Journal of Computer-Aided Molecular Design, 3 (1989) 85–94
ESCOM

J-CAMD 055

Computer simulation of the conformational behavior of cholecystokinin fragments: Conformational families of sulfated CCK8

M. Kreissler[a,*], M. Pesquer[a], B. Maigret[b], M.C. Fournié-Zaluski[c] and B.P. Roques[c]

[a]*Laboratoire de Physicochimie Théorique, U.A. CNRS 503, 351, Cours de la Libération, 33405 Talence Cedex, France*
[b]*Groupe de RMN et Modélisation Moléculaire, Laboratoire de Chimie Organique Physique, U.A. CNRS 422, Institut Lebel-Université L. Pasteur, 4, rue Blaise Pascal, 67000 Strasbourg, France*
[c]*Département de Chimie Organique, U 266 INSERM and U.A. CNRS 498, UER des Sciences Pharmaceutiques et Biologiques, 4, Avenue de l'Observatoire, 75006 Paris, France*

Received 15 December 1987
Accepted 18 August 1988

Key words: Cholecystokinin; Monte-Carlo method; Cluster analysis; Conformational analysis

SUMMARY

Display methods, such as principal component analysis, and clustering methods were applied to a sample of cholecystokinin, (sulfated CCK8) conformations obtained from a Monte Carlo simulation. It is shown that six families of conformations can entirely describe the sample. Each family represents a typical conformer. These theoretical models are in agreement with recent experimental results which stress the predominance of folded conformers in aqueous medium.

INTRODUCTION

A great deal of experimental studies has been and is still devoted to cholecystokinin, a gastrointestinal peptide, in order to make clear his conformational properties [1]. This 33-amino acid peptide produces pancreatic enzyme secretions and gallblader contactions during digestion. Among the shorter fragments of cholecystokinin, the C-terminal octapeptide Asp^{26}-$(SO_3)Tyr^{27}$-Met^{28}-Gly^{29}-Trp^{30}-Met^{31}-Asp^{32}-Phe^{33}-NH_2, designated CCK^{26-33} or CCK8, is of particular interest since it shows the same potency as the complete molecule [2]. Besides, high concentrations of CCK8 have recently been found in brain [3,4] and there are experimental facts which suggest that this peptide could act as a neuromodulator in the central nervous system [5,6]. In that context a

*To whom correspondence should be addressed.

0920-654X/89/$3.50

theoretical conformational analysis of the CCK8 fragment allows one to complete the experimental results obtained by proton NMR and fluorescence energy transfer [7,8].

As we want to focus on the comparison between calculated conformations and those observed in solution and owing to the great number of conformations to be explored for a linear peptide we have used a Monte Carlo sampling method devised by Metropolis et al. [9]. This approach has given satisfactory results for the conformational analysis of other peptides [10–13]. Applied to the simulation of the conformational behaviour of CCK8 in aqueous solution at physiological pH this procedure allows us to get a statistical sample of conformations which can be estimated representative of the overall population of conformations. A more detailed analysis of the information contained in the whole Monte Carlo sample implies the use of clustering analysis. We have followed the same methodology as proposed by Marchionini et al. for the investigation of angiotensin II [14].

METHOD (GENERATION AND DESCRIPTION OF THE CONFORMATIONS)

The Monte Carlo procedure is based on statistical variations of the molecular conformation. Random numbers are used to generate local variations of the conformation (change in a torsion angle) which lead to new states. If the energy of a new state is lower than that of a previous one, the new state is accepted, i.e., the next variation will start from this new state. On the other hand, if the energy is higher, the Boltzmann factor $P = \exp(-\Delta U/RT)$ is calculated and compared with a random number chosen between 0 and 1. If the random number is smaller than P, the new state is after all accepted; otherwise the previous state is retained. P is thus the transition probability for a state of higher energy and depends only on the directly preceding state. This type of random walk through the conformational space of the molecule is called a Markov chain [15]. In this way the molecule is allowed to fluctuate according to a Boltzmann distribution for temperature T. Using the Boltzmann factor as a transition probability is the essential feature of the Metropolis sampling method [9].

The conformational energies are computed by a molecular mechanics approach with the forcefield of Scheraga [16,17] and using rigid bond lengths and angles.

