Geometric Versus Topological Clustering: An Insight Into Conformation Mapping

Oren M. Becker[1]*
[1]*Department of Chemical Physics, School of Chemistry, Tel Aviv University, Ramat Aviv, Tel Aviv, Israel*

ABSTRACT **Clustering molecular conformations into "families" is a common procedure in conformational analysis of molecular systems. An implicit assumption which often underlies this clustering approach is that the resulting geometric families reflect the energetic structure of the system's potential energy surface. In a broader context we address the question whether structural similarity is correlated with energy basins, i.e., whether conformations that belong to the same energy basin are also geometrically similar. 'Topological mapping' and principal coordinate projections are used here to address this question and to assess the quality of the 'family clustering' procedure. Applying the analysis to a small tetrapeptide it was found that the general correlation that exists between energy basins and structural similarity is not absolute. Clusters generated by the geometric 'family clustering' procedure do not always reflect the underlying energy basins. In particular it was found that the 'family tree' that is generated by the 'family clustering' procedure is completely inconsistent with its real topological counterpart, the 'disconnectivity' graph of this system. It is also demonstrated that principal coordinate analysis is a powerful visualization technique which, at least for this system, works better when distances are measured in dihedral angle space rather than in cartesian space. Proteins 27:213–226** © **1997 Wiley-Liss, Inc.**

Key words: conformation space; potential energy surface; connectivity; topological mapping; family clustering; principal coordinate projections; visualization

INTRODUCTION

Conformational analysis is an important computational tool used to investigate structure and flexibility of small biomolecules. It is used both to locate the most stable structure of small to medium-sized molecules[1,2] and to analyze their overall flexibility.[3] It is also a major tool in mapping the conformational landscape of biomolecules, which is the key for understanding the protein folding process.[4,5]

Conformational analysis is often applied to peptides in the context of rational drug design, as it is well known that conformational considerations play an important role in determining the specificity and the potency of peptide drugs.[6–8] However, analyzing the precise relationship between the structure of these bioactive peptides and their activity is difficult due to their inherent flexibility, which allows them to adopt multiple conformations. It is believed that knowledge of the conformational characteristics of bioactive peptides gained through computational conformational analysis will provide a structural basis for the design of synthetic peptides and peptidomimetic molecules.

Analyzing conformation space, even for a relatively small molecule, is a very demanding computational task because conformation spaces are typically of an extremely high dimensionality. A molecule with N atoms has $3N$ degrees of freedom, and its corresponding conformation space is $3N$-6-dimensional. As a result, even small molecules have a very large conformation space (e.g., a small heptapeptide has a conformational space of about 100–150 dimensions). Understanding properties in such high dimensional spaces and analyzing the potential energy hypersurfaces defined over them is therefore a very complicated tasks.

A necessary first step in conformational analysis studies is the generation of a large sample of molecular conformations. Many alternative methods are available for sampling molecular conformations, each with its own advantages and limitations. In one common sampling procedure, known as "quenched trajectories," conformations are sampled by high-temperature molecular dynamics trajectories, which are quenched either by direct minimization or by simulated annealing.[3,9–11] Other sampling procedures include Monte Carlo sampling followed by minimization,[3,12,13] identification of conformations along least energy paths[14] and many more. A limitation common to all sampling procedures is that, although they generate large conformation samples,

Contract grant sponsors: Israel Science Foundation and Tel Aviv University.

*Correspondence to: Oren M. Becker, Dept. of Chemical Physics, School of Chemistry, Tel Aviv University, Ramat Aviv, Tel Aviv, Israel. E-mail: becker@sapphire.tau.ac.il.

Received 9 May 1996; Accepted 19 July 1996.

it is hard to assess the quality of the sampled ensemble, namely: What part of the molecular conformation space was actually sampled? How are the sampled conformation distributed in that space? Is it a representative sample?

After a large enough sample of conformations is generated the next step is to analyze them to extract the desired structural information. The conformation sampled are usually sorted according to two properties: the potential energy associated with the conformation and its distance from other molecular conformations. The quantity used for measuring conformational distances is often the root mean square distance (RMSD) between two conformations. The RMS distance d_{ij} between conformation i and conformation j of a given molecule is defined as the minimum of the functional

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

where N is the number of atoms in the summation (either all atoms or a subset of the molecule's atoms, e.g., only the heavy atoms), k is an index over these atoms and **r***k(i),* **r***k(j) are the cartesian coordinates of atom k* in conformations i and j. The minimum value is obtained by an optimal superposition of the two structures. The resulting RMSD data is compiled into a *distance matrix,* Δ, where the elements Δ_{ij} are the RMS distances between conformation i and conformation j. In a limited number of cases the distance matrix in itself is a useful analysis tool. For example, when sampling along a molecular dynamics trajectory the matrix can have a block diagonal form, indicating that the trajectory has moved from one conformational basin to another. Nonetheless, even in this case, the matrix does not yield reliable information regarding the size and shape of the respective basins. In general, the distance matrix requires further processing.

More detailed analysis of the conformation sample can be achieved by "family clustering," a method that clusters together conformations based on *geometric similarity.* Staring from a selected conformation (often that of lowest energy), all conformations that are within a given cutoff distance from this structure are grouped together into a "cluster" or a "family" of structures. Next, a conformation from the remaining structures is selected, and the process is repeated until no more families are found. Overlapping families are grouped together, and the resulting "clusters" are taken as the entities that characterize the molecular conformation space. Representative conformation from each structural family are sometimes analyzed further.

While family clustering is definitely a useful technique, it should be recognized that interpretation of its results (in terms of conformation space) often involves a hidden assumption. In general, the goal of conformational analysis is to probe the potential energy surface of the system because potential energy determines which conformations are favorable and which are not, and it determines the observed distribution of conformations. The real physical processes that determine which conformations are grouped together and which are not are the rapid kinetic transitions within the "energy basins," which are regions of structures connected by low barriers. Thus, an ideal conformational analysis procedure would identify conformations that are part of the same energy basin (i.e., are connected by low barriers) and distinguish between conformations that belong to different energy basins (i.e., are separated by high barriers). The implicit assumption behind family clustering is that the resulting geometrically defined conformation families reflect the real energy basins on the molecular potential energy surface. It is thus implicitly assumed that the barriers connecting conformations within a similarity cluster are significantly lower than the barriers separating one cluster from another, that is, that conformations that make up such a cluster are mutually accessible. This notion is supported by some partial results obtained from protein simulations[15] (L.S.D. Caves, J. Evenseck, and M. Karplus, unpublished observations).

Nonetheless, it is clear that structural similarity (which also depends on the similarity measure used) and kinetic accessibility or connectivity are not necessarily correlated. One cannot overrule the possibility of similar structures belonging to different energy basins (i.e., separated by high barriers) or of quite different structures belonging to the same energy basin (i.e., separated by a low barrier). Small deformations that have little effect on the structure can cause significant changes in energy. Therefore, the question of how well geometric clustering really reflects the energy structure of the potential energy surface of peptides and proteins remains open. Much of the difficulty associated with this question stemmed from the fact that until recently there were only limited tools for independent analysis of the energy structure of these surfaces. Attempts to partially characterize the underlying potential energy surfaces were mainly based on structural similarity of local minima.[9,15,16] Few more elaborate studies were able to characterize multidimensional potentials in terms of their barrier distribution[14] and partial connectivities,[11] but none of these was detailed enough to quantitatively assess the question posed above. It should be noted that, if conformation clustering is performed in a reduced subspace, such as Rackovsky's analysis of protein structures by using a four-Cα yardstick,[17] the potential energy analogy is completely avoided.

Recently, Becker and Karplus have introduced a new analysis method, named *topological mapping,* which analyzes the energy structure of multidimensional potential energy surfaces.[18] This analysis maps the multidimensional surface onto a simple two-dimensional graph, highlighting the energy basins on the surface and pointing to their overall connectivity. In this paper we use topological mapping of a small peptide system, the tetrapeptide isobutyryl-(ala)$_3$-NH-methyl (IAN), to quantitatively assess whether the geometry clusters generated by family clustering really reflect the underlying energy basins or whether they introduce errors into the analysis. In particular, clustering in cartesian space will be compared to clustering in dihedral angle space. Our analysis will make use of a powerful visualization technique, principal coordinate analysis, which enables a projection of the multidimensional conformation space onto a low-dimensional subspace. This technique will help us check whether proximity in conformation space is correlated with energy basins. Namely, whether conformations that are connected by low barriers are also geometrically similar.

