Principal coordinate analysis on a protein model

Nuran Elmaci and R. Stephen Berry
Department of Chemistry, The University of Chicago, 5735 S. Ellis Ave., Chicago, Illinois 60637

(Received 11 December 1998; accepted 23 February 1999)

A well-studied 46-bead protein model is the vehicle for examining principal coordinate analysis as a tool for interpreting topographies of complex potential surfaces. This study compares the effectiveness of several definitions of the comparison variable for revealing information about topographies. The extent of the information is ascertained by comparing the results of the various forms of principal coordinate analysis with results obtained from construction of interconnected monotonic sequences of linked stationary points (IMSLiSP) on the same surface. The conclusion is that the most powerful formulation of principal coordinate analyses for understanding protein folding and, in general, topographies of complex potentials, uses the changes in the set of interparticle distances as the definition of the comparison vector. However, even with this choice, the more efficient principal coordinate analysis is not able to reveal the extent of information contained in a more cumbersome IMSLiSP analysis. © *1999 American Institute of Physics.* [S0021-9606(99)51419-0]

INTRODUCTION

The analysis of conformations of proteins and the processes that change those conformations is central to the protein folding problem. How does a protein find one unique or a few three-dimensional configurations among the huge number of conformations? While the Leventhal paradox is essentially a facetious presentation of the issue, it caricatures the very general and fundamental question. ''What characteristics of its potential surface determine whether a system relaxes to some tiny fraction of all the local minima on that surface, or, alternatively, may find itself in virtually any of those many, many minima?'' In some diffuse way, it is generally recognized that the topography of the potential energy surface (PES) can reveal the answer to this question. However, since the number of local minima on the potential surface increases at least exponentially with the number of elements of the system, it is impossible—and undesirable to try—to obtain to full list of the minima, much less the far more numerous saddle points, of the surface and to visualize their relationships. We are forced to use some form of statistical sampling approach. But what should the elements of the statistical database be?

The first natural supposition is that the minima, or perhaps stationary points of all kinds, should be these elements. This approach with only the minima as the elements, coupled with a particular kind of distance measure, has been effective in categorizing the minima into ''basin'' structures.[1–5] This approach, the principal coordinate (PCoorA), is a variant of the principal component analysis (PCA), a means to select the degrees of freedom that are important in distinguishing one element of a set from others, and then using only those degrees of freedom to measure the differences among the elements—in this case, different structures. The application of these methods to a protein model, and establishment of its relationship to another, recently introduced method,[6–10] which we shall here call construction of ''interconnected

TABLE I. The accumulation of the percentage of the three largest eigenvalue fractions with the starting matrix taken as Cartesian coordinates, as pair distances, as bond angles, and as torsional angles for systems of various sizes: a, the largest; b, the sums of largest two; c, the sums of largest three eigenvalues.

Sample size	Upper energy	Case A			Case B			Case C			Case D		
		a	b	c	a	b	c	a	b	c	a	b	c
100	−0.5135	29.7	49.1	60.3	39.8	64.8	85.6	28.2	44.8	59.7	33.0	50.5	61
200	−0.5024	26.3	42.2	57.5	42.0	67.7	84.4	24.2	40.7	54.5	29.4	45.2	56
300	−0.4919	24.9	40.2	50.2	42.4	65.2	81.1	22.3	36.9	50.3	27.5	42.6	54
400	−0.4819	26.3	39.3	47.3	38.1	59.6	73.8	20.6	37.0	46.7	26.0	40.8	47
500	−0.4719	23.8	36.3	45.4	35.7	54.4	68.3	18.2	30.4	46.7	24.4	38.7	49
600	−0.4618	24.2	35.6	44.8	32.2	53.5	69.0	16.7	28.1	41.9	22.0	35.3	45
700	−0.4518	22.8	33.6	44.0	28.6	56.7	70.5	15.8	26.8	38.6	20.3	32.3	43
800	−0.4418	20.9	33.4	43.8	3.4	58.7	70.4	14.8	24.9	36.8	18.8	33.2	44

 © 1999 American Institute of Physics

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp

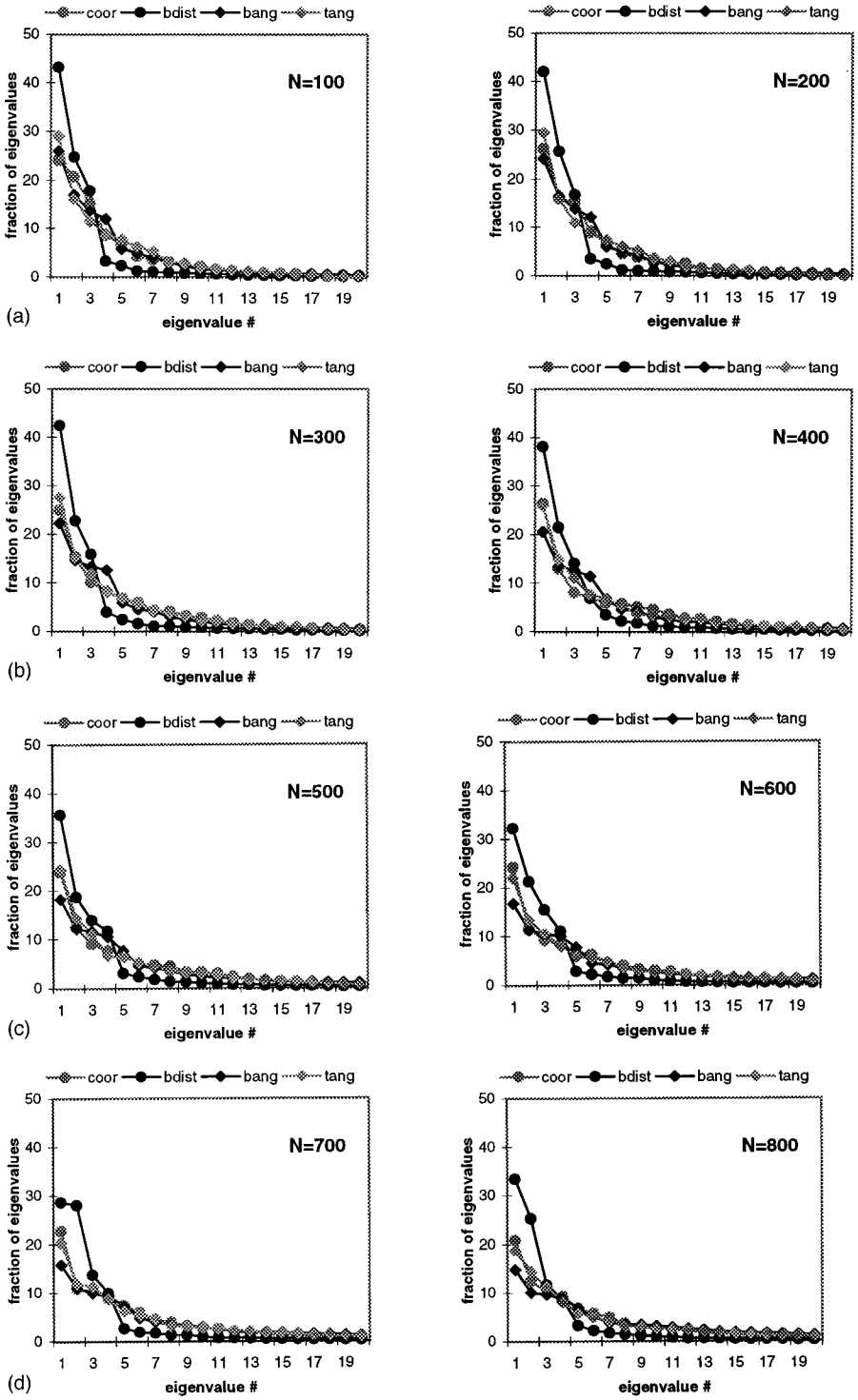

FIG. 1. Graphic representation of the fraction of the variance of each of the four choices of measures, for four sizes of samples: 100, 300, 500, and 800 elements, represented in (a), (b), (c), and (d), respectively. The upper energies of each selection are given in Table I.

monotonic sequences of linked stationary points'' (IMSLiSP), are the goals of this report.

The second, more elaborate approach takes as elements of the statistical database sequences of the stationary points. More specifically, the elements are sequences of alternating, geometrically linked minima and saddles, for which the energies of the minima increase monotonically from some lowest member from which no lower minimum can be reached without passing some higher-energy minimum. Such sequences are the natural constructions if one builds sets of linked ''triples,'' i.e., of two minima that share a common saddle between them. This method was first developed for analyzing why some atomic clusters readily take on amorphous structures from which it is difficult for them to escape, while others almost invariably go into crystalline or otherwise regular structures. The two kinds of clusters have been called ''glass-formers,'' for the former sort, and ''structure-seekers'' for the latter. The method was then used to test

