal/mol lower than that of the native SP.[32] We use this analogue to study the effects of a point mutation on the conformational properties of the SP peptide.

The backbone torsion angles (taken from the conformations in ref 30) of SP and SP-7 were clustered by the method described in the previous section. For each peptide, clustering results were stable at the different cutoffs. Similar geometries were found for both SP and SP-7 peptide restricted clusters, while differences were observed only for $\varphi 6$ and $\psi 6$, which flipped into different geometries in SP-7. Not surprisingly, these two dihedral angles are those of the mutated amino acid (Gln^6). The size of the examined clusters was found to be similar in most cases, while differences of 8−12% between cluster sizes of the two peptides were found for $\tau 2$, $\psi 4$, $\varphi 6$, $\varphi 9$, and $\iota 10$ (Figure 7(a)).

Following the classification scheme, the distribution of the SP rigid torsion angles is different at the N-terminal compared to the C-terminal. A sequence of rigid torsion angles defines a substantial part of the N-terminal ($\varphi 2-\varphi 7$). The C-terminal region, on the other hand, is characterized by a similar number of flexible and rigid torsion angles, arranged in an alternate pattern. For convenience, the distinct

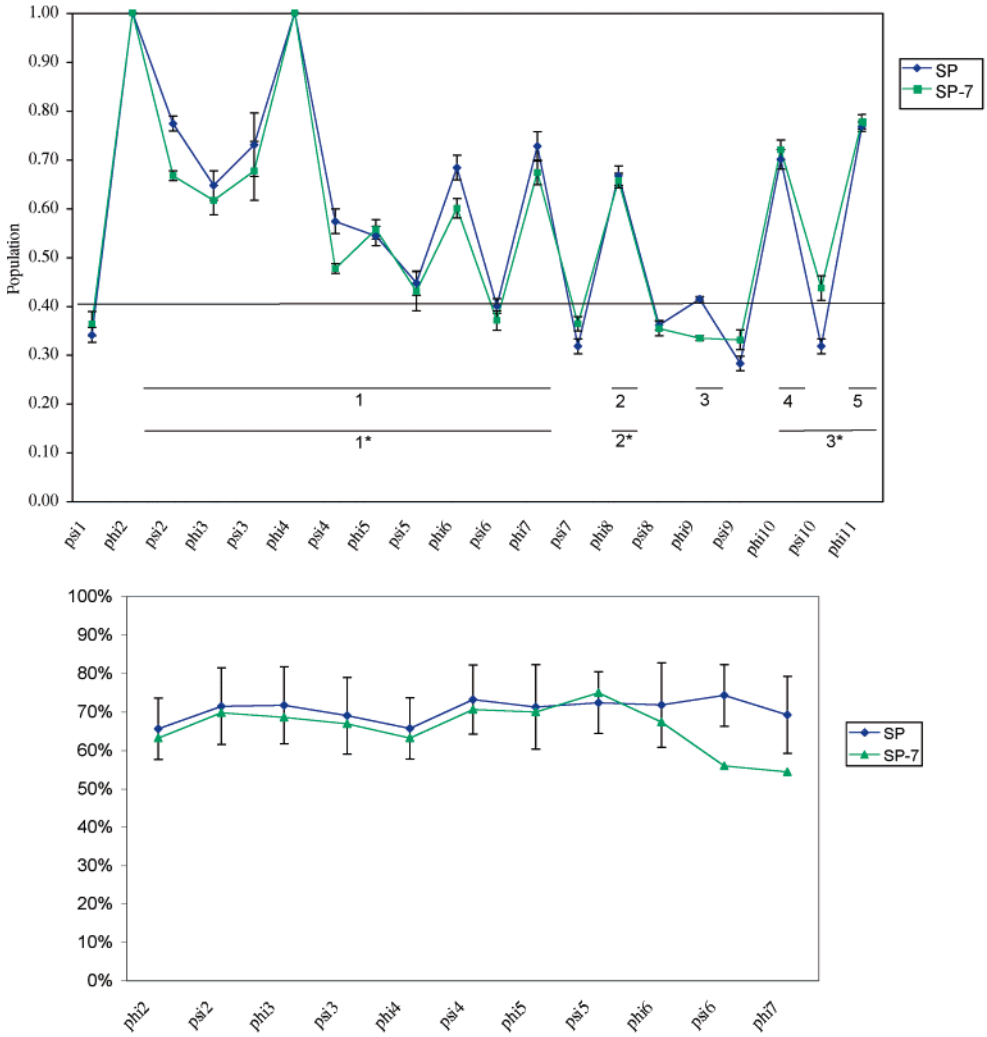


Figure 7. (a) The average size distribution of one-dimensional clusters of SP (blue), and of SP-7 (green), over three cutoffs of ±20°, ±25°, and ±30°. The 0.40 line separates rigid from flexible coordinates. The differences between the two distribution plots are discussed in the text. (b) The average correlation of one-dimensional clusters in region-1 with each of the others, as extracted from the correlation matrix S^b. A large change is encountered at the mutated amino acid Gln^6.

rigid regions at the structure of SP and SP-7 were marked from 1 to 5 and 1*−3* (Figure 7(a)), respectively. It should be noticed that for ψ7, in SP, the average cluster size is 40% while in SP-7 it is reduced into 37%. Since this size difference is rather small while comparing the two peptides (as can be seen also in the graph), ψ7 at SP-7 was considered also as rigid.

A recent theoretical study[30] of a series of SP analogues has shown that the N-terminal region of the SP peptide (Arg^1-Gln^5) is significantly more rigid than the C-terminal (Gln^6-Met^{11}). This observation was established by measuring the backbone root-mean-square distance (rmsd) of each part separately, after previously determining the boundary between them. As shown here, our classification is an alternative method that does not require any