The sample of molecular conformations obtained by the Monte Carlo algorithm comprises $n = 4635$ different conformations amongst a total of 10^6 conformations generated during the calculations. Each one of the n conformations may be characterized either by 36 dihedral angles (φ, ψ, $\{\chi\}$) or by a set of all variable interatomic distances. In their work concerning angiotensin II, the authors [14] have shown that a subset of 23, in our case $m = 25$ distances, carry the major part of the structural information: namely the 21 $C_i\alpha - C_i\alpha_{+2}$ distances, the N-C terminal end-to-end distance and the three $(SO_3)Tyr^{27}$-Trp^{30}, $(SO_3)Tyr^{27}$-Phe^{33} and Trp^{30}-Phe^{33} aromatic side chain centroïd distances. These 25 distances or descriptors will be used for the multivariate data analysis methods, such as principal component analysis and clustering.

DATA MATRIX

The $m = 25$ measurements (interatomic distances) for the $n = 4635$ objects (conformations) may be assembled in an $n \times m$ dimensional data matrix X with element X_{ij} (*i*th conformation and *j*th distance).

Each conformation i of the CCK8 molecule is characterized by a pattern vector $X_i = (X_{i1},...,X_{ij},...,X_{im})$ and the aim of the classification methods is to cluster the $n = 4635$ conformations and to obtain a small number of significant clusters. Inversely, one can also classify the interatomic distances j which are each characterized by a vector $X_j = (X_{1j},...,X_{ij},...,X_{nj})$. It is thus possible to consider the data space in a dual way, the objects being defined by the measurement values and in turn the measurements being characterized by the values they take for each object. We have used the two approaches for the study of our data set.

The multivariate analysis methods available are numerous and it is recommended to use several of them simultaneously to solve a given problem. By that way one improves the confidence one can have in the results of a data analysis. A combined use of principal component analysis and clustering methods is often fruitful.

DISPLAY OF THE OBJECTS(CONFORMATIONS)

With the help of principal component analysis (PCA), it is possible to get an idea about the aspect of the data set points without too much distortion, although it must be remembered that this is an 'approximate view' into 25D space. Fig. 1 represents the projection of 4635 conformation points onto the first three principal axes which preserve about 73% of total information. The data set cloud appears to be extended along the direction of the first principal axis. One can also distinguish the existence of some substructures in the data set cloud. They will be determined in a more precise and quantitative way by the use of clustering methods.

DETERMINATION OF CONFORMATIONAL CLASSES BY CLUSTER ANALYSIS

The principle of all clustering techniques consists in the possibility of reducing the number of data points by grouping together neighbour points. Methods of cluster analysis and their application to chemical problems are summarized by Duran and Odell [18] and by Massart and Kaufmann [19]. The interested reader can find a more general introduction to these methods in the book by Duda and Hart [20].

The goal of our study is to gain an understanding of the 4635-data structure and to determine the most likely number of clusters in the conformational space. It is not possible to reach the best possible partition of the analysed data by the use of only one clustering method. To avoid these difficulties we have used two clustering procedures which are included in the software package of Statistical Analysis System (SAS) [21]. These two methods are respectively a hierarchical SAS procedure CLUSTER and a non-hierarchical one FASTCLUS. The first procedure uses Ward's method [22] and operates in an ascending way, starting with all data points ($n = 4635$) as separate clusters and joining them, two at a time, until one cluster is left. The problem is to find the number of clusters which should be small, but the information content of each cluster should be large. In order to estimate the optimal number of clusters in the data set it is interesting to follow the evolution of the cubic clustering criterion (CCC) during the clustering process. The criterion has a very complicated, empirical definition and its use is not without pitfalls; no attempt will be made to discuss CCC here. The interested reader is referred to Sarle [23] for more detailed information. In the present context it is sufficient to know that the existence of peaks indicates good clusterings. In our case we conclude in favor of an optimal number of five or six clusters.