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp

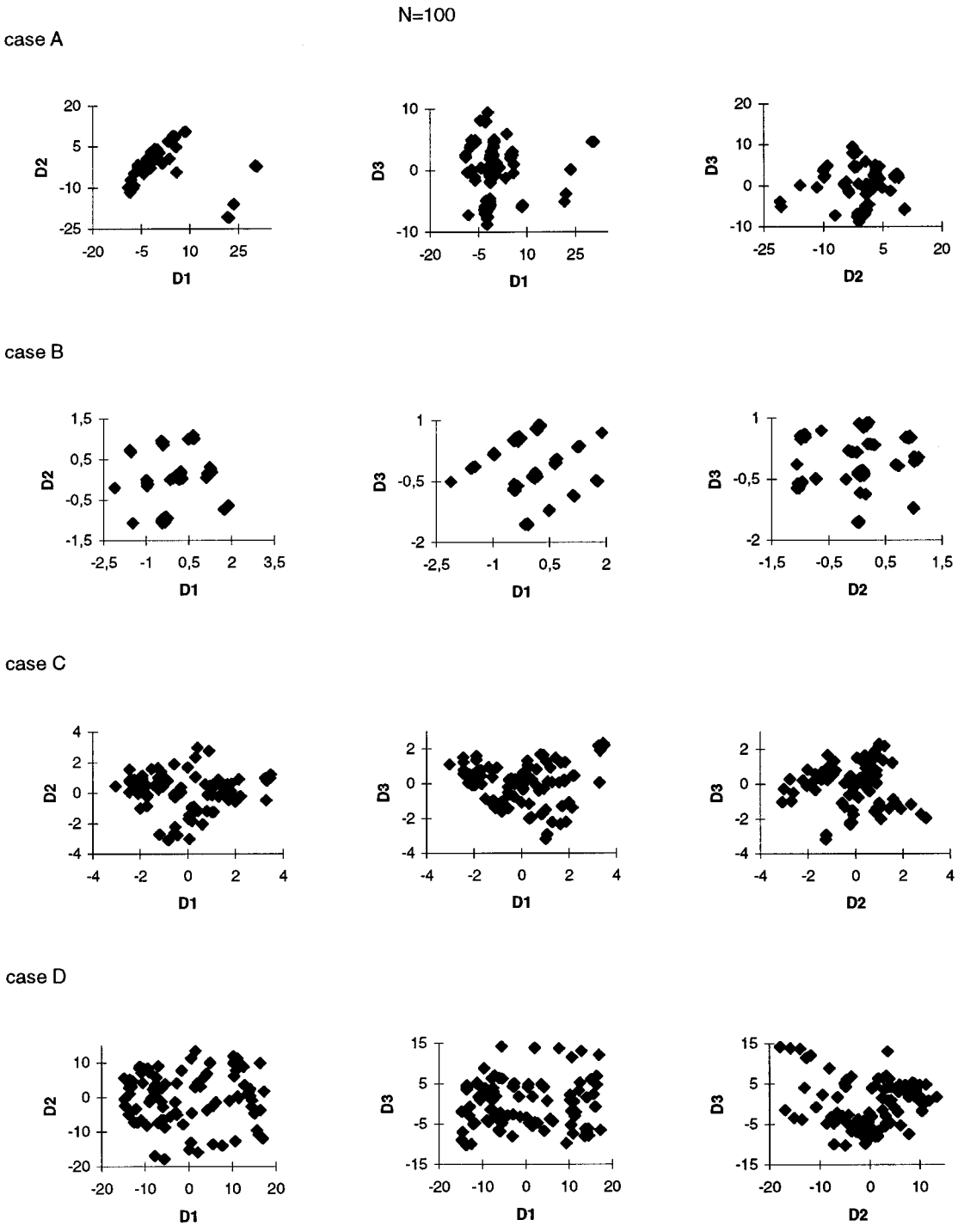

FIG. 2. Projection of the variance data onto planes of the two leading dimensions (principal coordinates with the largest variances), then of the first and third leading dimensions, and finally of the second and third leading dimensions, for all four choices of measures, A–D: Cartesian coordinates, interparticle distances, bending angles, and torsion angles, and for two sample sizes, 100 and 200. The distributions for larger samples are similar but denser.

whether the same criteria that distinguish structure-seeking clusters could apply to a protein model known to fold to a β-barrel.[11–14] The results seem, thus far, to be general: glass-formers have relatively short-range forces between their elements, sawtooth-like topographies with well-to-well energy drops small relative to heights of barriers separating the minima, and well-to-well motions involving only a few particles; structure-seekers have long-range forces between the elements (or effective long-range forces, such as imposed by correlations forced by a polymer chain), staircase-like topographies with sporadic large changes in energy from one minimum to the next, and highly collective motions in well-to-well passages. Clusters of particles interacting by Lennard-Jones forces, such as clusters of argon atoms, are in the class of glass-formers, with sawtooth potentials and few-body well-to-well motions. Alkali halide clusters and the 46-bead model that is the subject of this paper are structure-seekers, with staircase potentials and highly collective well-to-well

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp

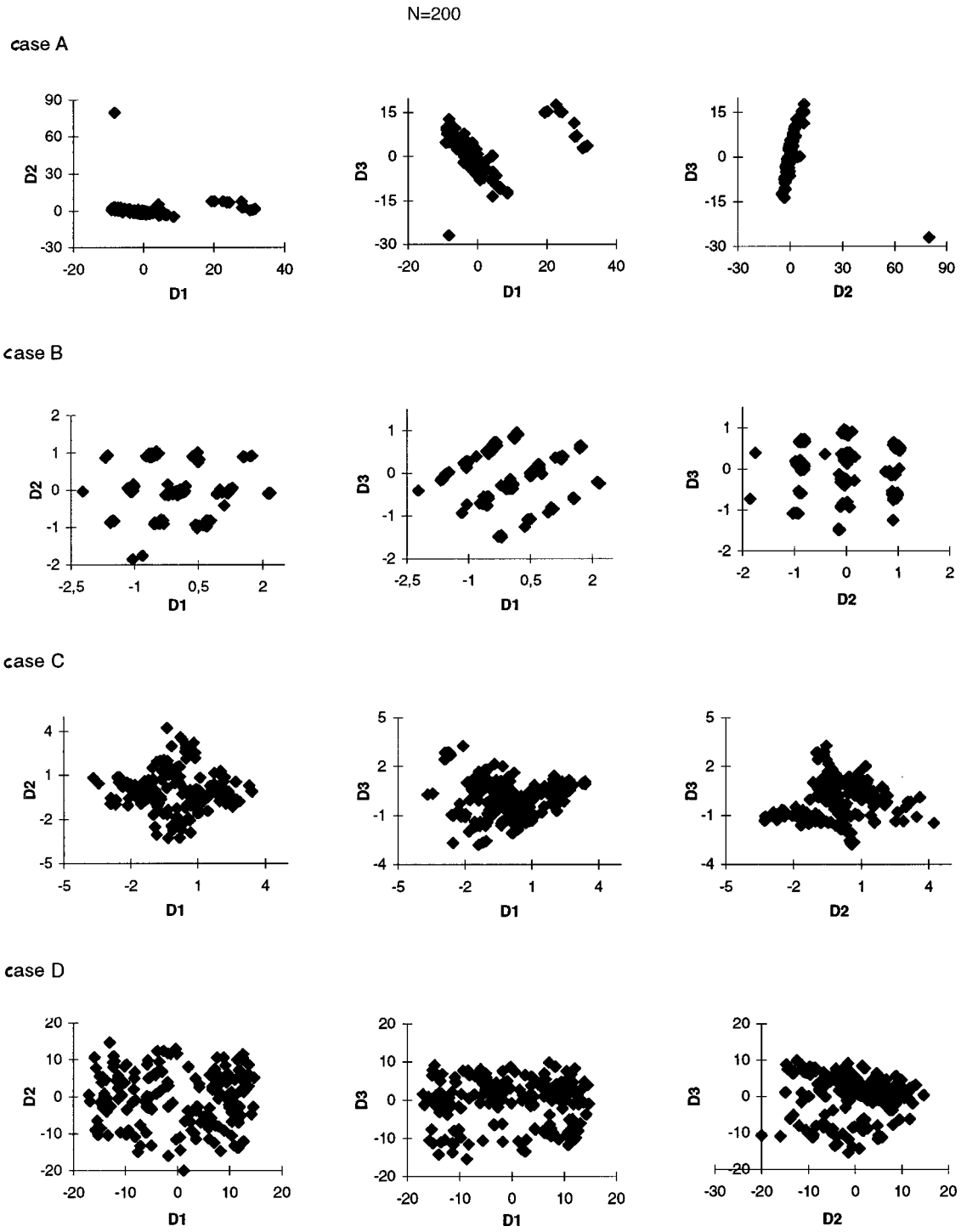

FIG. 2. (*Continued.*)

motions. The collective motions, the low energy barriers between minima and occasional large energy changes from one minimum to the next, are the common properties of the structure-seeker's potential surface. A planar graph of sequences is a highly simplified representation of the potential surface, which gives a vivid idea of the steepness of the potential surface.[9,10] Such a picture should be thought of as a two-dimensional projection of a many-dimensional surface.

There are a few examples of the use of PCoorA in the study of conformations of biomolecules. The method has been used to monitor the multidimensional distribution of reference conformations of Met-enkephalin by Abagyan and Argos.[15] They analyzed the influence of different parameters in a Monte Carlo minimization procedure using the stack-and-best-plane visualization.[15] The same method was used to study a protein conformational landscape.[16] Still more recently, Becker has presented quantitative visual representations of the potential surfaces of some proteins by principal coordinate (PCoor) analysis[1–5] as mentioned previously.