preliminary "regional classification" but only choosing an appropriate set of coordinates. Moreover, the automated way of summarizing the one-dimensional distributions as well as the coordinate classifications may be used for a robust analysis and comparative study of such distributions. This could be highly advantageous for analyzing larger systems (such as crambin vide infra) or for studying libraries of molecules.

To compare the differences between SP and SP-7 in internal correlations at region-1, the S^b type correlation matrix was constructed. This matrix is mostly suitable for this purpose, as its analysis can emphasize differences between the conformational spaces of the peptides, in terms of the coordinates. The average correlation of each cluster with the others, at region-1, was calculated. In the case of SP, a narrow range of average correlation, $70 \pm 5\%$, is established (Figure 7(b)). This is in contrast to the highly diverse cluster sizes (Figure 7(a)). A very similar situation is found for SP-7, except for ψ6 and φ7, where a decrease in the average correlation of clusters is observed.

The C-terminal region was previously established as the part that is mainly responsible for activity. In this region, φ-9 of SP-7 is more flexible than in SP, while ψ10 is more rigid in SP-7. This causes an increase in the correlation between ι10 and ι11 of SP-7 by 5% in region-3. The minimum correlation at this c-terminal region was compared for SP and SP-7, and a difference of 6% (30% in SP-7 vs 24% in SP) was found.

3. Crambin. Crambin is a 46 amino acid protein, which was extensively studied theoretically, near the X-ray crystal structure and during folding and unfolding.[7,15,27,31] Here, we apply our method to explore its conformational space at the very beginning of the folding process, by analyzing multiple conformations. One hundred conformations were simulated for a period of 3 ns and were subject to further analysis.

The average root-mean-square distance (rmsd) of the conformations from the structure of the folded protein, as a function of time, is very similar to the average rmsd between conformations (Figure 8). A slight structural convergence is observed for both, while the curvature of the lines is close to a plateau, with high rmsd values (9−10 Å). In addition, no native contacts such as disulfide bonds, or contacts between known secondary elements, were found. The simulation is analyzed here at two times of 1 and 3 ns. Both are post equilibrium ("production" phase), and the time difference between them is long enough to detect potential changes. In Figure 8, no such differences are detectable between the two points in time.