METHODS

Topological Mapping

The basic idea behind topological mapping, introduced by Becker and Karplus,[18] is to partition conformation space into its component energy basins. At any energy level E the molecular conformation space can be partitioned into disconnected regions, basins, each consisting of local minima that are connected by barriers lower than E. The topological mapping procedure follows the way these basins connect (disconnect) as a function of increasing (decreasing) energy E.

A molecule of N atoms has a 3N-6-dimensional conformation space $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_N) \neq \mathrm{R}^{3N\text{-}6}$, where the $\mathbf{r}_i$ are the vectors giving the position of atoms i in the three-dimensional space. The potential energy of the system, Φ is a function defined over conformation space $\Phi(\mathbf{r}_1, \ldots, \mathbf{r}_N)$ and is characterized by a large number local minima, defining a discrete set indexed by α. Direct minimization is the simplest way to get (i.e., map) from the multidimensional continuum $\mathrm{R}^{3N\text{-}6}$ to the discrete set of minima $\{\alpha\}$.

Following Stillinger and Weber,[16] let $R(\alpha) \subset \mathrm{R}^{3N\text{-}6}$ denote the set of system configurations $\mathbf{r}$ which map to a local minimum α by direct minimization. $R(\alpha)$ is a connected set and the different $R(\alpha)$ are disjoint. However, although this simple map partitions the potential energy surface in a physically meaningful way (i.e., it gives the relation between any point in $\mathrm{R}^{3N\text{-}6}$ and the nearest local minimum α), the information content of this map is limited as it does not contain information about the barriers between the minima. This limitation results in that the $R(\alpha)$ give dynamical information *only* at the $T \rightarrow 0$ limit, where any barrier, regardless of how small, traps the system in a given local minimum.

To better characterize the system at finite temperature, it is desirable to group minima that are connected by low barriers and separate them from minima connected by high barriers. This can be achieved by introducing superbasins

$$R^*(\alpha') = \cup R(\mathrm{a}) \tag{2}$$

which are the union of all $R(\alpha)$ sets connected by barriers lower than some defining energy. This can be formulated in terms of E in a microcanonic ensemble, $R^E(\alpha')$, or in terms of kT in a canonic ensemble, $R^T(\alpha')$. The symbol α' refers to the lowest minimum in $R^*(\alpha')$, that is,

$$\alpha' = \min \{\alpha \mid \alpha \in R(\alpha) \subset R^*(\alpha')\} \tag{3}$$

Using these definitions we introduce a new mapping procedure $M^*(\mathbf{r})$ that partitions the multidimensional conformation space $\mathrm{R}^{3N\text{-}6}$ into superbasins at energy E (or temperature T) and map it onto the smaller set of minima $\{\alpha'\}^*$ defined by Equation (3),

$$M^*(\mathbf{r}) \colon R^{3N-6} \rightarrow \{\alpha'\}^* \tag{4}$$

where the asterisk stands for E in a microcanonic ensemble (i.e., $M^E(\mathbf{r})$ and $\{\alpha'\}^E$) or T in a canonic description (i.e., $M^T(\mathbf{r})$ and $\{\alpha'\}^T$). The resulting superbasins are also disjoint open sets.

The new mapping procedure is very similar to the results obtained by simulated annealing minimization. Its main advantage, especially in comparison with the direct minimization map, is that, in addition to partitioning the potential energy surface in a physically meaningful way, it also reflects the basins' connectivity at different energies or temperatures. From this map and its variation as a function of the energy one can obtain the system's disconnectivity graph $G^E(\Phi)$. This graph, which characterizes the multidimensional potential surface Φ, indicates how the energy basins split as the overall energy E is decreased. Since different branching characteristics represent different topographic features this two-dimensional treelike graph can be interpreted in terms of the *shape* of the potential hypersurface.

Despite the similarity between the above mapping and simulated annealing, it is not practical to use simulated annealing to construct the map. These maps are better reconstructed from large collections of local minima and information regarding the barriers that connect them. The minima sample can be generated by any suitable sampling procedure (e.g., systematic, biased, Monte Carlo, high-temperature molecular dynamics) and the local barriers can be obtained by employing path-finding algorithms between pairs of local minima.[14,19] Although, this data collection effort is computationally intensive, it is feasible on today's computers. The actual construction of the graph from the data is very fast.

Using this method Becker and Karplus[18] analyzed the potential energy surface of the alanine tetrapeptide (138 local minima, about 400 distinct local barriers) resulting in some interesting conclusions. The overall shape of the potential surface of this peptide was determined and found to be that of a multidimensional *funnel* (although the surface has some nonfunnel regions too, for details see the subsection Sampling and Energy Basins below). It was also found that in this system the funnel is centered on the global minimum, a fact that was speculated yet never proven before. Analyzing a variant of this map with regard to temperature enabled them to calculate the basin-to-basin kinetics in this system, which proved to have a reach structure.

Principal Component Analysis

How to make sense of large bodies of data has always been a main concern for statisticians, and methods developed to address it are often used in molecular analysis. A unique aspect of this general issue is the problem of visualization of the multidimensional molecular conformation spaces that are under consideration here. The method of *principal coordinate analysis,* which was introduced by Gower in 1966[20,21] but only recently applied to molecular systems, is a powerful tool that helps address this problem.

Phrased in molecular terms, the idea behind principal coordinate analysis is that cartesian coordinates may not be the best set of coordinates for analyzing conformation space. As in many other cases, it may be that by selecting a different set of coordinates we will be able to reveal properties that were hard to observe in the original coordinate system. For example, a one-dimensional distribution of points set along the diagonal in an X-Y plane, may seem two-dimensional when projected on the original coordinates. A simple 45° rotation of the axes, which aligns the points with a new X′ axis, yields a broad projection on the X′ direction and a very narrow projection on the new Y′ axis, revealing the true one-dimensional character of the distribution. Understanding this concept, the question is what would be the best coordinate set to choose. Gower's solution was to choose an axes set that maximizes the variance of the projection along orthogonal directions. In practice, this amounts to diagonalizing the matrix of squared distances, d_{ij}^2, after centering it. The resulting eigenvalues (normalized) give the percentage of the projection of the original distribution on the new coordinate set, while the eigenvectors indicate the position of the original points in the new coordinate system.

This analysis technique was recently successfully applied by Caves, Evenseck, and Karplus (unpublished observations) to conformations of the protein crambin (a 46-amino acid molecule) based on their cartesian distances. It was found that after this transformation almost all of the information about the distribution was contained in less than 10 dimensions, with 55% of it projected into merely two dimensions. This means that the distribution of conformation in conformation space, at least in this case, was extremely nonhomogeneous and plotting its projection on the suitable plane (plane of maximum variance) contained a very large portion of the information about it. In other applications this method was used by Abagyan and Argos[12] to visualize trajectories of the peptide Met-enkephalin based on distances in dihedral space and by Troyer and Cohen[22] for visualizing a 1-ns molecular dynamics trajectory of BPTI.

Technically, principal coordinate analysis, which is closely related to principal component analysis, starts with a matrix of pairwise distances d_{ij} between n reference conformations P_i ($i = 1, 2, \ldots, n$) and results in the new set of coordinates. The procedure is as follows[20]:

1. Define a $n \times n$ symmetric matrix A such that

$$A_{ij} = -1/2 d_{ij}^2 \quad \text{and}$$

$$A_{ii} = 0 \quad (i,j = 1, 2, \ldots, n) \tag{5}$$

This definition constructs A as a special case that conforms to the relationship

$$d_{ij}^2 = A_{ii} + A_{jj} - 2A_{ij} \tag{6}$$

which is derived from a well-known property of normalized latent vectors.

2. To guarantee that the matrix A has a zero root it is "centered," so that the sum of every row and of every column of A is zero. This centering is performed by the following transformation

$$\mathbf{A}_{ij}^* = \mathbf{A}_{ij} - \langle \mathbf{A}_{ij} \rangle_i - \langle \mathbf{A}_{ij} \rangle_j + 2\langle \mathbf{A}_{ij} \rangle_{ij} \tag{7}$$

where $\langle \ldots \rangle_k$ is the mean over all specific indices $k = i, j, ij$. This transformation does not effect the distance between P_i and P_j given by Equation (1).