The goal of PCoor analysis is reducing the number of dimensions that describe a complex system from its complete set to a lower-dimensional representation that contains the most important data characterizing that system. In the situation discussed here, those data are the generalized

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp

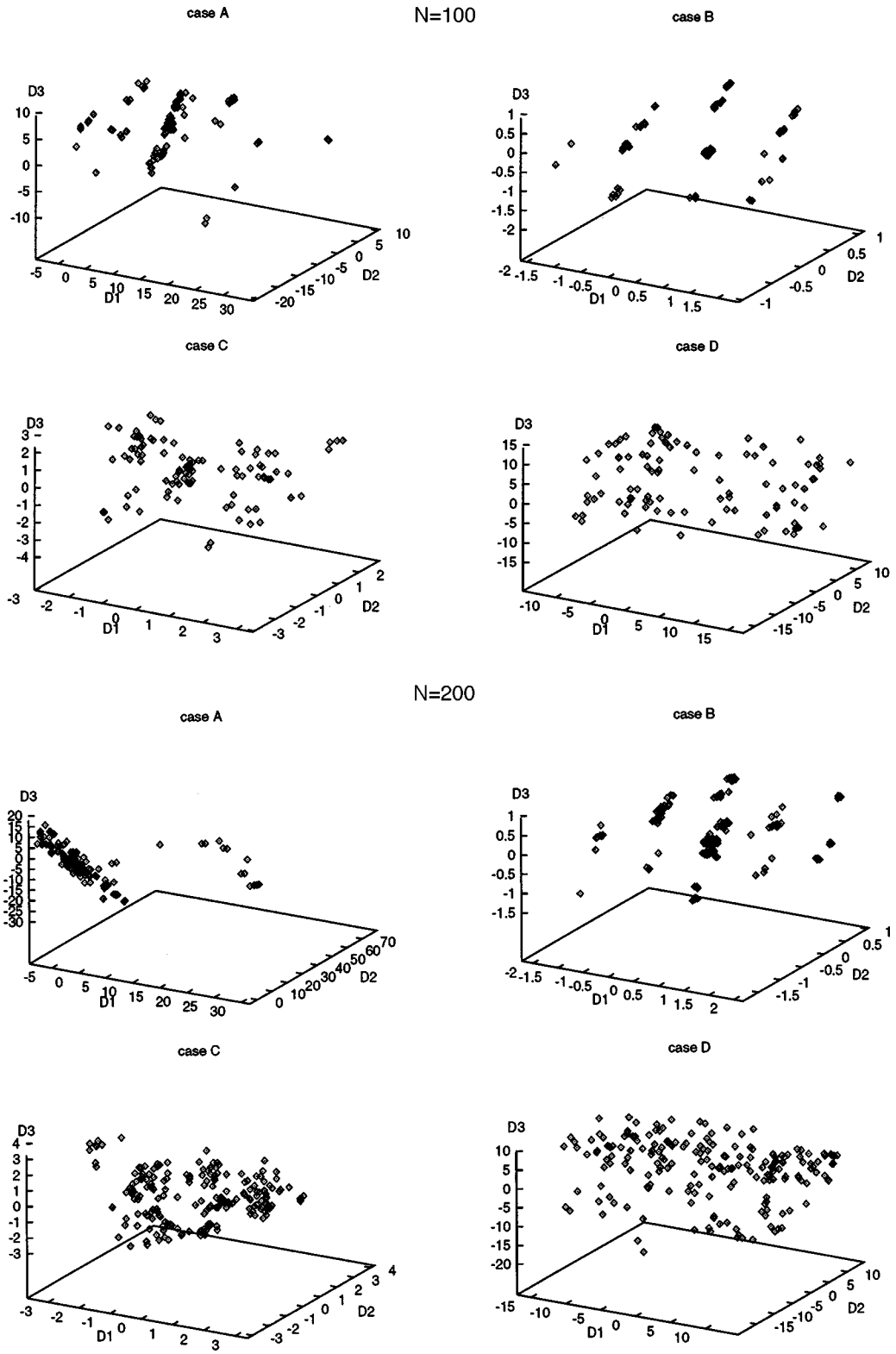

FIG. 3. Three-dimensional representations of the distributions, with the three most important components as the axes, for six sample sizes, 200–700 elements, and for three choices of criterion variables, A–D: Cartesian coordinates, interparticle distances, bending angles, and torsion angles.

coordinates that distinguish structures of clusters or polymers; more specifically, these should be coordinates that show the essential distinctions between structures that lie far from each other on their many-dimensional potential surface. The starting point of this method is the principal component (PComp) method, which is a statistical technique that linearly transforms an original set of variables into a substantially smaller set of uncorrelated variables that represents most of the information in the original data set.[17,18]

PCoor analysis was developed by Gower in 1966.[19,20] He showed the duality between PCoor and PComp methods.

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp

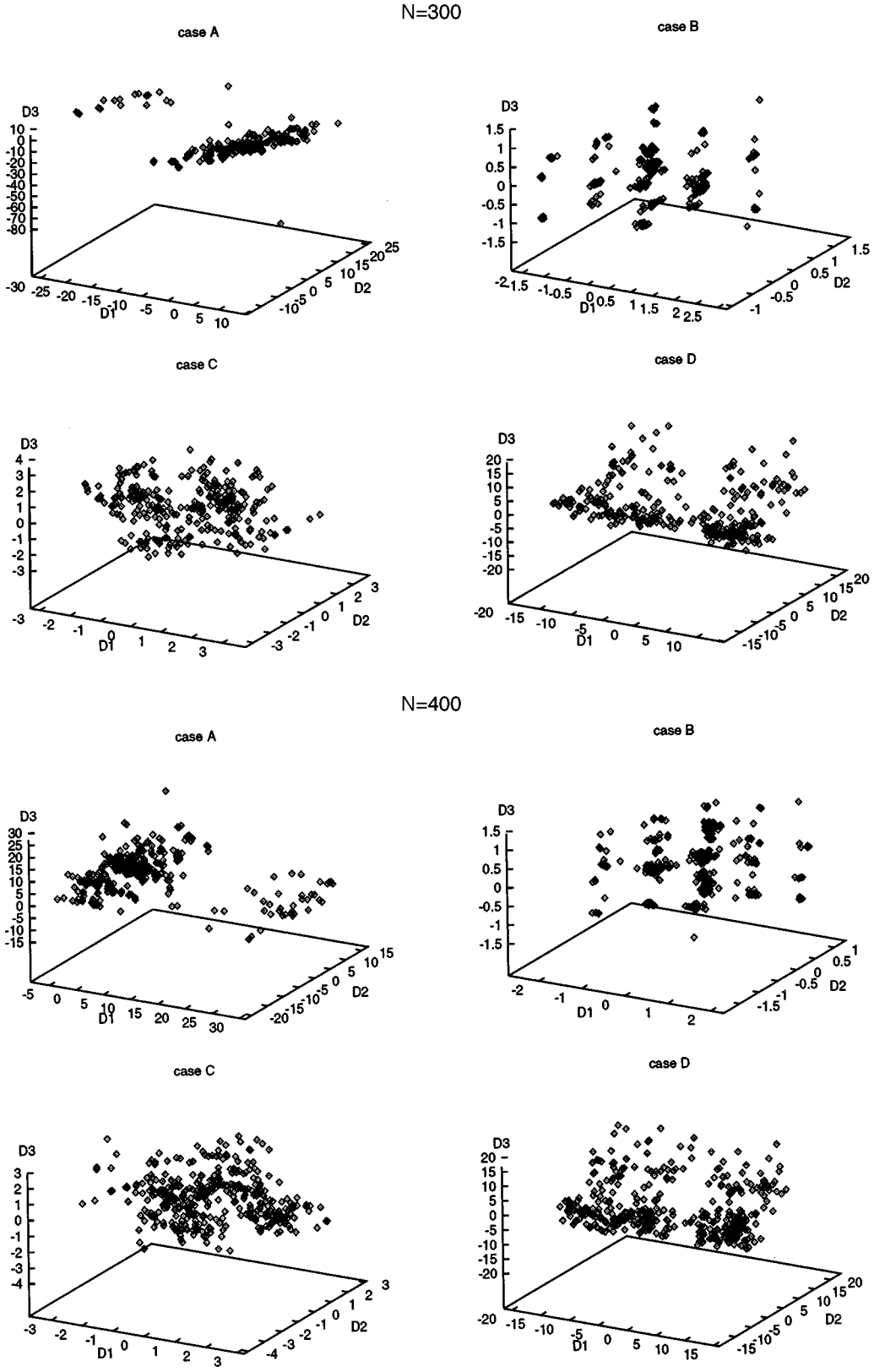

FIG. 3. (*Continued.*)

These methods start from a matrix of dissimilarities between sets of observations and produce a low-dimensional graphical representation of the data in such a way that the distances between points in the plot are close to the original, full-dimensional distances which are truly measures of dissimilarity.

In this work, we apply several versions of PCoorA to the 46-bead protein model that originated as a lattice model with Skolnick,[21–24] which was then developed by Honeycutt, Guo, and Thirumalai.[11–14] We compare several choices ovariables that can differentiate among locally stable strucf variables that can differentiate among locally stable structures of