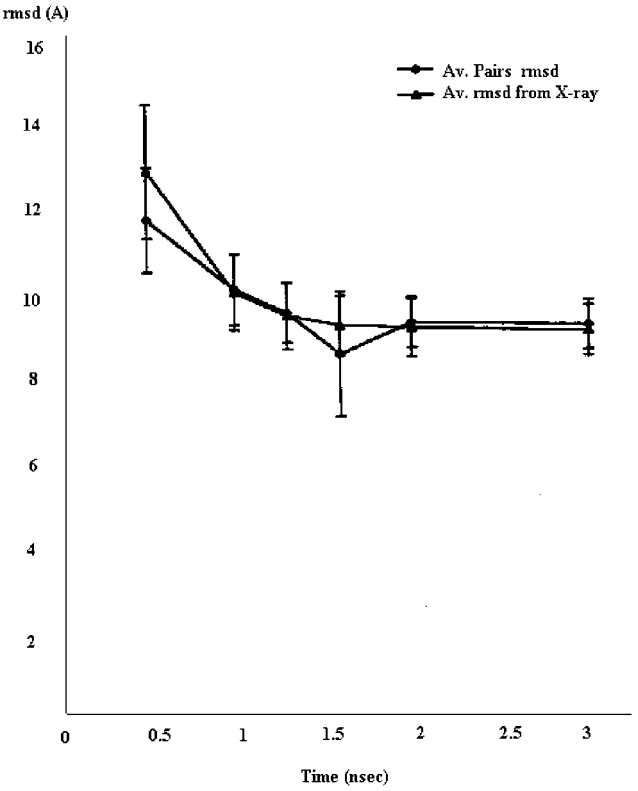


Figure 8. The average rms distance between (a) pairs of conformers and (b) conformers and the crystal structure of Crambin.

Each of the 90 backbone dihedral angles of Crambin was analyzed using geometrical cutoffs of 35° and 45°. From an analysis of the type presented for the peptides and applied to Crambin, nine continuous (at least two consecutive nonflexible dihedral angles) regions could be identified (Figure 9). The differences between values at 1 and 3 ns are displayed in Figure 10 (red line). The respective average correlation values of each cluster with the other regions, extracted from the matrix type S^b, are depicted also in Figure 10 (blue line). This figure suggests an increase in the inter-region correlation between 1 and 3 ns, while the cluster sizes fluctuate along the main chain of Crambin.

A possible connection between the cluster size and its geometrical difference from the crystal structure was examined as well. The clusters (collected at 3 ns) were grouped into four categories, containing 20−40%, 40−60%, 60−80%, and 80−100% of the conformations. At each category, differences between the average geometries of the clusters and the simulation crystal geometry were calculated

$$\Delta\mathrm{dih} = |\mathrm{dih}_{\mathrm{crystal}} - \mathrm{dih}_{\mathrm{cluster}}|$$

where $\mathrm{dih}_{\mathrm{crystal}}$ stands for the X-ray geometry found for a dihedral angle and $\mathrm{dih}_{\mathrm{cluster}}$ stands for the average geometry of the major cluster of the same dihedral angle.

The distributions around the geometrical difference of the different cluster size categories are plotted in Figure 11. Accordingly, rigid coordinates tend to cluster near the native value, while this tendency is reduced when moving to coordinates that are more flexible. Clusters with sizes above 60%, which are found to be concentrated mainly near the crystal geometry, appeared in 49 out of 90 dihedral angles. Eleven of these clusters differ by more than 50° from the

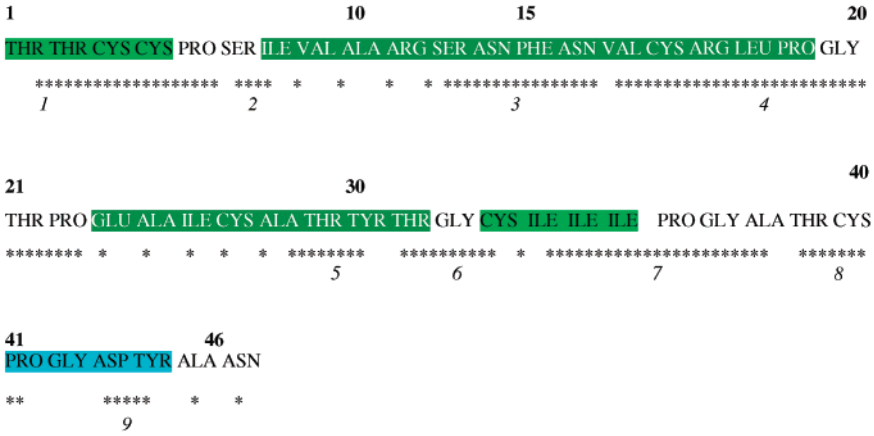


Figure 9. (a) Secondary structures along the Crambin Sequence: strand (light green), helix (dark green), turn (blue), Undefined shape (noncolored). (b) Rigid dihedral angles are marked in *. Two and more continues rigid dihedral angles are numbered (below).