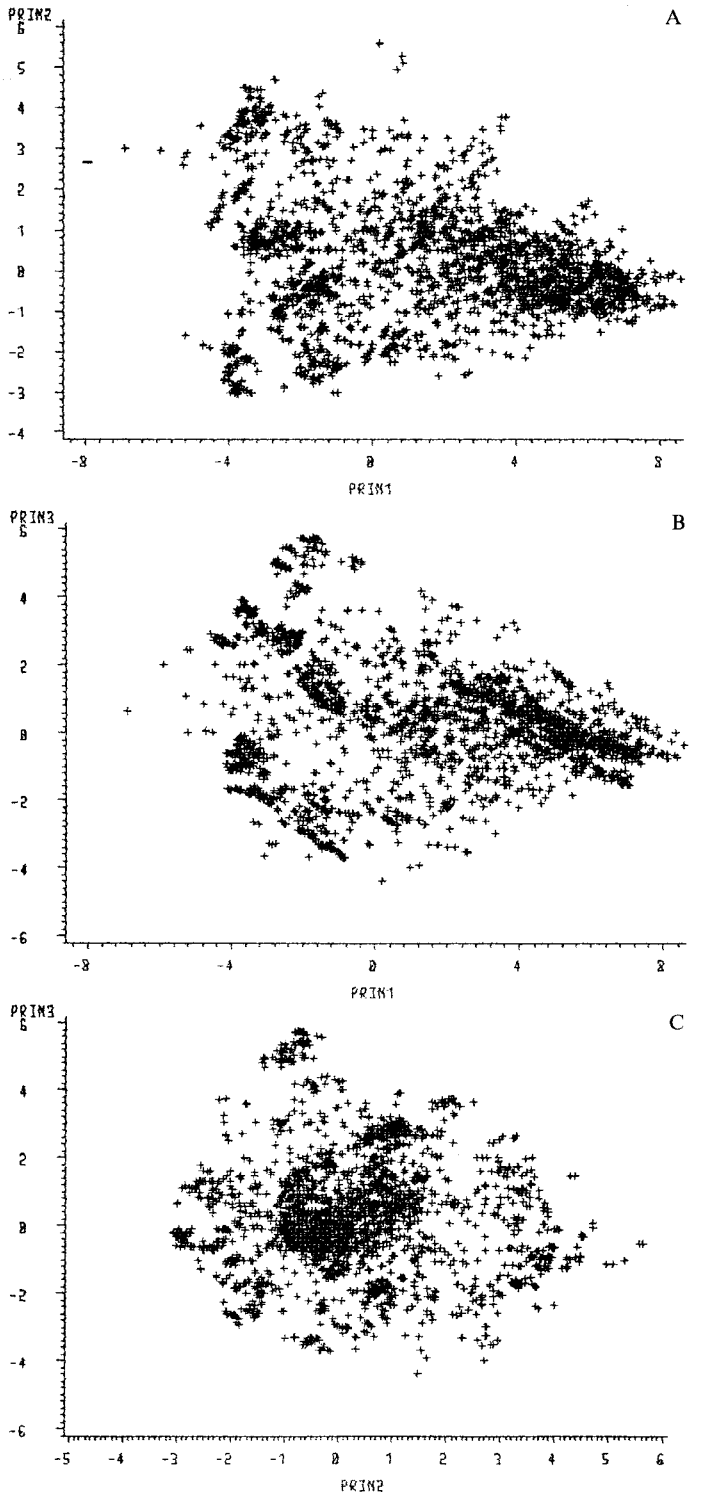

Fig. 1. Projections of the 4635 conformers on the three first principal axes of the sample.

Furthermore, the CLUSTER procedure allows one to get the number of data points (conformations) for each one of the six clusters. We can now compare these values with those obtained by the second method, the non-hierarchical FASTCLUS procedure, which is designed for disjoint clustering of large data sets. We have run this method with different numbers of clusters and we will specially examine the results for five and six clusters as suggested by the CCC from the first method. In Table 1 a comparison of the number of data points in each cluster obtained by the two methods shows that these values are very close for six clusters, whereas for five clusters the discrepancy (Δ) is significant. The agreement between the results obtained at least for two clustering methods may be taken as evidence that a cluster number of six is inherent to the data set and not an artifact of the clustering method used.

A canonical discriminant anlysis performed on the clustering results gives a better insight into the cluster structure. In a similar way to PCA, this method reduces the dimensionality of the variable space without losing too much of the initial information content of the original data set. The canonical variables (linear combinations of the initial quantitative variables) obtained by that way are strongly correlated to the six clusters and especially the first canonical variable presents the strongest multiple correlation with each of the clusters. This first canonical variable allows a good separation of the clusters and gives a clear picture of the relative position of the clusters in the data set of conformations.