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp

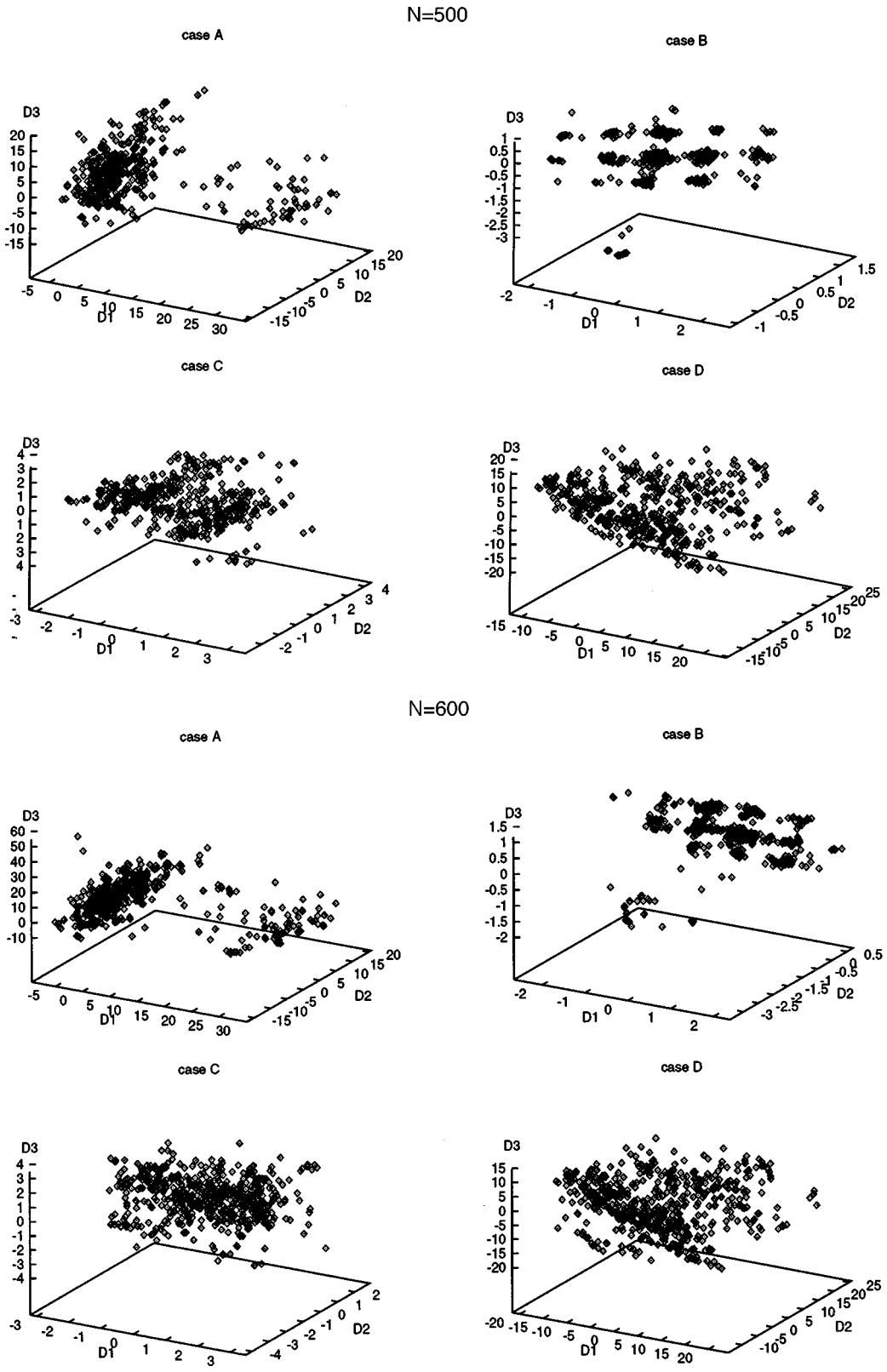

FIG. 3. (*Continued.*)

this system, to determine which contain the most information of interest. We also compare the use of databases of different sizes, and databases with different ''energy-lids,'' and determine how well PCoorA corresponds to the topographic information revealed by IMSLiSP analysis.

METHOD OF PCoor ANALYSIS

The starting point of PCoorA (sometimes called the ''Q-test'' or ''classical scaling method'') is principal component analysis (PCA), which uses a correlation matrix formed from

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp

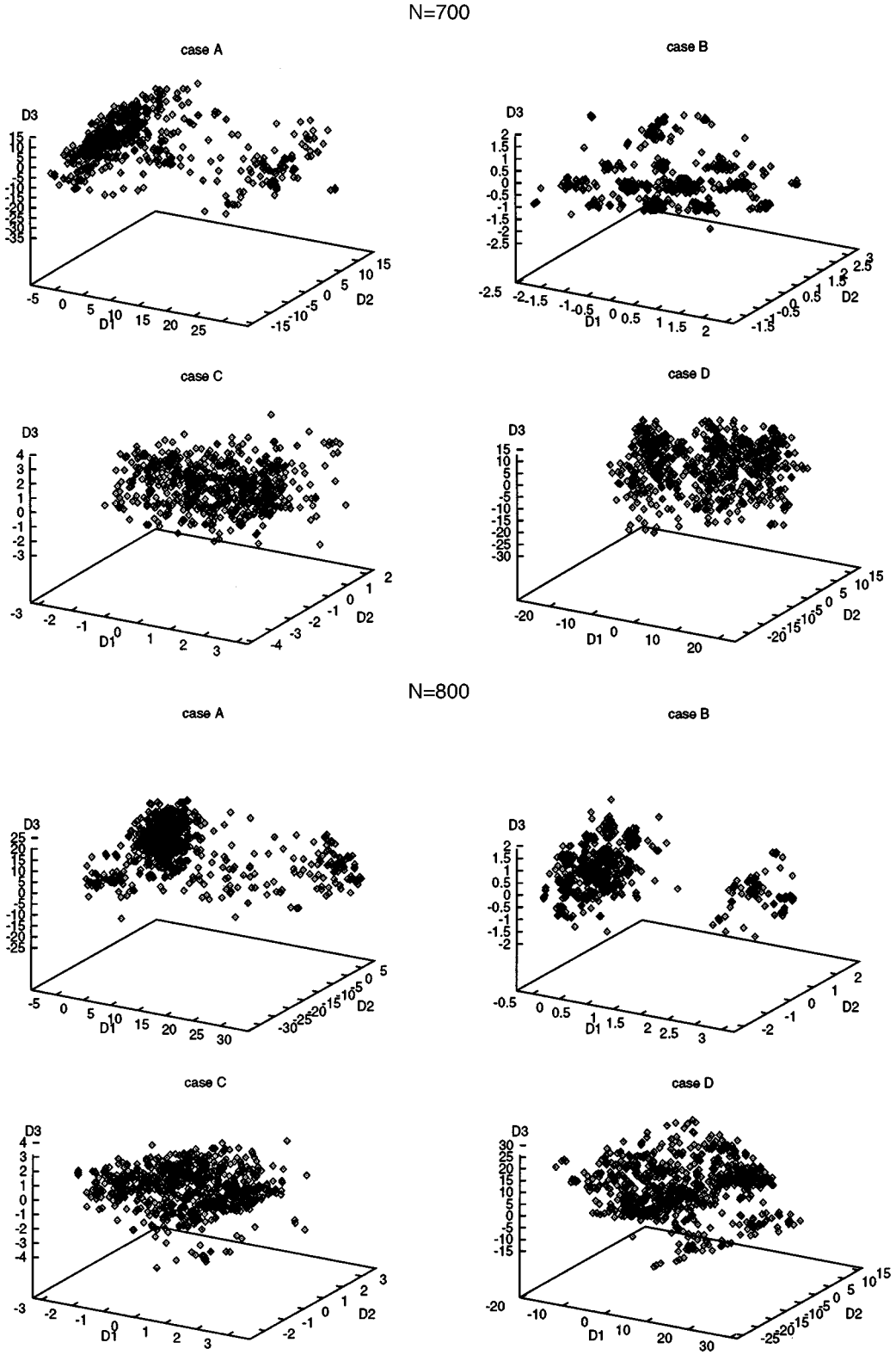

FIG. 3. (*Continued.*)

comparisons between the variables that distinguish the individuals of a set; in our case this is a set of coordinates that distinguish among structures. Assume we have a sample matrix, **X**, with dimension *nxp*, which has *p* variables and *n* observables or individuals. PCA uses the *pxp* matrix ($\mathbf{X}^T\mathbf{X}$, where $\mathbf{X}^T$ is the transpose matrix of **X**) as a starting (correlation) matrix and determines the most fluctuating variables among the full set of *p* members. On the other hand, PCoorA uses the *nxn* matrix, $\mathbf{X}\mathbf{X}^T$, as its starting matrix to find which groups of observables have the same character. Gower showed the duality between these two methods.[19] If the number of variates is larger than the number of observables, it is

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp

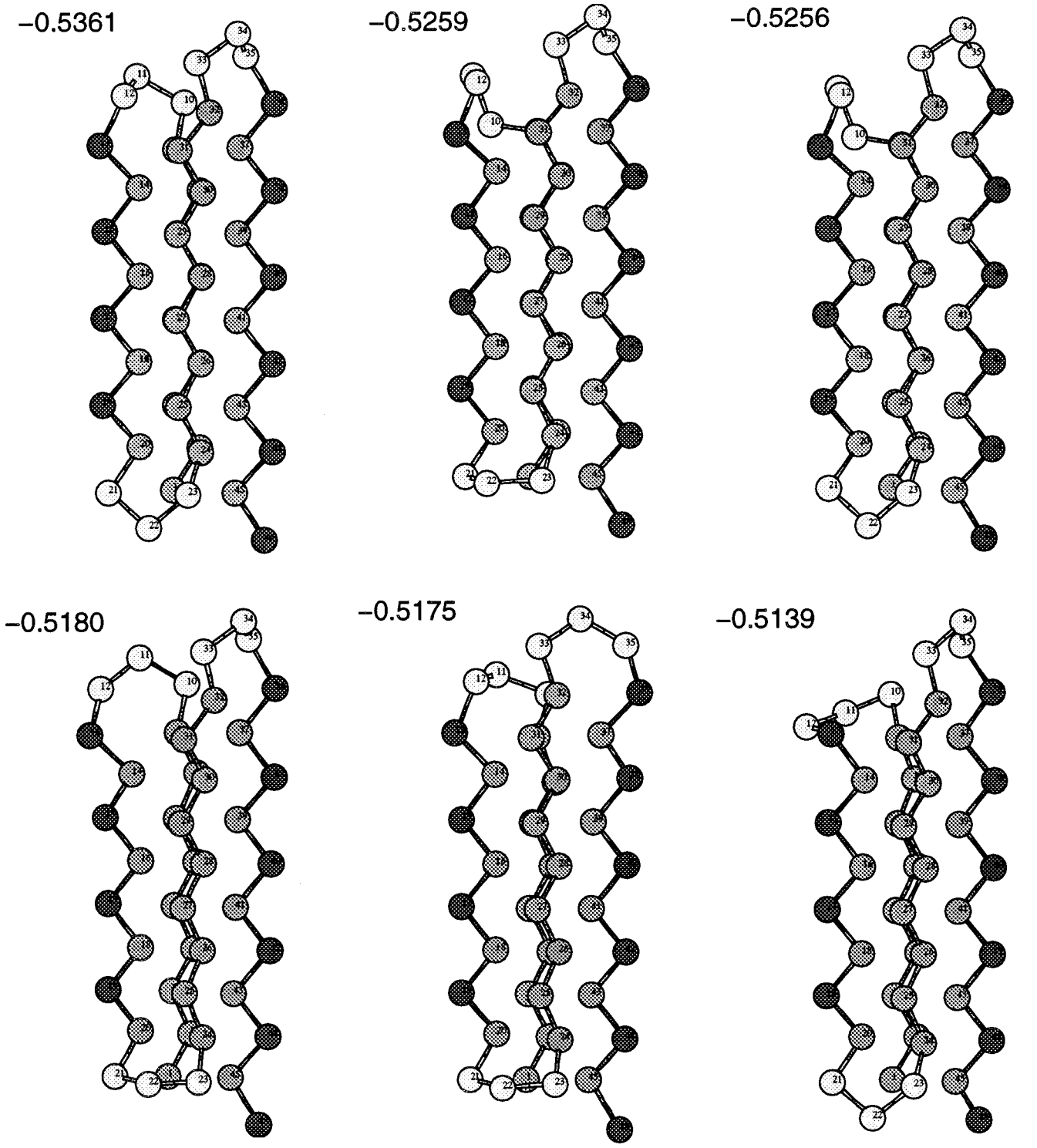

FIG. 4. Structures of some of the stable forms of the protein model that fall into common groups according to criterion B, the criterion of interparticle distance variability, as the choice of basis for the principal coordinates. Each of these structures lies in a single basin, of which there are probably at least eight.

better to use the PCoor technique since this makes finding eigenvalues easier.