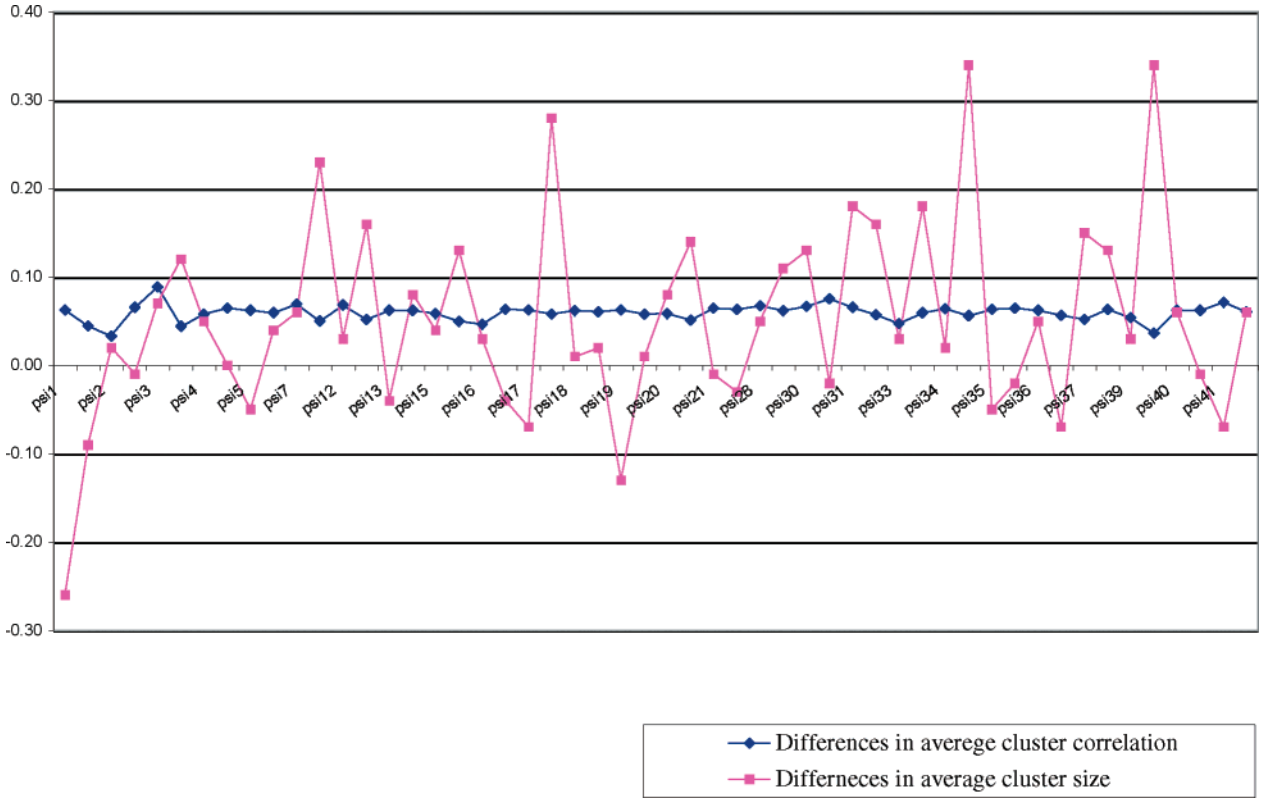


Figure 10. The size difference between subset (see text for details) of the major one-dimensional clusters, that appear in Figure 9 (red). The average correlations of each cluster with the other, as extracted from the matrix of type S^b appears in blue.

crystal geometry (four of them are due to the presence of Gly or Pro), while six of them are found at regions that connect α-helices or *β*-sheets.

VI. DISCUSSION

A new method for correlation analysis was introduced and applied to the analysis of molecular conformations. This method uses ideas taken from three current analysis methods. *Cluster analysis* is applied to summarize each *single coordinate distribution*. A classification scheme is then applied to filter only the desired region(s) at each distribution. *Correlation matrices*, in the form of a covariance matrix, are then constructed. Each matrix cell $S_{(i,j)}$ contains the normalized number of conformations common to a pair of filtered clusters. Normalization can be made either by the size of the sample or by the size of cluster i. The new type of correlation matrix has a general advantage in comparison to the covariance matrix, for structural analysis, as follows. Nonrelevant coordinates are eliminated from the data set. The correlations are only between the "regions of interest" that remain at each coordinate due to a filtration step, which reduces much "noise". Now, since correlations are between structural meaningful regions, structural conclusions may be drawn, as shown in our examples.