3. Now, similar to principal coordinate analysis, solve for $A^* = B\Lambda B^T$ using standard matrix algebra, where B is the matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues.

4. Scale each eigenvector (the column vectors of B) so that its sum of squares is equal to its corresponding latent root (eigenvalue), that is, $B^T B = \Lambda$

The ith row of the scaled matrix B may now be regarded as the new coordinates of the set of points P_i whose distances are given by the best approximations to $(A_{ii} + A_{jj} - 2A_{ij})^{1/2}$ in the chosen number of dimensions. The eigenvalues represent the variation along the corresponding axis. If the eigenvalues and corresponding eigenvectors are sorted in decreasing order the first eigenvector represents the axis of maximal variance, the second is axis with the second largest variance and so forth. Projection of the distribution onto the first two or three dimensions represents the best planar and best 3D projections of the distribution.

Whether or not a valuable reduction in dimensionality is obtained depends on the eigenvalues. It is easy to see why good representation can be achieved in a reduced number of dimensions when some of the λ_i are small. If λ_n is small than the contribution $(B_{in} - B_{jn})^2$ to the distance between P_i and P_j will also be small. In fact the sum of squares of the coordinates along the nth axis is λ_n by definition. As in all principal component analysis, the sum of squares of the residuals (i.e., the contributions perpendicular on to the reduced k-dimensional representation) will be the difference between the trace of Λ and the sum of its k largest eigenvalues. It turns out that in all molecular systems studied by this method (the proteins crambin [Caves, Evenseck, and Karplus, unpublished observations] and BPTI[22] and the peptides Met-enkephalin[12] and RGDS analogues [Becker, unpublished observations]), there are rarely more than 10 eigenvalues of significant magnitude. In fact, in all these cases projection onto two-dimensions contained 55%-75% of the trace of Λ, and projection onto three dimensions contained 65%-85% of the trace. It should be noted however that the *partial* sampling, that was performed on these molecular systems, is at least partially responsible for the observed anisotropy. In general, due to the large size of polypeptide conformation spaces, principal coordinate analysis cannot be performed on the full space, and some sort of sampling will always be necessary for this kind of analysis.

Conformation Families

A common analysis method for molecular conformations is family clustering, in which sampled conformations (usually associated with local minima) are grouped into families of similar structures. In this work, local minima conformers were grouped into families based on the rms distance [Eq. (1)] using a hierarchical clustering procedure. This clustering procedure starts with the lowest energy conformer and groups together all the conformers that are within a given *cutoff distance* from any other member of that conformation cluster. The next cluster is based on the lowest minima outside of the first cluster and so forth. The hierarchical procedure ensures that the resulting clusters are disjoint, and that each cluster is more than the cutoff distance away from its neighboring clusters.

This clustering procedure was used to cluster the local minima conformations of IAN based on two rms measures: (1) the all atom rms distance in cartesian coordinates [Eq. (1)], and (2) the rms distance in dihedral space (details are given in the subsection Visualization in Dihedral Angle Space below).

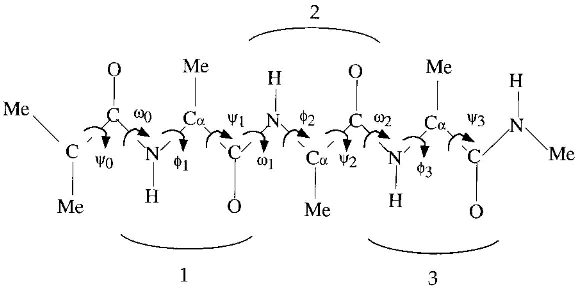

Fig. 1. The tetrapeptide IAN (isobutyryl-(ala)$_3$-NH-methyl). The soft torsions (ϕ,ψ) are on each side of the C(*i*) carbons. The ω torsion angle is associated with the peptide bond.

RESULTS

Sampling and Energy Basins

A molecular system for which extensive knowledge of the minima and local barriers exists is the tetrapeptide isobutyryl-(ala)$_3$-NH-methyl (IAN) in vacuum (Fig. 1), which was studied by Czerminski and Elber.[14] This peptide, which is a derivative of the (ala)$_4$ tetrapeptide, is the simplest model system that can form a full α-helical turn, including the stabilizing hydrogen bond. In the extended atom representation IAN has a 78-dimensional configuration space, but to a large extent the conformations of this peptide can be described in terms of a much smaller manifold, defined by the 7 soft torsions that correspond to its (ϕ, ψ) dihedral angles.

In their IAN study Czerminski and Elber identified 139 local minima and more than 500 barriers on the potential energy surface of this tetrapeptide.[14] The local minima were obtained using the polar hydrogen CHARMM potential energy function,[23] in which the methyl groups are represented as single extended carbon atoms. Czerminski and Elber's search algorithm started from the minimal energy path between the α helix and a β sheet conformations of IAN and then recessively determining minimal energy paths between any two minima encountered during the search. Each pathway calculation involved several refinement steps and the search ended when no new minima were found. The resulting set of 139 minima can be considered a good representation of the potential energy surface of this molecule, although some local minima are probably missing. These 139 minima are connected by the 502 transition states, of which only 393 are barriers that define connectivity between minima (the other barriers are either self-connecting paths or alternative higher lying saddle points between already connected minima).

Based on the data of Czerminski and Elber,[14] Becker and Karplus[18] performed a topological mapping analysis of IAN's conformation space. As mentioned above (in the subsection Topological Mapping), this analysis revealed that the potential energy

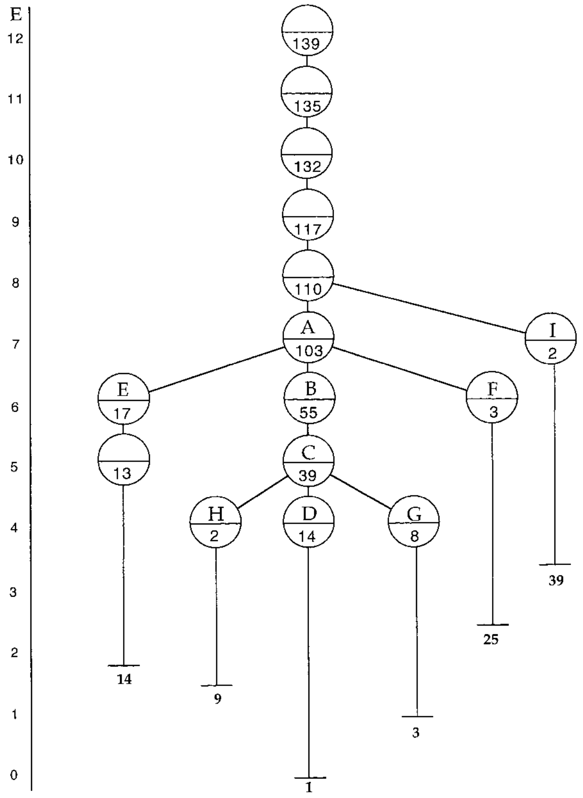

Fig. 2. A simplified disconnectivity graph of IAN (with level spacing of $\Delta E = 1.0$ kcal/mol), indicating how the main energy basins on IAN's potential energy surface split as a function of decreasing energy. The overall structure is that of a funnel. Each node in the graph represents a basin on the potential energy surface, defined as a set of local minima connected by barriers not higher than the energy defining that basin (the energy scale is given relative to the global minimum). All nodes specify the number of local minima included in the corresponding basins and the letters (A, B, . . . , I) are given for future reference. The vertical lines, stretching down from each of the low-level nodes, extend to the energy value of the deepest minima in those basins (the number associated with them are that minima's rank among the 139 local minima of IAN sorted by energy).

surface of this tetrapeptide has an overall funnel shape.

Figure 2 is a simplified version of the molecular disconnectivity graph of IAN (generated with level spacing of $\Delta E = 1.0$ kcal/mol), indicating how the main energy basins on IAN's potential energy surface split as a function of decreasing energy. Each node on the graph represents a basin on the molecular potential energy surface. Following the tree from its root (the ergodic basin at the top) down the energy scale two observations can be made: (1) As the energy decreases one sees a contraction in the size of the central basin (from all 139 minima at the root down to 103 minima at A, 55 minima at B, and 39 minima at C); and (2) there are smaller basins which branch off the central basin (e.g., the 2 minima of basin I or the 17 minima of basin E). These two observations are the marks of a multidimensional funnel on potential energy surfaces.