The first step of the method is setting up the similarity or dissimilarity matrix, the distance matrix between pairs of observables. Originally, we have n different conformations—our observables—with each conformation defined by p variables, usually coordinates of each atom in the system and equal to three times the number of atoms, N.

The distance matrix defined as

$$d_{ij}=\sum_{r=1}^{p}(X_{ir}-X_{jr})^2,$$

where i and j are the conformation indices, r is the coordinate index, and d_{ij} is the measure of dissimilarity between the ith and jth conformations. In this work we have used different X

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp

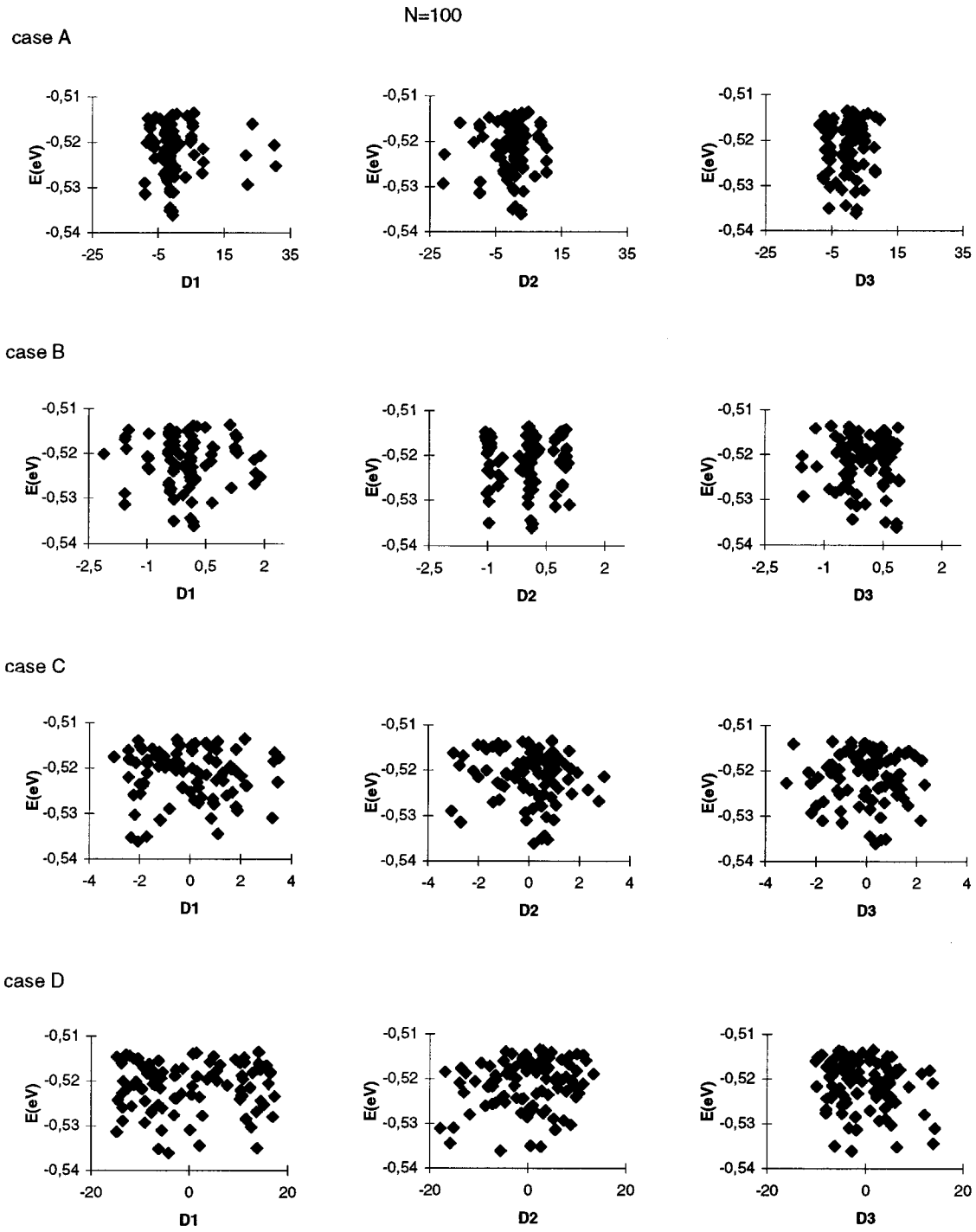

FIG. 5. Distributions of energies as functions of each of the three principal coordinates, for the smallest set of samples, 100, and for the four choices of criteria, A–D: the Cartesian coordinates, the interparticle distances, the bond angles, and the torsion angles.

matrices, based on different choices of variables. The four cases are

Case A, Cartesian coordinates $p=3N$;

Case B, pair distances between two atoms $p=N(N-1)/2$;

Case C, bond angles $p=N-2$;

Case D, dihedral angles $p=N-3$.

When the distance matrix is ready, the steps of the method ensue, in summary, thus

(1) Defining the matrix **A** whose elements are $-d_{ij}/2$;
(2) Obtaining the **B** matrix (the centralized form of **A**) as,
$b_{ij}=a_{ij}-a_{i.}-a_{j.}+a_{..}$,
where $a_{i.}$ is the average of the elements of the ith row of the **A** matrix, $a_{j.}$ is the average of the elements of the jth

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp

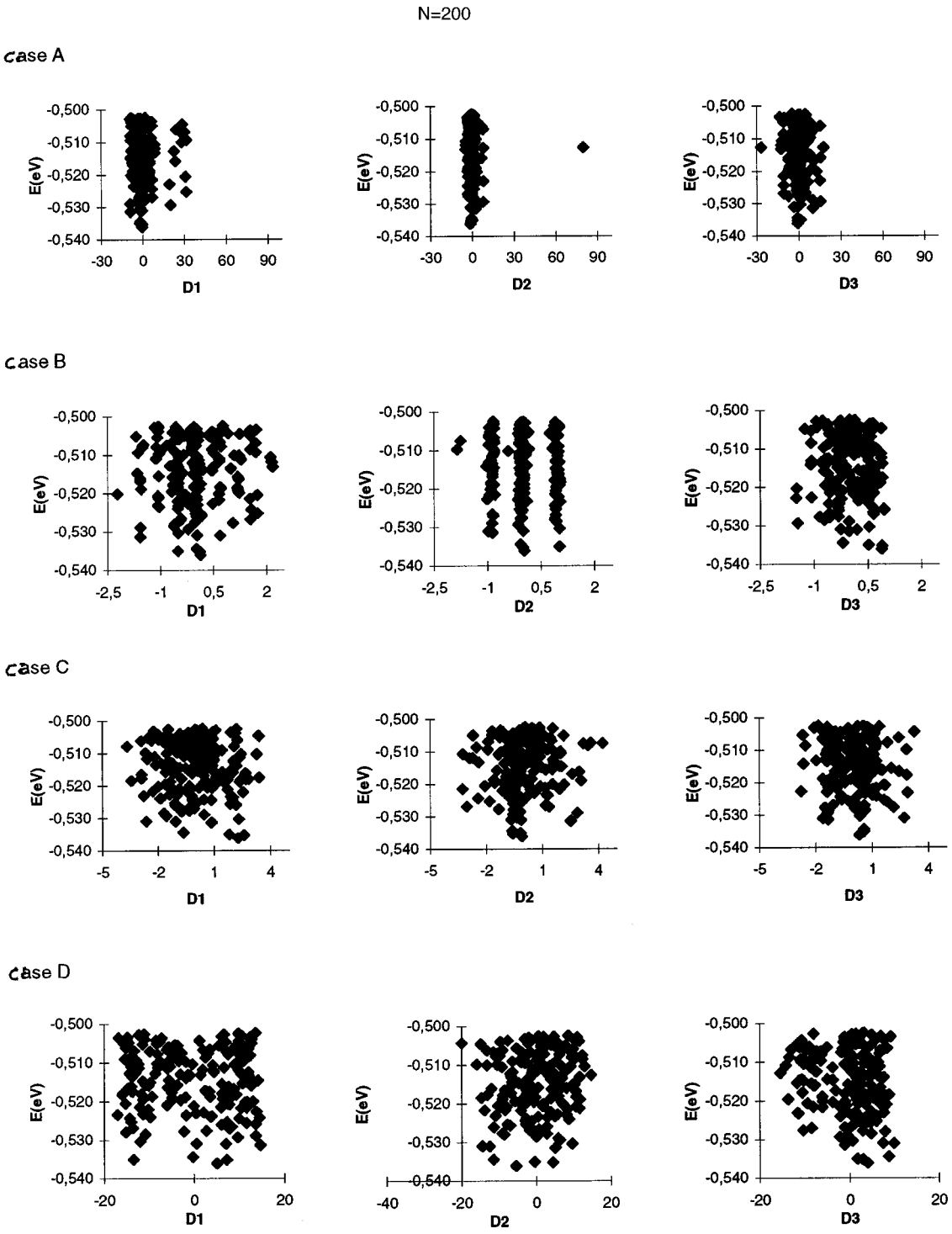

FIG. 5. (*Continued.*)

column of the **A** matrix, and $a_{..}$ is the average of all elements of the **A** matrix;

(3) Diagonalizing the **B** matrix and finding its eigenvalues. Then one scales the eigenvectors to make the sum of squares of elements of each eigenvector equal to the corresponding eigenvalue. The eigenvalues of **B** give the new dimensions and its eigenvectors are the coordinates corresponding to those eigenvalues. The largest eigenvalue represents the maximum variance of the original set of variables and carries the most information about the system. This eigenvector can be named as the 1st dimension, the eigenvector corresponding to the second largest is the 2nd dimension, and so on.