The method was first applied to the conformational space of a cyclic peptide [GSAGPV]. Three separated regions were identified in the peptide structure, and one of them contains two alternative geometries. The correlation matrices were constructed and resulted in a Venn diagram, which illustrates the relations between the three regions. We demonstrated the ability to differentiate between the contributions of two alternative geometries found at the third region to the other

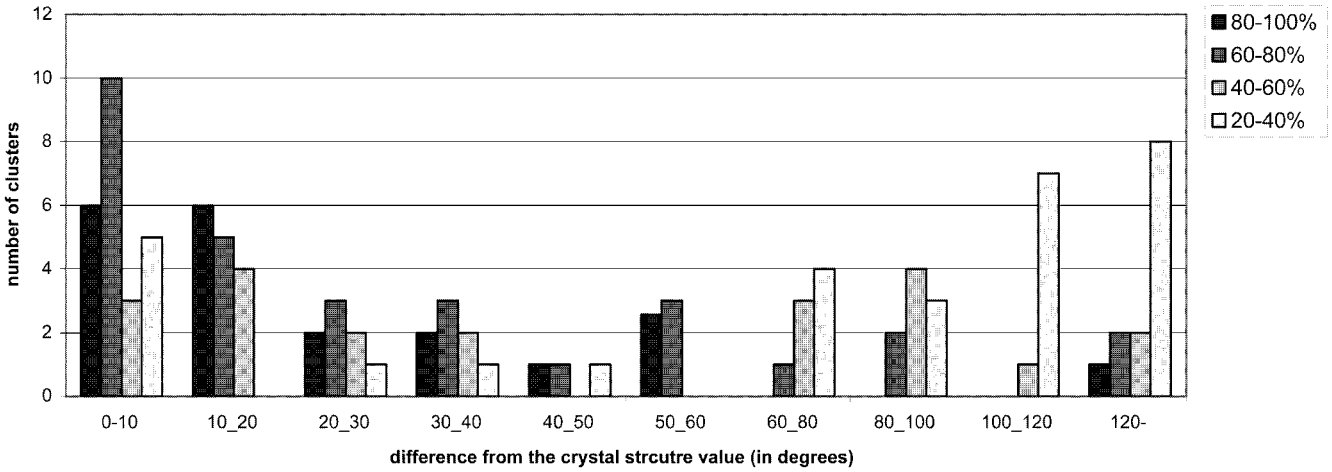


Figure 11. The major clusters (one from each coordinate) are grouped into four categories, contain 80−100%, 60−80%, 60−40%, and 20−40% of the conformations. The distribution of the categories, as a function of their average distance from the crystal value, is plotted.

regions. Although the alternative geometries are almost equal in their cluster size, one of them is significantly more correlated than the other with the other regions. This type of information is unique to the results gained from our method. It has, as is shown in this example, the potential of illuminating nontrivial structural relations between different regions.

In the second example, the conformational spaces of two peptides that differ by a single mutation were compared. The one-dimensional clusters of interest were found to have similar geometries, except at the region of mutation. Several differences in their size were also identified, while most clusters have identical sizes. The S^b matrix was used to gain further insight into the average correlation of each coordinate with the others. Using this method, differences at the peptide N-terminal that were not observed by the one-dimensional comparison were detected. The C-terminal region of the peptides was compared by their minimal correlation, finding that the region in SP-7 has a higher correlation. Our analysis is able to detect similarities and dissimilarities between two conformational spaces, based on structural regions. Following that, the impact of the mutation introduced into the peptide sequence can be evaluated in such terms. Such a conformational analysis may be applied for instance within the frame of peptide structure−activity relationships.