INTERPRETATION OF THE CLUSTERS OF CONFORMATIONS

The sample of 4635 molecular conformations built from the Monte Carlo procedure is quite well separated into six stable clusters. A preliminary idea of the typical conformational feature for each cluster is given by the mean values of the end-to-end and Tyr^{27}-Phe^{33}, Tyr^{27}-Trp^{30}, Trp^{30}-Phe^{33} side-chain centroïd distances (Table 2). Cluster 1 which represents 21% of all the analysed

TABLE 1
COMPARISON OF THE CLASSES OBTAINED FROM A HIERARCHICAL CLUSTERING ALGORITHM (PROCEDURE CLUSTER OF THE SAS PACKAGE) AND A PARTITION ALGORITHM (PROCEDURE FASTCLUS OF THE SAS PACKAGE)

Cluster	n=6			n=5		
	Number of conformations			Number of conformations		
	Hierarch.	Partition	Δ[a]	Hierarch.	Partition	Δ[a]
1	952	964	12	872	952	80
2	1133	1096	37	521	429	92
3	429	472	43	1694	1240	454
4	641	656	15	651	881	230
5	881	898	17	897	1133	236
6	599	549	50			

[a]Difference in number of conformations between the hierarchical and the partitioned clustering.

TABLE 2
TYPICAL MEAN DISTANCES OBTAINED IN THE CLASSES AND IN THE WHOLE SAMPLE

Cluster	End-to-end distance (Å)	Side-chain centroïd distances (Å)		
		Tyr-Phe	Tyr-Trp	Trp-Phe
1	20.8	19.4	13.9	9.3
2	13.9	13.3	11.5	8.7
3	14.1	13.6	10.6	7.8
4	14.8	12.5	6.9	6.4
5	8.3	5.2	6.7	6.4
6	8.4	9.6	10.2	10.6
In the whole sample	13.7	12.5	10.7	10.6

structures consists of the most extended conformations with the aromatic side chains pointing well apart from the backbone. Cluster 2 which predominates (24% of the sample conformations) as well as clusters 3 (10%) and 4(14%) are characterized by conformations with an S-folded backbone or containing helical fragments. In cluster 2 the aromatic side chains do not present any privileged orientation, as do the Trp^{30} and Tyr^{27} rings in cluster 3 which lie quite near and parallel to one another. The situation is similar for the Trp^{30} and Tyr^{27} rings in cluster 4. Clusters 5 and 6 (19% and 12% of the sample conformations) include the most folded conformations, but they differ in the relative orientation of the three aromatic side chains. In cluster 5 the aromatic rings are quite close to one another with a tendency for Tyr^{27} and Phe^{33} to be parallel, while in cluster 6 the aromatic side chains are more apart from each other and extend in different directions.

Having used the intramolecular distances for the cluster analysis we now look back to the usual (φ, ψ, $\{\chi\}$) dihedral angles which allow a three-dimensional representation of the CCK8 molecule for each of the six clusters (Figs. 2A–F). These typical conformations will be considered as representative of the behavior of the CCK8 fragment in aqueous medium at neutral pH. To obtain these conformations for each cluster, we have proceeded in the following three ways:

(i) the repesentative conformation of the cluster is the one with the lowest conformational energy;
(ii) this conformation is the one nearest to the centroïd of the cluster;
(iii) we calculate a 'mean' conformation which is defined by the mean values of the dihedral angles of the conformations belonging to each cluster.

A useful representation of these typical conformations is to plot the (φ, ψ) angles for each residue of the CCK8 molecule on the usual Ramachandran (φ, ψ) map (Fig. 3). A conformation of CCK8 is depicted by eight points on the map, – one for each residue. We use different symbols to distinguish between the six clusters. To interpret the results, we consider the maps obtained with the centroïd conformations nearest to the cluster. Torsion angles and standard deviations (for the backbone dihedral angles) are collected in Table 3. The values of the standard deviations are important because of the non-unimodal angular distributions in each cluster.

By examining the map for cluster 1, it appears that the (φ, ψ) angles of the residues belong to the β-extended region; this explains the extended conformation of the CCK8 molecule. Conformations in cluster 2 are also defined by (φ, ψ) angles in the β-extended region, except for the