In our study, the set of locally stable configurations has been obtained by molecular dynamics simulation with constant temperature. Samples have been obtained from simulations at various temperatures, and minima corresponding to different geometric structures combined to form the statistical dataset. The final samples were then sorted in order of

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp

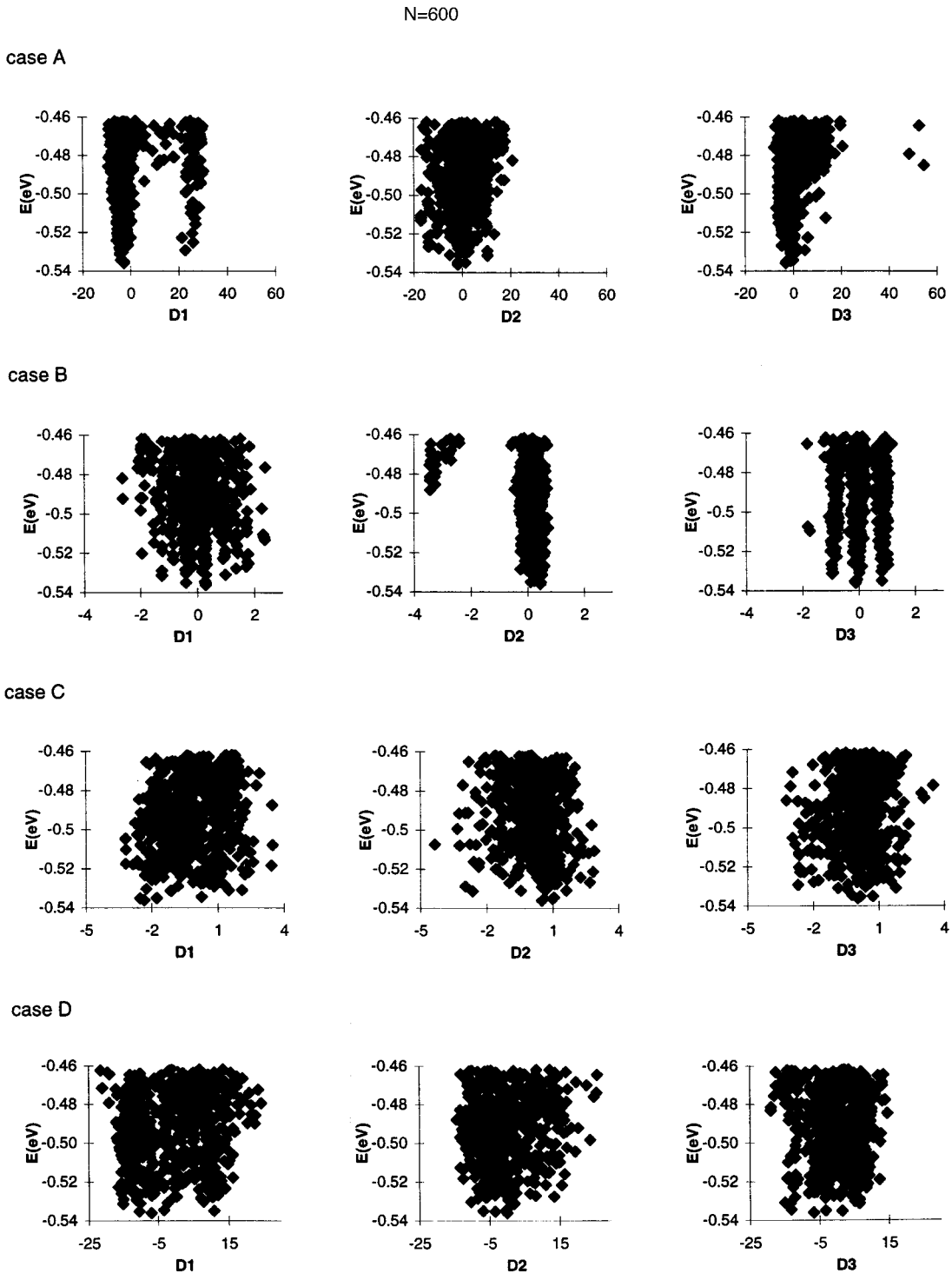

FIG. 5. (*Continued.*)

ascending energy. We then applied PCoor analysis to subsets of these data, with each subset defined by the upper limit of its energy. Hence these subsets necessarily differed in the number of structures included. This kind of sorting is akin to the ''lid method'' of Sibani and Schön.[25,26]

RESULTS AND DISCUSSIONS

In this section, we examine what the PCoor analysis reveals with different starting matrices, **X**, based on the four cases described previously and on the sample size and upper limit of energy.

The fractions of each eigenvalue coming from the most important (most varying) dimensions, as a function of the number of dimensions included, i.e., the cumulative contributions to the eigenvalues, are given for cases A through D and for various sample sizes in Table I and Fig. 1. Since each eigenvalue is a measure of fluctuations in the original data, the largest eigenvalue carriers most information, and projection is the best for that eigenvalue. Table I summarizes the accumulation of the three largest eigenvalue fractions for all four cases, corresponding in turn to the starting matrix as Cartesian coordinates, pair distances, bond angles, and tor-

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp

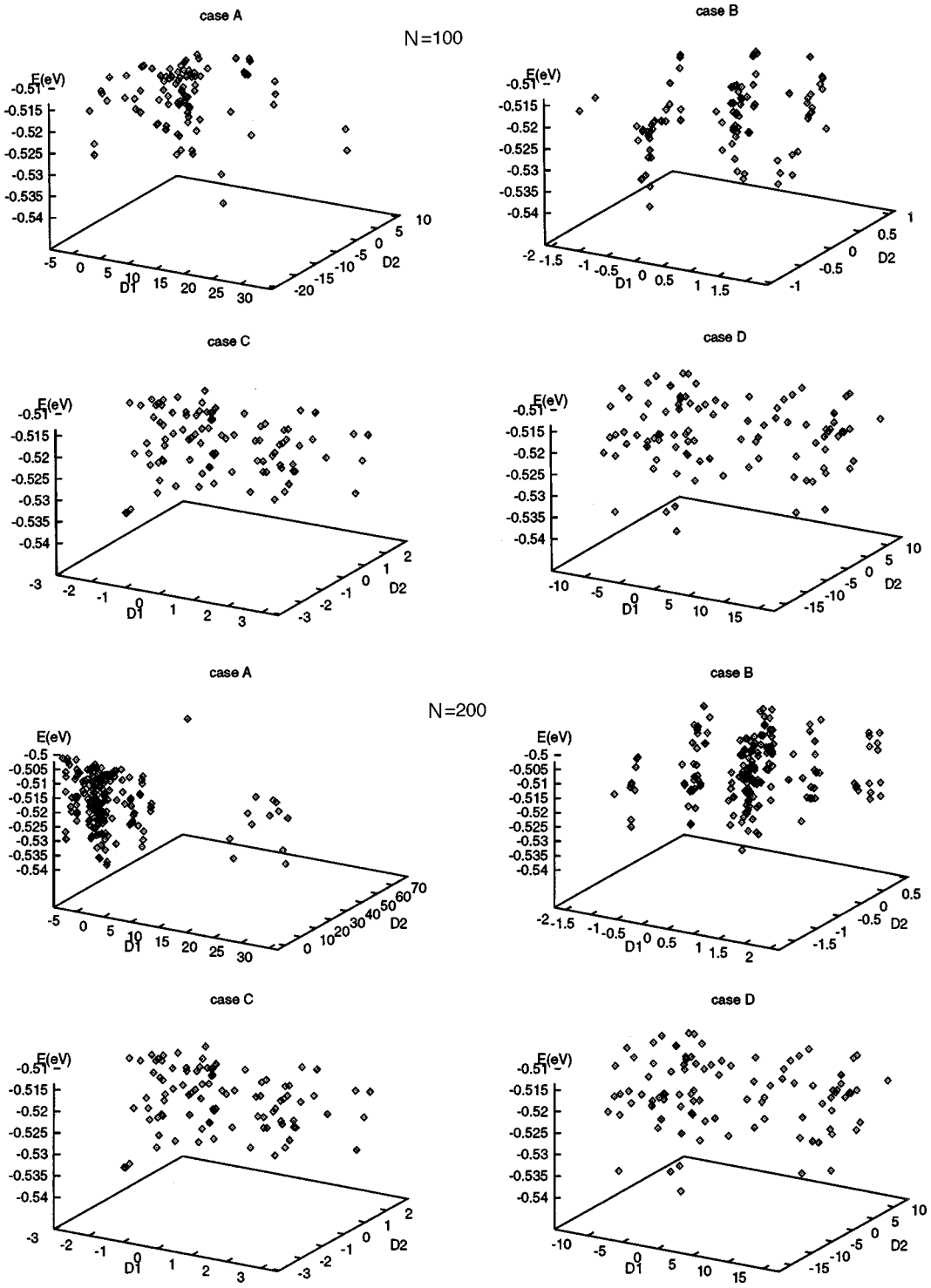

FIG. 6. Three-dimensional plots of the energy as a function of the two most important principal coordinates, for sample sizes 100 and 200, and for the four choices of criterion, A–D, Cartesian coordinates, interparticle distances, bending angles, and torsion angles.

sional angles at various system sizes. Case B always shows the largest accumulation of variance in this system. The first two and first three dimensions contain about 60%–65% and 70%–85%, respectively, of the information on variance that is contained in the data. In cases A and D, the leading eigenvectors contain about 33%–50% and 40%–60% of the variability information in the leading two and leading three variables. Case C has the lowest values at all sample sizes; that is, the bond angle variations contain the least concentrated, most diffusively spread information among the four choices of coordinates.