Since the energy basins depicted in this graph are the real physical partitioning of the system, that is, the partition that governs the system's kinetics, they will be used in the present study as a reference both for the visualization techniques and for assessing the family clustering technique.

Visualization in Cartesian Space

Visualization of molecular conformation spaces is highly desirable as it could support insights into the role of conformations in protein folding and rational drug design. Unfortunately, the size of these spaces make simple visualization techniques impractical, and conformational analysis studies must resort to nonvisual measures (alongside some pictures of characteristic conformations). Even the topological disconectivity graph discussed above, which reflects the overall structure of the potential energy surface, cannot qualify as a readily available visualization technique due to the elaborate calculations required for its construction.

Recently a few research groups used principal coordinate analysis[20] to project dynamical trajectories[12,22] and molecular conformations [Caves, Evenseck, and Karplus, unpublished observations] on the plane of maximal variance (see subsection Principal Component Analysis above). We apply this method here to the tetrapeptide IAN in order to study whether conformations that belong to the same energy basins are also similar geometrically.

To obtain the projections of principal coordinate analysis it is first necessary to construct a distance matrix for the system. We used the rms distance measure in cartesian coordinates (Eq. 1) to construct a 139×139 distance matrix, which was then centered and diagonalized. The resulting eigenvalues, depicted in Figure 3(d) and Table I, represent the relative variance along the corresponding axis (normalized to the sum of all positive eigenvalues; the negative eigenvalues amount to 10% of the total). The results indicate that this system shows a highly nonisotropic distribution with 83.8% of the total variance contained within a 10-dimensional subspace. Specifically, the projections along the first three dimensions are 27.9%, 16.7%, and 12.8% of the total variance, holding 57.4% within the three-dimensional subspace of maximal variance. This breaks down to 44.6% for the plane of maximal variance, 40.7% for the plane defined by dimensions 1 and 3, and 29.5% on the plane defined by dimensions 2 and 3 (see Table I).

Although the distribution of conformations in IAN is clearly not isotropic, it is less so in comparison with the distributions obtained for other molecules (crambin [Caves, Evenseck, and Karplus, unpublished observations] BPTI,[22] Met-enkephalin[12] and RGDS analogues [Becker, unpublished observa-

TABLE I. Eigenvalues of the Principal Coordinate Analysis Projection of the IAN Tetrapeptide Using Different Definitions of rms Distances*

	Cartesian space (%)	Dihedral space (full) (%)	Dihedral space (without termini) (%)
1st dimension	27.9	21.3	28.7
2nd dimension	16.7	18.7	20.8
3rd dimension	12.8	15.5	18.0
4th dimension	9.7	13.5	7.7
1–2 plane	44.6	40.0	49.5
1–3 plane	40.7	36.8	46.7
2–3 plane	29.5	34.2	38.8
1-2-3 space	57.4	55.5	67.5
First 10 dimensions	83.8	94.8	97.9

*The eigenvalues, which correspond to the variance in the specific dimensions, are given relative to the trace of the eigenvalue matrix Λ, excluding the small contribution of the negative eigenvalues (for details see text).

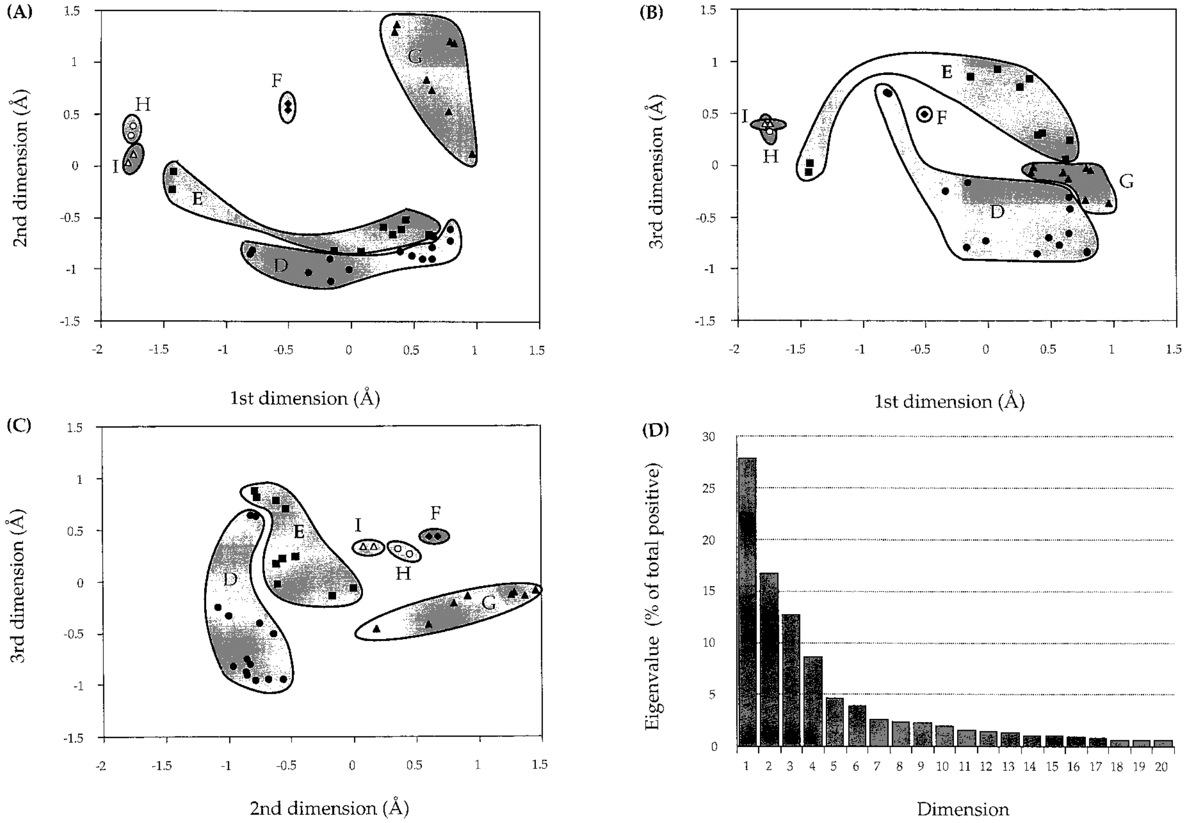

Fig. 3. Principal coordinate analysis of IAN based on all atom rms distances in cartesian space. Projected are all the conformations that belong to the basic energy basins on IAN's potential energy surface (basins D, E, F, G, H, I, as defined in Figure 2). Each point in these projections is a molecular conformations, and the conformations are grouped according to their basins assignments. The contours around the groups are schematic, and introduced only for clarity. **A:** Projection on the plane of maximal variance (defined by the 1st and 2nd dimensions). **B:** Projection on the plane defined by the 1st and 3rd dimensions. **C:** Projection on the plane defined by the 2nd and 3rd dimensions. **D:** The 20 largest eigenvectors (as a percent of the sum of positive eigenvalues).

tions]). This is reflected by the fact that the above numbers are smaller than those reported so far for other molecular systems (about 55%-70% for the first two dimensions), and by the fact that there is a substantial tail of not-negligible eigenvalues beyond the first 10 dimensions (which hold only 83.8% of the total variance). This behavior may be attributed to at least two factors: (1) In IAN we have an almost complete description of conformation space, while in the other systems the analyzed conformation samples

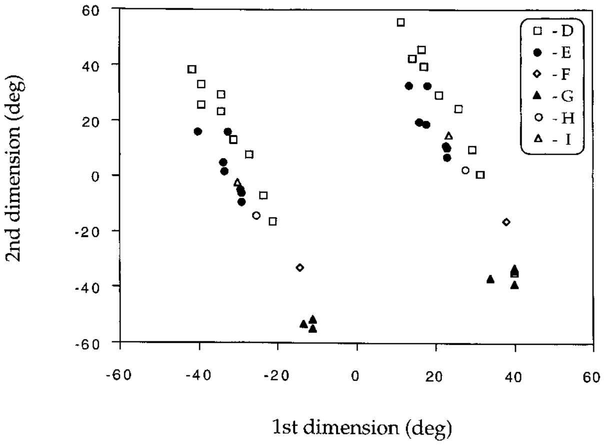

Fig. 4. Principal coordinate analysis of IAN based on full backbone dihedral angle rms distances (10 torsion angles). Shown is the projection on the plane of maximal variance (defined by the 1st and 2nd dimensions). The splitting in the projection is due to the effect of the terminal dihedral angles.

held only a partial sample of this space. Full conformation spaces are likely to be more isotropic than partial samples due to the possible anisotropy introduced by the sampling procedure (especially when sampling along dynamical trajectories). (2) The simplicity and inherent flexibility of the IAN tetrapeptide may be associated with a more isotropic conformation space in comparison the other more complicated and less flexible molecules.