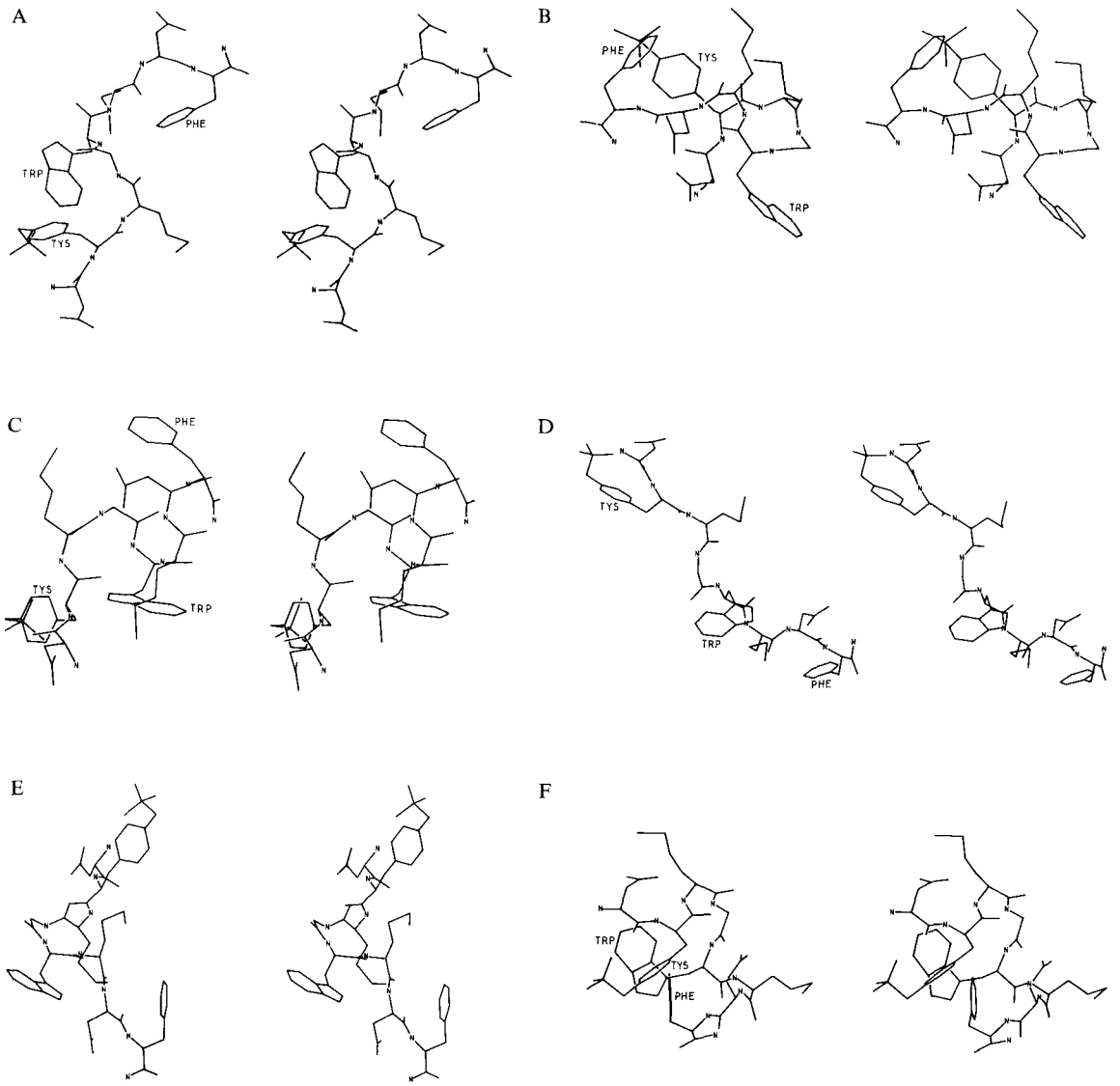

Fig. 2. Typical conformers in each of the six obtained clusters (A: cluster 1,B: cluster 2., etc.). These conformations are the ones found to be the nearest to the centroid of each class.

Gly^{29} residue which is in the α_R region. Because of its particular conformation, this residue could play a central role in the formation of the folded conformations of the CCK8 molecule.

In cluster 3, the Met^{31}, Asp^{32} and Phe^{33} residues lie in neighbourhood of the α_R region and we notice that these residues form the C-terminal moiety of the CCK8 fragment.