Projection of the data on the first three dimensions is shown in Figs. 2 for Cases A–D for two sample sizes. These smaller samples were taken with energy ''lids'' so the structures of the samples were restricted to the lower-energy reaches of the surface. With sample sizes of 100–400 structures, corresponding to an energy range between the minimum at -0.536 and -0.48 eV, there are too many clus-

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp

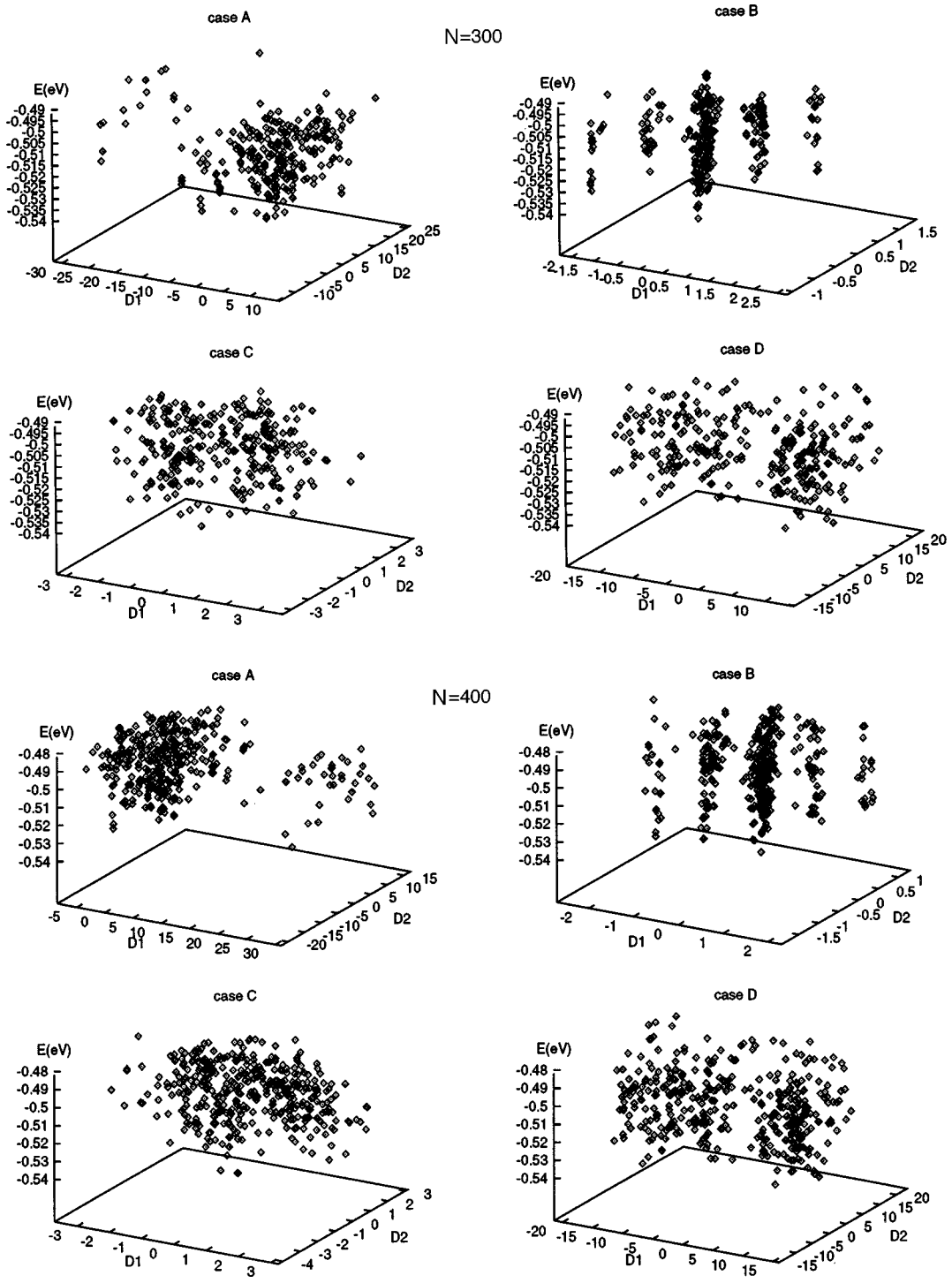

FIG. 6. (*Continued.*)

ters of configurations to be informative in the case B projection. On the other hand, case A shows two or three clusters of structures and case D shows no correlation or clustering among the structures. This can be seen better in the three-dimensional representations of these plots, in Fig. 3. In this figure, case B shows at least five different structural groups. We have examined the structural similarity of these groups; some examples are given in Fig. 4. The same observations are seen in a different way in Fig. 5, a representation of energy versus dimensions, and Fig. 6, a three-dimensional plot of energy versus the first two dimensions.

Using distance space reduces the number of variables, p, by a factor of $N/6$ from that of a representation in coordinate space and has the advantage of shorter computational time in both PCoorA and PCA. Very recently, Abseher and Nilges[27] used the covariance matrix in distance space to carry out the analysis of their data by PCA. They compared these results from PCA in coordinate space with reduced eigenvectors and

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp

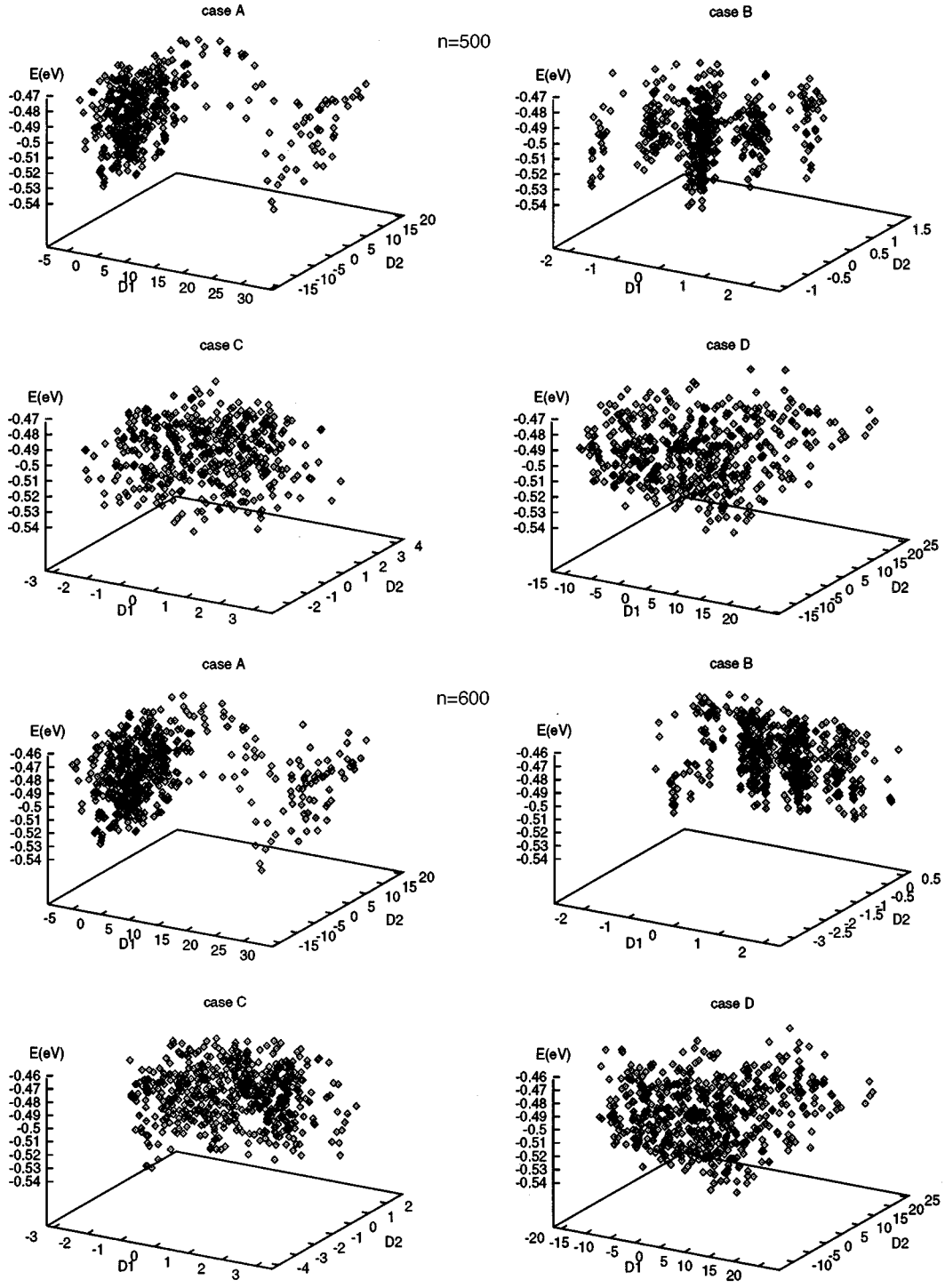

FIG. 6. (*Continued.*)

pointed out the similarity of the results from the two approaches.

If the sample size is larger than 400 configurations, structural groups merge at higher energy levels. With about 800 configurations, we find just two groups due to inclusion of very wild, high-energy structures.