Figure 3A-C show the projection of the basic energy basins in IAN's conformation space (Fig. 2) onto the three planes of maximal variance. Each point in these projections is a molecular conformations, and the conformations are grouped according to their basins assignments (The contours around the groups are schematic, and introduced only for clarity). The projections include basins D (containing the global minimum), G and H which are part of the central funnel, and basins E, F and I which branch off the central 'funnel'. This projection helps us visualize spatial proximity (i.e., structural similarity) in conformation space, defined in this case by the all atom rms distance in cartesian space. Given the nature of the analysis we expect similar structures to appear close to each other in this projection. Keep in mind that although this projection contains only 57.4% of the total variance, the rest of it is distributed in diminishing quantities along the other 136 dimensions of the distribution.

Using this visualization we can address the question whether there is a correlation between structural similarity and energy basins, that is, whether conformations that belong to the same energy basin are similar to each other. It was previously observed by Becker and Karplus[18] that in the IAN tetrapeptide there is a strong correlation between spatial proximity in conformation space and being a member of the same energy basin. This general observation is confirmed by the projections of Figure 3. The fact that almost all the points in a given energy basin can be easily grouped together is a clear indication of their structural similarity. However, we also see that this correlation is not absolute nor a necessary condition as there are clear exceptions to it. All three large basins (D, E and G) are spread over quite large areas, with some minima lying far away from the rest. Namely, some conformations that belong to these energy basins (i.e., connected to the other minima by low barriers) bear little structural similarity to the other conformations in them. The difference between structural similarity (in Cartesian space) and energy basins is more striking if we step up one level on the disconnectivity graph and look at basin C, which is a union of basins D, G, and H (Fig. 2). From the relative location of these basins we could not have guessed that these three basins are connected. The implications of these findings on the possibility of family analysis to identify energy basins on the surface is discussed in the subsection Family Clustering and Energy Basins below.

Visualization in Dihedral Angle Space

The results in the previous section indicate that IAN's conformation space is less nonisotropic in comparison to other studied systems. This may be a real property of the system, but it may also be a consequence of the measure we used to gauge structural similarity. Namely, it is possible that the all-atom rms distance in cartesian space is not the best distance measure for this system. A better choice may be a distance which is defined in the dihedral angle subspace, since the conformations of this tetrapeptide are easily specified by the (ϕ, ψ, ω) backbone dihedrals (Fig. 1).[14,18]

We repeated the principle coordinate analysis of IAN with a back-bone dihedral angle rms distance defined as

$$d_{ij} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\min[(\theta_k^{(i)} - \theta_k^{(j)})^2, (2\pi - \theta_k^{(i)} + \theta_k^{(j)2}]} \quad (8)$$

where N is the number of dihedral angles, and $\tau_k{}^{(i)}$, $\tau_k{}^{(j)}$ are the values of the dihedral angle τ_k in the two structures. In IAN the full (ϕ, ψ, ω) backbone dihedral space includes 10 torsion angles. The peptide-bond dihedral angles ω were included for completion and to cover the possibility of cis-trans isomerization of these bonds. The results of the principle coordinate analysis with the full dihedral distance matrix are given in Table I. We see that the overall quality of the projection in this space is better in comparison to the projection which was based on cartesian distances. The percent of the the total variance contained in the first 10 dimensions is now 94.8% (up from 83.8%) although the variance contained in the first three dimensions is similar, 55.5% in compari-

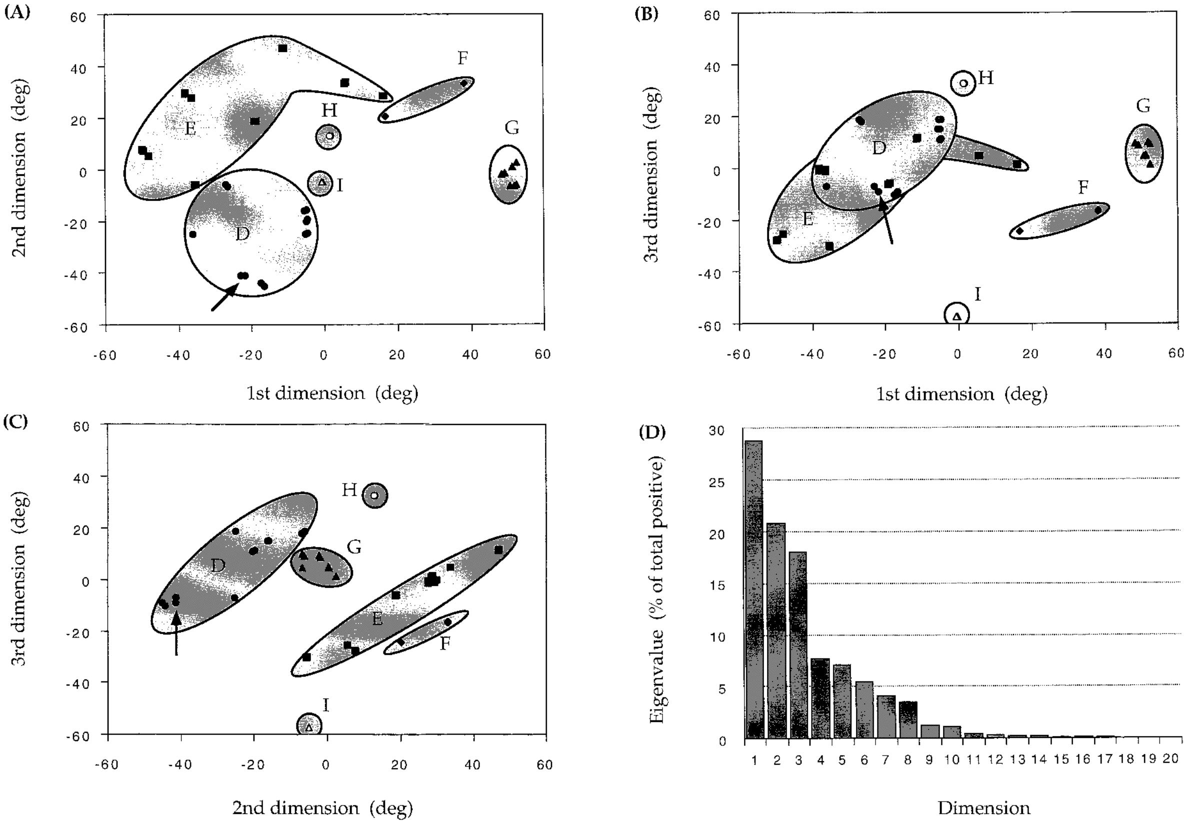

Fig. 5. Similar to Figure 3 but based on the backbone dihedral angle rms distances which excludes the two terminal dihedral angles (ψ_0 and ψ_3, leaving 8 torsion angles), with the letters indicating basin assignments. The global minimum is indicated by an arrow.

son to 57.4%. The better result in the first 10 dimensions is not surprising given the size of the original dihedral angle space.

Figure 4 depicts the projection of IAN's basic energy basins on the plane of maximal variance in dihedral space. This projection shows a splitting of conformation space into two very similar mirror images. A splitting so pronounced that it masks all other structural features. This splitting is easily attributed to the dual character of the N-terminal dihedral angle ψ_0. The detailed analysis of Becker and Karplus[18] showed that many IAN conformations come in pairs, differing only by a 180° rotation of the ψ_0 dihedral angle (between the values $+120°$ and $-60°$) and yielding a small energy difference of about 0.05–0.10 kcal/mol. The large 180° difference dominates the distance measure [Eq. (8)] and appears as the most prominent feature in the projection.