Cluster 4 reveals residues which are in particular conformations: a C_7 structure for Gly^{29} which induces a γ-bend including the sequence Met^{28}-Gly^{29}-Trp^{30} and a folding like a β-bend concerning the Gly^{29}-Trp^{30}-Met^{31}-Asp^{32} sequence where the two Trp^{30} and Met^{31} residues act as a joint. Such a tridimensional structure based on a double folding, a β-bend for the C-terminal moiety followed

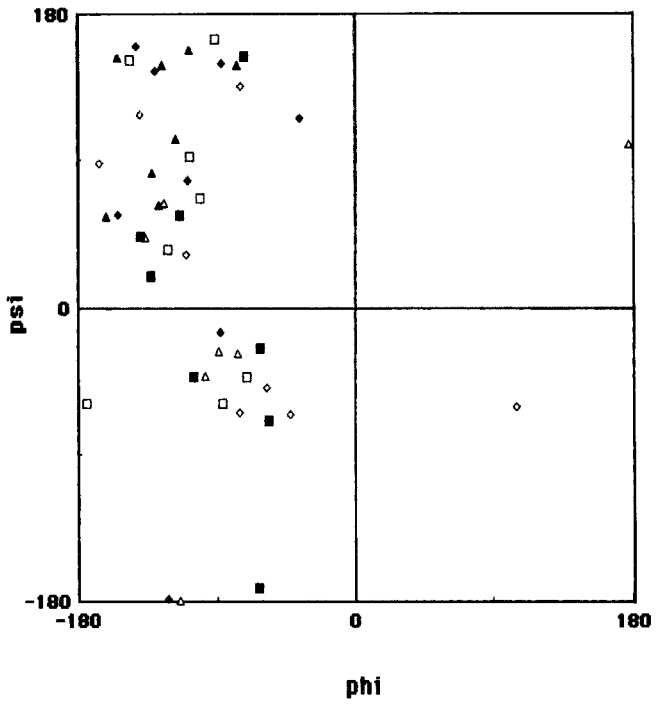

Fig. 3. Ramachandran plot of the typical conformation in six clusters. Each cluster is characterized by a different symbol: ▲ cluster 1, ◆ cluster 2, △ cluster 3, ◇ cluster 4, ■ cluster 5 and □ cluster 6. A conformation is depicted by eight points corresponding to the eight residues of the sulfated CCK8 molecule.

by a γ-bend for the N-terminal one, has been proposed by Durieux et al. [8]who have studied the CCK7 (CCK^{27-33}) fragment by NMR.

The conformations of cluster 5 do not exhibit any β-bend, but a tendency for the Met^{28}, Met^{31}, Asp^{32} residues to adopt an α_R conformation which explains the compactness of the structure of CCK8 in that cluster. Similar conclusions can be drawn for cluster 6 where the amino acids Met^{28} and Gly^{29} are in the α_R region. It is interesting to note the role of the Gly^{29} residue to act as a joint as already mentioned before.

The experimental results obtained by Fournié-Zakuski et al. [7] using NMR spectroscopy and fluorescence energy transfer measurements on CCK8 and analogs allow one to conclude in favor of folded conformations which are preferentially stabilized in aqueous medium. Furthermore, the sulfonated Tyr^{27} residue points clearly away from the backbone of the CCK8 molecule.

If we compare their results with ours, obtained for the six conformational cluster types of CCK8, it appears that the folded conformations are predominant, except for cluster 1, insofar as clusters 2 to 6 represent about 80% of the whole sample of 4635 conformations generated by the Monte Carlo method. Furthermore the Tyr^{27}-Trp^{30} distance measured by fluorescence energy transfer [7] is of the order of 15 Å which is also in favor of folded conformations. This value is in quite good agreement with those calculated for the clusters 2 to 6 (Table 2). We also want to stress the particular orientation (outside) of the sulfonated Tyr^{27} in clusters 5 and 6 (Figs. 2E and F) which could explain the cholecystokinic properties of the CCK8 fragment.