Becker and Karplus pointed out the correlation between sharing conformational similarity and being members of same energy basin for the IAN tetrapeptide.[1] Becker confirmed the correlation for the same system by using principal component analysis, although it is not certain that correlation with that method is complete.[2] In our previous work we identified four different basins of attraction for the 46-bead model by monotonic sequence analysis; there are probably at least eight on the full surface.[10,12] These four basins appear as four different groups of structures in the PCoorA. In addition to these four, there are other groups. We have checked the lowest energy configurations in each group, to determine

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp

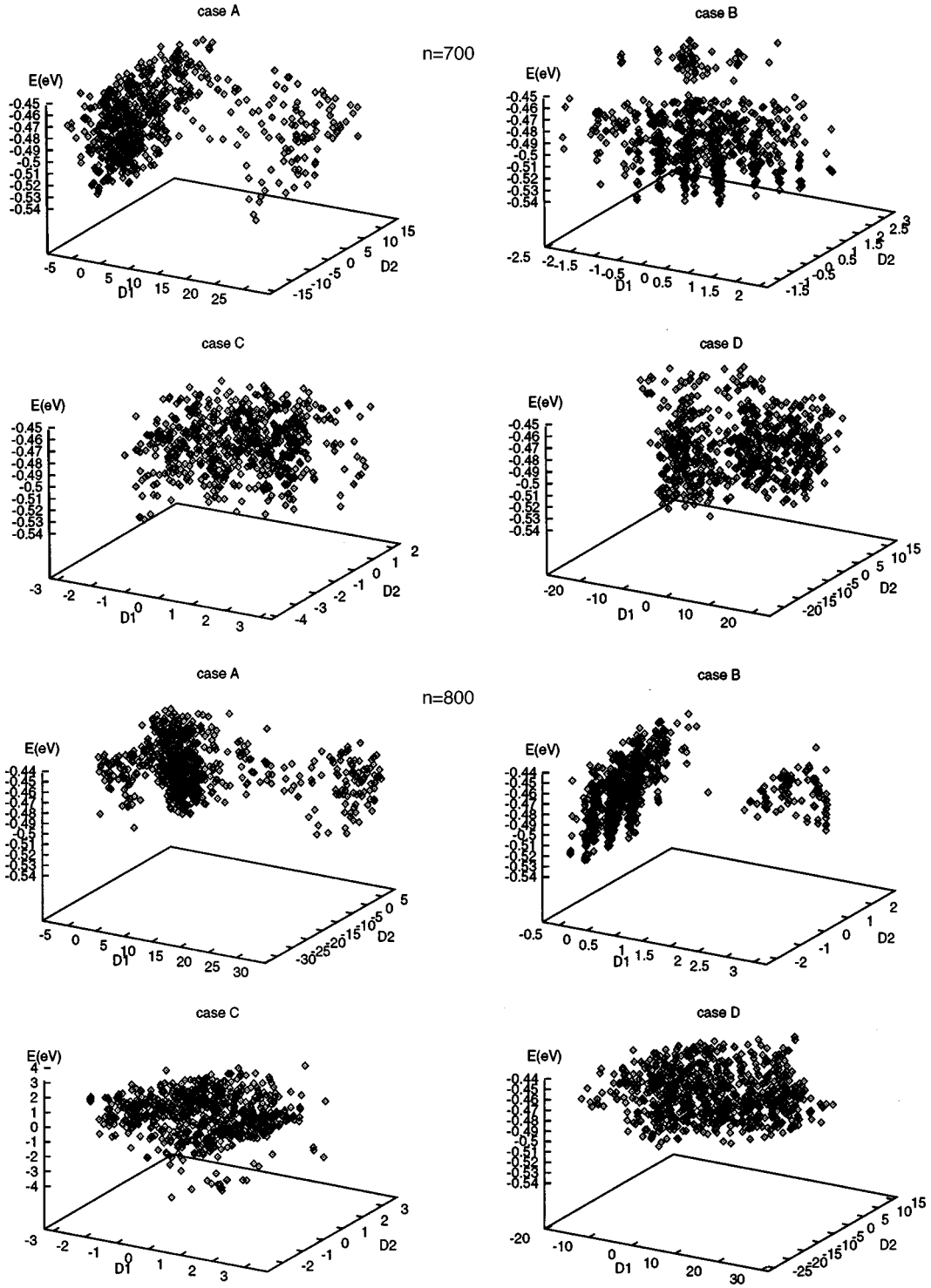

FIG. 6. (*Continued.*)

whether or not they belonged to any of the four fully identified basin sequences. The minimum points with energies −0.5285 and −0.5268 eV fall into different groups according to PCoorA, but the IMSLiSP analysis shows that they all lie in one of the four basins but with relatively large energy barriers (as intrabasin barriers go) separating them. See Fig. 7, showing sequences from IMSLiSP.

We conclude that PCoor analysis based on the changes of interparticle distances is an efficient and useful tool for categorizing minima on complex potential surfaces, especially for differentiating structures separated by moderately high energy barriers. Comparison of patterns based on different upper bounds of energy helps to reveal the topography of the multidimensional surface, as had been proposed by Becker and Karplus.[1] The importance of PCoor analysis will probably grow with the complexity of the system, because larger and larger systems will place greater and greater demands on the databases for IMSLiSP. Nevertheless,

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp

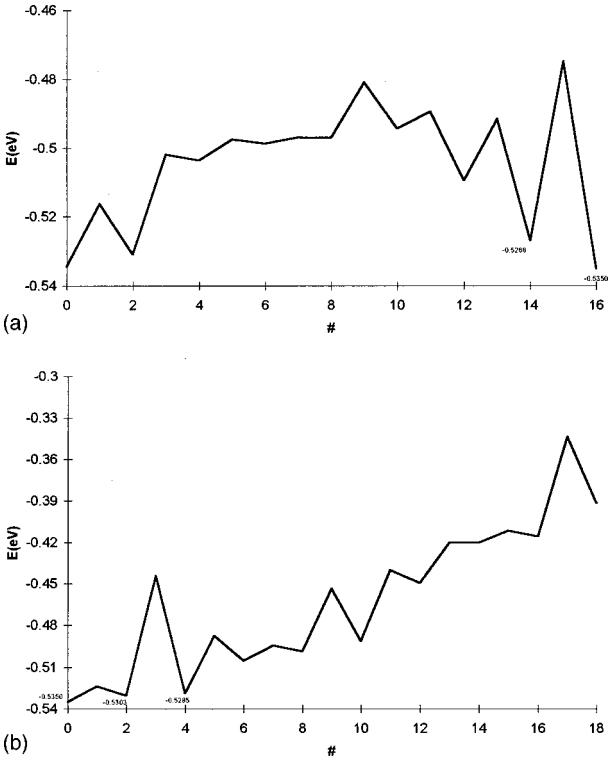

FIG. 7. Typical sequences of linked minima and saddles, monotonic in the energies of the minima, on the potential surface of the 46-bead model; the upper figure shows two basins with a divide at the saddle between minima 8 and 10; the lower is a longer monotonic sequence on the potential surface of the 46-bead model.

IMSLiSP will probably be necessary to establish the pattern of saddle points and to implement kinetic studies by master equation methods. We can hope that PCoorA will make it possible to carry out IMSLiSP in more systematic, efficient ways than have been possible heretofore.

ACKNOWLEDGMENTS

We would like to thank Dr. Oren Becker for introducing us to principal coordinate analysis and for very helpful discussions. This research was supported by a Grant from the National Science Foundation.

[1] O. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).
[2] O. M. Becker, J. Mol. Struct.: THEOCHEM **398–399**, 507 (1997).
[3] O. M. Becker, Proteins: Struct., Funct., Genet. **27**, 213 (1997).
[4] O. M. Becker, J. Phys. Chem. (submitted).
[5] O. M. Becker, J. Comput. Chem. **19**, 255 (1998).
[6] R. E. Kunz and R. S. Berry, J. Chem. Phys. **103**, 1904 (1995).
[7] R. E. Kunz, R. S. Berry, and T. Astakhova, Surf. Rev. Lett. **3**, 307 (1996).
[8] R. E. Kunz, P. Blaudeck, K. H. Hoffmann, and R. S. Berry, J. Chem. Phys. **108**, 2576 (1998).
[9] K. D. Ball, R. S. Berry, A. Proykova, R. E. Kunz, and D. J. Wales, Science **271**, 963 (1996).
[10] R. S. Berry, N. Elmaci, J. P. Rose, and B. Vekhter, Proc. Natl. Acad. Sci. USA **94**, 9520 (1997).
[11] Z. Guo, D. Thirumalai, and J. D. Honeycutt, J. Chem. Phys. **97**, 525 (1992).
[12] J. D. Honeycutt and D. Thirumalai, Biopolymers **32**, 695 (1992).
[13] Z. Guo and D. Thirumalai, in *Protein Folds*, edited by H. Bohr and S. Brunak (CRC, Cleveland, 1995), p. 233.
[14] Z. Guo, C. L. Brooks III, and E. M. Boczko, Proc. Natl. Acad. Sci. USA **94**, 10161 (1997).
[15] R. Abagyan and P. Argos, J. Mol. Biol. **225**, 519 (1992).
[16] J. M. Troyer and F. E. Cohen, Proteins **23**, 97 (1995).
[17] G. H. Dunteman, *Principal Components Analysis* (Sage, London, 1989).
[18] I. T. Jolliffe, *Principal Component Analysis* (Springer-Verlag, New York, 1986).
[19] J. C. Gower, Biometrika **53**, 325 (1966).
[20] J. C. Gower, Biometrika **55**, 582 (1968).
[21] J. Skolnick, A. Kolinski, and R. Yaris, Biopolymers **28**, 1059 (1989).
[22] J. Skolnick and A. Kolinski, Science **250**, 1121 (1990).
[23] A. Sikorski and J. Skolnick, Biopolymers **28**, 1097 (1990).
[24] A. Sikorski and J. Skolnick, J. Mol. Biol. **212**, 819 (1990).
[25] P. Sibani, J. C. Schön, P. Salamon, and J. O. Andersson, Europhys. Lett. **22**, 479 (1993).
[26] J. C. Schön, Ber. Bunsenges. Phys. Chem. **100**, 1388 (1996).
[27] R. Abseher and M. Nilges, J. Mol. Biol. **279**, 911 (1998).

Downloaded 23 Sep 2003 to 128.135.233.75. Redistribution subject to AIP license or copyright, see http://ojps.aip.org/jcpo/jcpcr.jsp