To filter out the effect of the terminal dihedral angles (mainly the N-terminal dihedral angle ψ_0 but also the C-terminal dihedral angle ψ_3) a third distance matrix was constructed. This time we used a backbone dihedral angle distance which excludes the two terminal dihedral angles (i.e., equation 8 with $N = 8$ torsions). The results of the principle coordinate analysis of this matrix are shown in Figure 5. From Figure 5D and Table I, we see that this projection is significantly better than the projection based on the cartesian distance. The percent of the total variance contained in the first 10 dimensions is 97.9% (compared to 83.8 using the cartesian distance) with 67.5% in the first three dimensions (compared to 57.4%) and 49.5% in the plane of maximal variance (compared to 44.6%). This means that the dihedral distance, after excluding the effect of the terminal torsion angles, is a better measure for structural similarity in IAN than the all atom rms in cartesian coordinates. The projection of energy basins on the three principle planes (Figures 5A-C), which shows a simpler clustering pattern, also supports this conclusion. Nonetheless, even in this projection basins D and E are spread over relatively large volumes, with only basin G showing up as a compact cluster.

Projection of the Topological Tree

So far we have seen that there is a strong, though not absolute, correlation between structural similarity and topological connectivity; that is, conformations that belong to the same energy basin often bear structural similarity to each other. This correlation

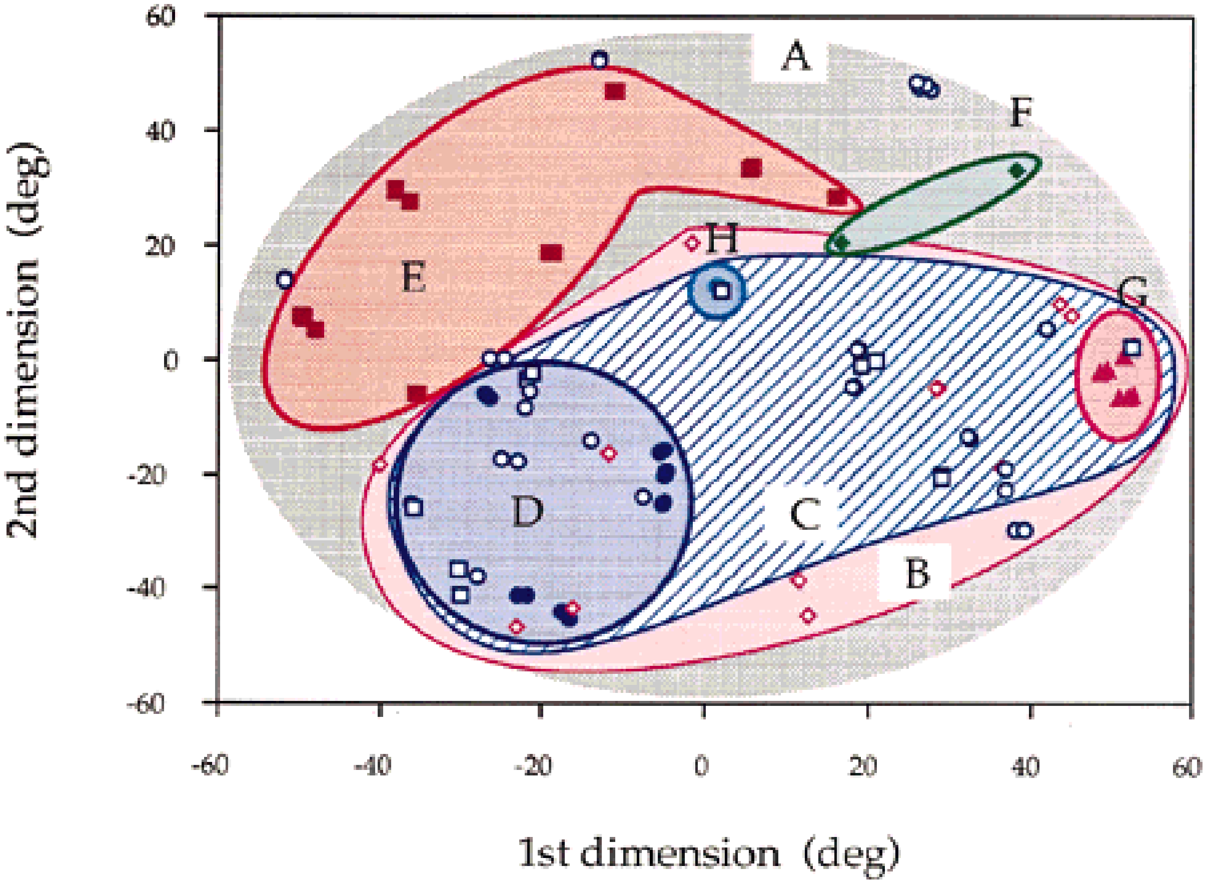

Fig. 6. A projection of the topological tree (Fig. 2), from basin A down, on the plane of maximal variance using the dihedral angle without termini projection (similar to Fig. 5A). The contours group minima according to their basin assignments (given by the letters defined in Figure 2). Blue square, minima that are part C but outside of sub-basins D, G and H; red diamond, minima that are part B but outside of C; blue circle, minima that are part A but outside of subbasins B, E, and F.

is more pronounced in dihedral angle space (after excluding the terminal dihedrals) than in cartesian space. It would be interesting to see whether this correlation is retained when going up the topological connectivity tree (Fig. 2). Figures 3 and 5 indicate that this is not a trivial question as basins that belong to the same 'parent' basin do not appear very close to each other in the projections.

Figure 6 shows the projection of the whole topological tree, from basin A down, on the plane of maximal variance using the dihedral angle without termini projection (similar to Figure 5A). In addition to the basic basins (D, E, F, G, H) this projections includes all the local minima that are included in the parent-basin but are outside of the lower level basins (totaling 103 conformations). For example, among the 39 minima of basin C there are 15 local minima that are not included in subbasins D (14 minima), G (8 minima), or H (2 minima). In a similar way, basin B includes 16 minima in addition to the 39 minima of basin C, and basin A includes 28 minima in addition to the minima already accounted for as part of basins B (55 minima), E (17 minima) and F (3 minima). The resulting picture shows that principal coordinate analysis is indeed a powerful visualization technique. Although this projection contains only 50% of the total variance it enables a clear visualization of the whole topological tree. Each successive level up the topological graph is clearly represented in this plane by a continues region. Basin C with its additional minima engulfs its sub-basins D, G and H, basin B slightly expands the geometrical boundary of basin C, and finally super-basin A engulfs the whole projection space. It should be stressed that despite the fact that this is only a partial visualization, this picture is a *real visualization* of the geometrical features of the attraction basins in molecular conformation space.

Family Clustering and Energy Basins

'Family clustering' is a common method for grouping together molecular conformations. With this