Several methods were previously applied to analyze the conformational space of folding and unfolding intermediate ensembles of protein conformations.[25−27] Such tools are a direct analysis of the rms matrix, contacts analysis, and Principal Component Analysis (PCA).[25,27] These methods, that were found to be very useful for characterizing the molten globule state, were found to be less useful for understanding the early folding (late unfolding) stages, where the conformational space may not be easily characterized.[23−25] In these cases, it was argued that alternative methods might be required.[26−27] The method described by us could be a step in that direction. The method was applied here to a sample of Crambin conformations (100) that was simulated for 3 ns, while conformations were taken after 1 and 3 ns and were analyzed for structural differences. The results indicate that all regions behave similarly with respect to the change of correlation along time. This happens despite the local fluctuations in size of the individual clusters, observed along time. Therefore, despite the fact that no significant energetic or overall similarity (RMSD) changes occur in the ensemble throughout this simulation, an increase in the overall internal arrangement may be detected. This ability to observe local changes suggests that this method could be helpful for analyzing early protein folding ensembles. This is also in contrast to the method described above, which usually measures the overall similarity between conformations, and therefore may miss local changes. In addition, the larger the cluster, the greater is its probability to resemble the dihedral angle conformation of the X-ray crystal structure.

In the case of peptides, especially conformationally constrained ones, structural analysis as described here could have several advantages. The method is not limited to explore only the valleys of the peptide energy landscapes, and following that it may capture structural differences which may be important in cases where a peptide is subject to a structural modification as it binds to a bimolecular target. In addition, as we have shown in the example of the second peptide, the effect on the conformational space has a "structural address", namely the structural region that the modification affects. This "structural address" is unique to our approach.

ACKNOWLEDGMENT

We would like to thank Yaakov Levy and Orr Ravitz for making their data on the Substance P analogues available to us. We thank Profs. Amiram Goldblum and Mordechai Bixon for helpful discussions and suggestions and for the critical assessment of this paper. We also would like to thank Ernesto Joselevich and Andrea Zaliani for comments on this manuscript.

REFERENCES AND NOTES

(1) Leach, A. R. In *Reviews in Computational Chemistry;* Lipkowitz, K. B., Boyd, D. B., Ed.; New York, 1991; Vol. 2, pp 1−55.
(2) Howard, A. E.; Kolmann, P. A. An analysis of current methodologies for conformational searching of complex molecules. *J. Med. Chem.* **1988**, *31*, 1669−75.
(3) Hampel, J. C.; Fine, R. M.; Hassan, M.; Ghoul, W.; Guaragna, A.; Koerber, S. C.; Li, Z.; Hagler, A. T. Conformational analysis of endothelin-1: Effects of solvation free energy. *Biopolymers* **1995**, *36*, 282−301.
(4) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, pathways, and energy landscape of protein folding: A synthesis. *Proteins* **1995**, *21*, 167−195.