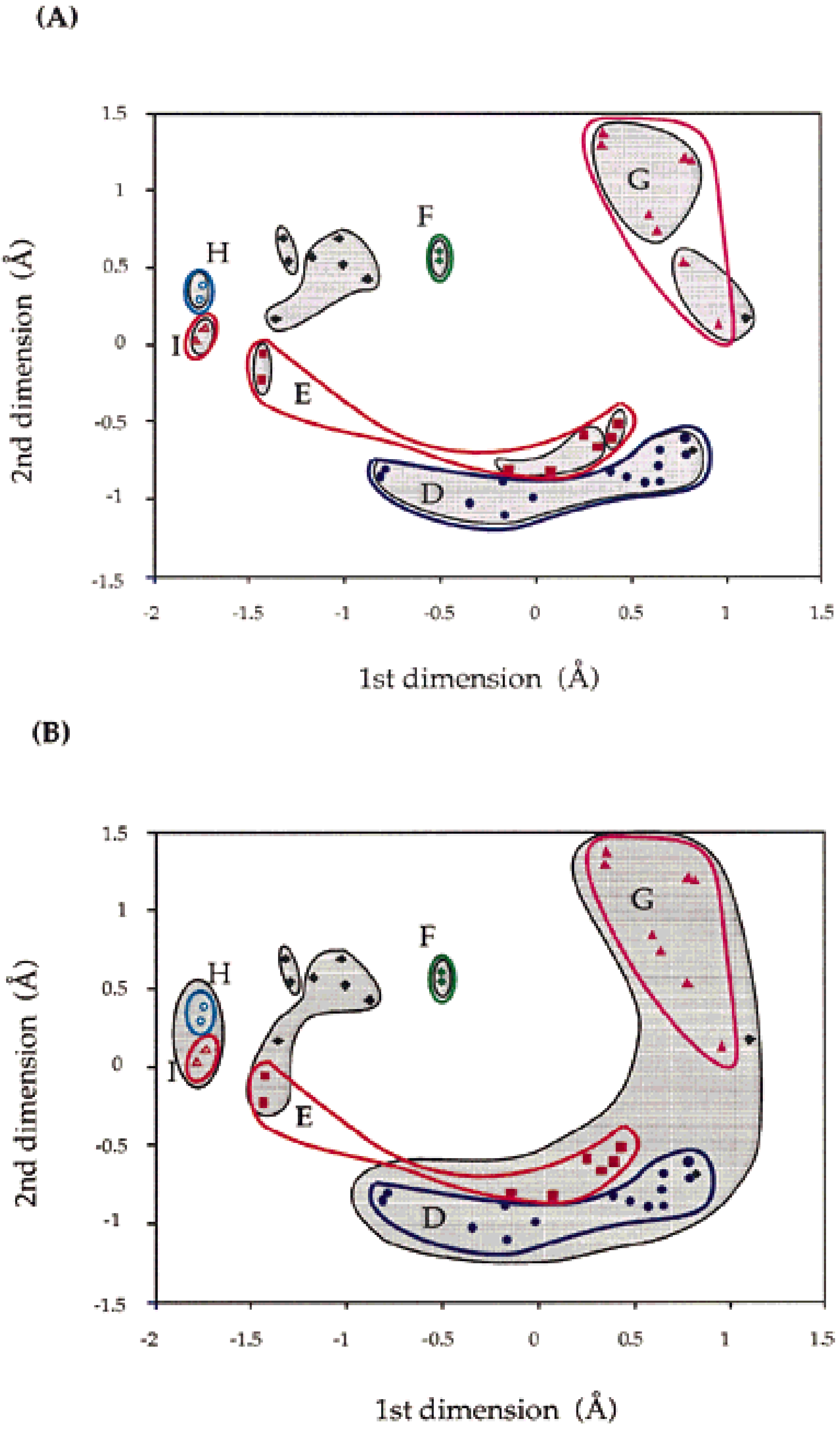

Fig. 7. A projection of the lowest 45 minima of IAN on the plane of maximal variance, based on the all atom rms distances in cartesian coordinates (similar to Figure 3A). The minima are grouped according to the energy-basin assignments and according to the results of family clustering. The color contours and letters minima but outside the basic energy basins are indicated by black diamonds (they are assigned to basins B, C or A). The shaded areas indicate minima grouped together by family clustering. **Top:** Family clustering with a cutoff distance of 1.00 Å. **Bottom:** Family clustering with a cutoff distance of 1.10 Å.

TABLE II. Distribution of Geometry Clusters Obtained by Family Clustering in Cartesian Coordinates Among the Actual Energy Basins of the System*

	Basin	D	E	F	G	H	I	Other
Clust	size	14	8	2	8	2	2	9
Cutoff distance 1.00 «								
1	15	14						1
2	6				6			
3	2					2		
4	3				2			1
5	2		2					
6	2		2					
7	4		4					
8	2			2				
9	2							2
10	5							5
11	2						2	
Cutoff distance 1.10 «								
1	30	14	6		8			3
2	4					2	2	
3	7		2					5
4	2			2				
5	2							2

*Clustering based on all-atom rms distances in cartesian coordinates of the lowest 45 minima of IAN: Cutoff distance 1.00 «, cutoff distance at 1.10 «.

TABLE III. Distribution of Geometry Clusters Obtained by Family Clustering in Dihedral Angle Space Among the Actual Energy Basins of the System*

	Basin	D	E	F	G	H	I	Other
Clust	size	14	8	2	8	2	2	9
Cutoff distance 35°								
1	4	4						
2	9				8			1
3	13	8				2		3
4	4		4					
5	2	2						
6	4		4					
7	2			2				
8	4							4
9	2						2	
10	1							1
Cutoff distance 40°								
1	20	14				2		4
2	15			2	8			5
3	8		8					
4	2						2	

*Clustering based on the backbone dihedral angle rms distances, excluding the two terminal dihedral angles of the lowest 45 minima of IAN: cutoff distance of 35°, cutoff distance of 40°.

method all conformations that are within a given cutoff distance from a reference conformation (or from other members of the family) are grouped into a conformation family. As discussed in the introduction, an assumption underlying the analysis of this clustering procedure is that the resulting families reflect the energy basins on the molecular potential energy surface. Namely, it is implicitly assumed that conformations that belong to the same 'family' are connected by low barriers, forming a kinetically connected set of structures. The results we obtained so far from the IAN system are inconclusive in terms of the likelihood of 'family clustering' to abide by this assumption. On the one hand, energy basins and structural similarity were found to be correlated, but on the other hand there are exceptions to this trend, and structures are widely spread even within the low lying basins.

To find whether family clustering can reproduce the actual energy basins we compare, on the same projected plane, the energy basins with the clusters obtained by family clustering. Figure 7 shows a projection of the lowest 45 minima of IAN on the plane of maximal variance. The projection is based on the all atom rms distance matrix in cartesian coordinates. Note, there are some differences between this figure and Figure 3A. While the previous figure showed all the local minima that are incorporated in the basic energy basins, here the lowest 45 minima are shown. As a result, some minima which are included in Figure 3A are not present here (e.g., one of the minima in basin F which is ranked number 88), while some minima that are included here are not included in Figure 3A (e.g., minima number 26 and 27 that are assigned to basin B).

Figure 7A and Table II show the results of family clustering obtained with a cutoff distance of 1.00 Å. Comparing these clusters to the energy basins we see that the clustering procedure was successful in reproducing the deepest energy basin (basin D), which includes the global minimum in this system. The 15 conformations of this cluster contain the 14 conformations of basin D and one very similar conformation that is assigned to basin A (outside of its subbasins). The geometrical clustering procedure was, however, unsuccessful in grouping together all the minima of basins G or E. These are split among several different clusters, without any indication that they are kinetically connected. Beside the large 15 member cluster of basin D, the other clusters were much smaller consisting of 6, 4, or less minima (Table II).

Family clustering is often used to generate a tree-like graph showing how the clusters connect as a function of increasing cutoff distance.[3] To see whether this family tree can reproduce the actual topological disconnectivity graph (Fig. 2) we repeated the same clustering procedure at several cutoff values. At values smaller than 1.00 Å the clusters appeared even more fragmented and at 1.05 Å the picture was identical to that at 1.00 Å. Increas-

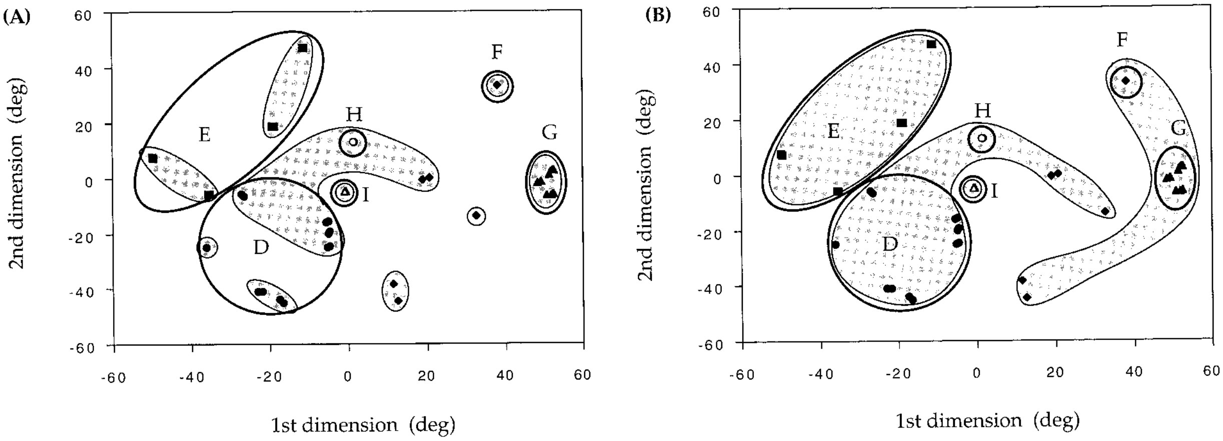

Fig. 8. Similar to Figure 7 but with a projection based on the backbone dihedral angle rms distances excluding the two terminal dihedral angles (similar to Figure 5A). **A:** Family clustering with a cutoff distance of 35°. **B:** Family clustering with a cutoff distance of 40°.

ing the cutoff distance to 1.10 Å, however, produced a substantial change in the clustering pattern (Fig. 7B). At this cutoff value the largest cluster includes 30 local minima, and engulfed all of basins D and G and large portion of basin E (6 of the 8 minima) (Table II). This clustering is inconsistent with the real 'disconnectivity' graph of the system (Fig. 2), which shows that basins D, G, and H are the ones that are kinetically connected, while basin E joins them only higher (at the level marked A). A similar misrepresentation of the real connectivity is found in the cluster that groups together basins H and I, which are very far away on the disconnectivity graph. Therefore, the family tree that is generated by family clustering of this system is completely inconsistent with the real disconnectivity graph of this system.