(5) Karplus, M.; Shakhnovich, E. In *Protein Folding*; Creighton, T. E., Ed.; W. H. Freeman: New York, 1992; pp 127−195.
(6) Elber, R.; Karplus, M. Multiple conformational states of proteins: A molecular dynamic analysis of myoglobin. *Science* **1987**, *235*, 318−321.
(7) Caves, L. S. D.; Evanseck, J. D.; Karplus, M. Locally accessible conformations of proteins: Multiple dynamics simulations of Crambin. *Protein Sci.* **1998**, *7*, 649−666.
(8) Veber, D. F.; Holy, F. W.; Paleveda, W. J.; Nutt, R. F.; Bergestrand, S. J.; Torchiana, M.; Glitzer, M. S.; Saperstein, R.; Hirshmann, R. conformational restricted bicycle analogues of somatostatin. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *262*, 2636−40.
(9) Pierschbacher, M. D.; Rouslahti, E. Influence of stereochemistry of the sequence Arg-Gly-Asp-Xaa on binding specificity in cell adhesion. *J. Biol. Chem.* **1987**, *262*, 17294−98.
(10) Shenderovich, M. D.; Nikiforovich, G. V.; Golbraikh, A. A. Conformational features responsible for the binding of cyclic analogues of enkephalin to opioid receptors. *Int. J. Peptide Protein Res.* **1991**, *37*, 241−51.
(11) Gordon, H. L.; Somoraji, R. L. Fuzzy cluster analysis of molecular trajectories. *Proteins* **1992**, *14*, 249−264.
(12) Shenkin, P. S.; Mcdonald, D. Q. Cluster Analysis of molecules conformations. *J. Comput. Chem.* **1994**, *15*, 899−916.
(13) Torda, A. E.; Van Gunstern, W. F. J. Algorithm for clustering molecular dynamic configurations. *J. Comput. Chem.* **1994**, *15*, 1331−1340.
(14) Leach, A. R. In *Molecular Modeling Principles and Applications Essex: AWL*; 1996; pp 445−451.
(15) Elber, R.; Karplus, M. Multiple conformations in proteins: A molecular dynamic analysis of myoglobin. *Science* **1987**, *235*, 318−321.
(16) Noguti, T.; Go, N. Structural basis of hierarchical multiple sub states of a protein. *Proteins* **1989**, *5*, 97−138.
(17) Brooks, C. L., III; Case, D. A. Simulations of peptide conformational dynamics and thermodynamics. *Chem. Rev.* **1993**, *93*, 2487−2502.
(18) Wolyness, P. G.; Onuchic, J. N.; Thirumalai, D. Navigating the Folding Routes. *Science* **1995**, *267*, 1619−1620.
(19) Shakhnovich, E. Folding nucleus: Specific or multiple? Insights from lattice models and experiments. *Folding Design* **1998**, *3*, 108−111.
(20) Kedem, K.; Chew, L. P.; Elber, R. Unit-vector RMS (URMS) as a tool to analyze molecular dynamic. *Proteins* **1999**, *37*, 554−64.
(21) Torda, A. E.; Van Gunstern, W. F. Algorithm for clustering molecular dynamic configurations. *J. Comput. Chem.* **1994**, *15*, 1331−40.
(22) Bravi, G.; Gancia, E.; Zaliani, A.; Pegna, M. SONHICA (simple optimized nonhierarchical cluster analysis): A new tool for analysis of molecular conformations. *J. Comput. Chem.* **1997**, *18*, 1295−1311.
(23) Abagyan, R.; Argos, P. Optimal protocol and trajectory visualization for conformational searches of peptides and proteins. *J. Mol. Biol.* **1992**, *225*, 519−532.
(24) Troyer, J. M.; Cohen, F. E. Protein conformational landscapes: energy minimization and clustering of a long molecular dynamic trajectory. *Proteins* **1995**, *23*, 97−105.
(25) Kazamirski, L. S.; Li, A.; Dagget, V. Analysis methods for comparison of multiple molecular dynamic trajectories: Applications to protein unfolding pathways and denatured ensembles. *J. Mol Biol.* **1999**, *290*, 283−304.
(26) Sullivan, D. C.; Kuntz, I. D. Conformation spaces of proteins. *Proteins* **2001**, *42*, 495−511.
(27) Lazaridis, L.; Karplus, M. "New View" of protein folding reconciled with the old through multiple unfolding simulation. *Science* **1997**, *278*, 5345, 1928−31.
(28) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamic calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.
(29) MacKerrel, A.; Bashford, J. D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kucszera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J. All-atom empirical potential for molecular modeling and dynamic studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586.
(30) Becker, O. M.; Levy, Y.; Ravitz, O. Flexibility, conformational spaces, and bioactivity. *J. Phys. Chem. B* **2000**, *104*, 2123−35.
(31) Lazaridis, T.; Karplus, M. Effective energy Function for Proteins in Solution. *Proteins* **1999**, *35*, 2, 133−152.
(32) Wang, J. X.; Dipasquale, A. J.; Bray, A. M.; Maeji, N. J.; Spellmeyer, D. C.; Geysen, H. M. Systematic study of substance P analogues. *Int. J. Pept. Protein Res.* **1992**, *41*, 1096.
(33) Cascieri, M. A.; Huang, R. R.; Fong, T. M.; Cheung, A. H.; Sadowski, S.; Ber, E.; Strader, C. D. Determination of the amino acid residues in substance P conferring selectivity and specificity for the rat neurokinin receptors. *Mol. Pharm.* **1992**, *41*, 6, 1096−9.
(34) Cascieri, M. A.; Ber, E.; Fong, T. M.; Sadowski, S.; Bansal, A.; Swain, C.; Seward, E.; Frances, B.; Burns, D.; Strader, C. D. Characterization of the binding of a potent, selective, radioiodinated antagonist to the human neurokinin-1 receptor. *Mol. Pharm.* **1992**, *42*, 3, 458−63.
(35) Sadowski, S.; Huangm, R. R.; Fong, T. M.; Marko, O.; Cascieri, M. A. Characterization of the binding of [125I-iodo-histidyl, methyl-Phe7] neurokinin B to the neurokinin-3 receptor. *Neuropeptides* **1993**, *24*, 6, 317−9.
(36) Barker, F. S. *The Elements of Logic*; McGraw-Hill Higher Education: 1988.

CI0103471