To check whether family clustering does better with the dihedral angle distance we repeated the family clustering analysis with the backbone dihedral angle rms distances (excluding the two terminal dihedral angles). The results, presented in Figure 8 and Table III, are given for two cutoff values: 35° and 40°. At the smaller cutoff value the clustering procedure was able to group correctly basin G (which in this projection is very tightly packed) but failed to group together the conformations that make up basins D and E. In particular, the global minimum is not even part of the largest cluster (a 13-member cluster) but rather of a small four member cluster. In addition it groups together part of basin D with basin H. At the higher cutoff distance, basin E is clustered correctly, but basin D is still connected to basin H and basin G is joint together with basin F (which is far off on the actual 'disconnectivity' graph).

CONCLUSIONS

The work presented in this paper indicates that principle coordinate analysis is a very powerful visualization method for presenting and discussing molecular conformation spaces. Although the projections are only partial, these low dimensional subspaces captures much of the geometric properties of the full conformation space. Troyer and Cohen[22] are correct in their comment that these projections must be carefully interpreted, as distortions in the relationship between the points can result. Conformational volumes or reconstruction of exact distances are among the properties that suffer most from these distortions. However, this study has shown that at least in the IAN system these distortions are not very severe when trying to visualize the *overall* organization of conformation space. In particular, it was shown projections of the geometry defined clusters, that result from family clustering, show up in the projections as continuous areas. One of the reasons why the projection gives an adequate picture, in spite of the fact that only about 60%-70% of the variance is contained within a three-dimensional subspace, is that most of the rest of the variance is distributed in small quantities over a very large number of dimensions.

Using this visualization method with the alanine tetrapeptide derivative (IAN), for which a relatively complete understanding of the potential energy surface is known, we were able to demonstrate to what extent is structural similarity correlated to the actual energy basins on the potential surface. It was shown that for this system there is a relatively strong qualitative correlation between the two, although it is not absolute. Most conformations that belong to the same energy basin are indeed similar to each other, but the spread of conformations (in conformation space) can be quit wide and there are exceptions to this general trend. There are conformations that belong to a given energy basin, but bear little structural similarity to the other conformations in the basin. It was also shown that for the IAN system simpler projections are obtained using the (ϕ, ψ, ω) backbone dihedral angle distances (excluding the terminal dihedrals) rather than with the all-atom rms distance measure in cartesian coordinates.

The principal coordinate visualization technique enables us to check whether the often used procedure of family clustering can reproduce the energy-basin structure of this system, or whether the above structural correlation is not strong enough to be correctly used in a clustering procedure. Comparing the clusters obtained by the hierarchical geometrical clustering procedure at different cutoff distances (both the cartesian distance and the dihedral angle distance) it was shown that the purely geometrical clustering procedure was *unable* to reproduce the correct connectivity and basin structure of the surface. With both distance measures while some energy basins were correctly reproduce other basins were either split into several smaller clusters or grouped incorrectly to other basins. It was also shown that the family tree, obtained by repeating the clustering procedure at different cutoff values, could not reproduce the correct topological disconnectivity graph of the system. Nonetheless, it should be stressed that using cartesian distances the family clustering procedure was able to identify the main basin, which includes the global minimum of the system. This is highly desirable task in conformational analysis.

To conclude, we see that it is necessary to be cautious when interpreting the results obtained from family clustering procedures. While this clustering may be useful, it should not be taken to reflect the underlying structure of the potential energy surface. In particular one should be careful in interpreting the resulting family tree and not equate it with the topological connectivity on the molecular potential energy surface. This connectivity can be represented only by the appropriate topological graph.

ACKNOWLEDGMENTS

I thank Dr. Leo S. D. Caves for stimulating discussions. This work is supported in part by a grant from the Israel Science Foundation and in part by a grant from the research authority of the Tel Aviv University.

REFERENCES

1. Howard, A.E., Kollman, P.A. An analysis of current methodologies for conformational searching of complex molecules. J. Med. Chem. 31:1669, 1988.
2. Leach, A.R. In: “Reviews in Computational Chemistry,” Vol. 2. Lipkowitz, K.B., Boyd, D.B. (eds.). New York: VCH Publishers, 1991:1–55.
3. Hempel, 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 36:282–301, 1995.
4. Bryngelson, J.D., Onuchic, J.N., Socci, N.D., Wolynes, P.G. Funnels, pathways, and energy landscape of protein folding: A synthesis. Proteins 21:167–195, 1995.
5. Karplus, M., Shakhnovich, E. In: “Protein Folding.” Creighton, T.E. (ed.). New York: W.H. Freeman, 1992:127–195.
6. Veber, D.F., Holly, F.W., Paleveda, W.J., Nutt, R.F., Bergstrand, S.J., Torchiana, M., Glitzer, M.S., Saperstein, R., Hirschmann, R., Conformationally restricted bicyclic analogues of somatostatin. Proc. Natl. Acad. Sci. USA 75:2636–2640, 1978.
7. Pierschbacher, M.D., Ruoslahti, E. Influence of Stereochemistry of the Sequence Arg-Gly-Asp-Xaa on binding specificity in cell adhesion. J. Biol. Chem. 262:17294–17298, 1987.
8. Shenderovich, M.D., Nikiforovich, G.V., Golbraikh, A.A. Conformational features responsible for the binding of cyclic anlogues of enkephalin to opioid receptors. Int. J. Peptide Protein Res. 37:241–251, 1991.
9. Stillinger, F.H., Weber, T.A. Dynamics of structural transitions in liquids. Phys. Rev. A 28:2408–2416, 1983.
10. Bruccoleri, R.E., Karplus, M. Conformational sampling using high-temperature molecular dynamics. Bioploymers 29:1847–1862, 1990.
11. Kunz, R.E., Berry, R.S. Statistical interpretatin of topographies and dynamics of multidimensional potentials. J. Chem. Phys. 103:1904–1912, 1995.
12. Abagyan, R., Argos, P. Optimal protocol and trajectory visualization for conformational searches pf peptides and proteins. J. Mol. Biol. 225:519–532, 1992.
13. Li, Z., Scheraga, H.A. Monte-Carlo minimization approach to the multiple-minim problem in protein folding. Proc. Natl. Acad. Sci. USA 84:6611–6615, 1987.
14. Czerminski, R., Elber, R., Reaction path study of conformational transitions in flexible systems: Application to peptides. J. Chem. Phys. 92:5580–5601, 1990.
15. Noguti, T., Go, N. Structural basis of hierarchical multiple substates of a protein: I-V. Proteins 5:97–132, 1989.
16. Stillinger, F.H., Weber, T.A. Packing structures and transitions in liquids and solids. Science 225:983–989, 1984.
17. Rackovsky, S. Quantitative organization of the known protein x-ray structures. I. Methods and short-length-scale results. Proteins 7:378–402, 1990.
18. Becker, O.M., Karplus, M. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. J. Chem. Phys. 105(24) (in press). 1996.
19. Fischer, S., Karplus, M. Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom. Chem. Phys. Lett. 194:252–261, 1992.
20. Gower, J.C. Some distance properties of latent root and vector methods used in multivarient analysis. Biometrika 53:325–338, 1966.
21. Gower, J.C. Adding a point to vector diagrams in multivariate analysis. Biometrika 55:582–585, 1968.
22. Troyer, J.M., Cohen, F.E. Protein conformational landscapes: Energy minimization and clustering of a long molecular dynamics trajectory. Proteins 23:97–110, 1955.
23. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplys, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. 4:187–217, 1983.