TABLE 3
DIHEDRAL ANGLES OF THE REPRESENTATIVE CONFORMERS IN EACH OF THE SIX CLASSES FOR EACH RESIDUE (STANDARD DEVIATIONS ARE GIVEN IN PARENTHESES)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Asp[26]						
φ	−76	−35	178	−166	−63	91
	(57)	(60)	(65)	(69)	(62)	(60)
ψ	149	116	100	88	−172	165
	(124)	(111)	(125)	(109)	(71)	(115)
(SO_3)Tyr[27]						
φ	−153	−121	−114	−139	−114	−175
	(36)	(32)	(29)	(31)	(20)	(46)
ψ	153	−179	−180	118	56	−59
	(116)	(146)	(124)	(104)	(61)	(116)
Met[28]						
φ	−131	−86	−124	−74	−56	−86
	(36)	(37)	(17)	(22)	(26)	(80)
ψ	82	150	63	136	−69	−59
	(89)	(79)	(45)	(43)	(82)	(73)
Gly[29]						
φ	−127	−87	−136	105	−139	−71
	(88)	(93)	(120)	(68)	(102)	(111)
ψ	62	−16	42	−61	44	−42
	(108)	(76)	(38)	(55)	(25)	(52)
Trp[30]						
φ	−125	−129	−72	−57	−72	−102
	(33)	(26)	(27)	(31)	(14)	(21)
ψ	148	145	154	−49	154	93
	(49)	(88)	(113)	(22)	(91)	(88)
Met[31]						
φ	−116	−108	−88	−75	−105	−101
	(66)	(30)	(28)	(22)	(23)	(79)
ψ	103	77	−27	−65	−43	67
	(30)	(110)	(36)	(43)	(16)	(83)
Asp[32]						
φ	−107	−141	−76	−42	−62	−146
	(30)	(33)	(28)	(48)	(20)	(30)
ψ	158	160	−29	−66	−25	152
	(61)	(66)	(60)	(100)	(59)	(87)
Phe[33]						
φ	−161	−154	−97	−109	−132	−122
	(28)	(24)	(23)	(28)	(19)	(43)
ψ	55	57	−43	32	19	35
	(55)	(51)	(60)	(48)	(42)	(54)

CONCLUSION

Clustering methods applied to a sample of 4635 conformations of CCK8 generated by a Monte Carlo algorithm, allow the determination of molecular models which confirm that the CCK8 molecule in aqueous medium at physiological pH adopts preferentially folded conformations, in good accordance with experimental measurements.

ACKNOWLEDGEMENTS

We would like to acknowledge the Laboratoire de Cristallographie, University of Bordeaux I, for the molecular drawings and the Centre National Universitaire Sud de Calcul (Montpellier, France) for computational facilities. We also thank Dr. M. Cotrait for stimulating discussions.

REFERENCES

1 Vanderhaegen, J.J. and Cranley, J.N. (Eds.), Ann. N.Y. Acad. Sci., 448 (1985).
2 Gardner, J.D., Walker, M.D., Martinez, J.,Priestly, G.P., Natarajan, S. and Bodanszky, M., Biochim. Biophys. Acta, 630 (1980) 323–329.
3 Dockray, G.J., Gregory, R.A., Hutchison, J.B., Harris, J.I. and Runswick, M.J., Nature, 270 (1977) 359–360.
4 Robberecht, P., Deschodt-Lanckmann, M. and Vanderhaegen, J.J., Proc. Natl. Acad. Sci., USA, 75 (1978) 524–528.
5 Shirboll, L.T., Grace, A.A., Hommer, D.W., Rehfeld, J., Goldstein, M., Hockfelt, T. and Bunney, B.S., Neuroscience, 6 (1981) 2111–2124.
6 Gaudreau, P., Quirion, R., St. Pierre, S. and Pert, C.B., Eur. J. Pharmacol., 87 (1983) 173–174.
7 Fournié-Zaluski, M.C., Belleney, J., Lux, B., Durieux, C., Gérard, D., Gacel, G., Maigret, B. and Roques, B.P., Biochemistry, 25 (1986) 3778–3787.
8 Durieux, C., Belleney, J., Lallemand, J.Y., Roques, B.P. and Fournié-Zaluski, M.C., Biochem. Biophys. Res. Commun., 114 (1983) 705–712.
9 Metropolis, N.A., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E.J., Chem. Phys., 21 (1953) 1087–1092.
10 Premilat, S. and Maigret, B., J. Chem. Physiol., 66 (1977) 3418–3420.
11 Paine, G.H. and Scheraga, H.A., Biopolymers, 24 (1985) 1391–1436, 25 (1986) 1547–1563 and 26 (1987) 1125–1162.
12 Betins, J., Nikiforovitch, G.V. and Chipens G., J. Mol. Struct., 137 (1986) 129–132.
13 Li, Z. and Scheraga, H.A., Proc. Natl. Acad. Sci. USA, 84 (1987) 6611–6615.
14 Marchionini, C., Maigret, B. and Premilat, S., Biochem. Biophys. Res. Commun., 112 (1983) 339–346.
15 Lowry, G.G. (Ed.), Markov Chain and Monte Carlo Calculations in Polymer Science, Dekker, New York, NY, 1970.
16 Scheraga, H.A., Adv. Phys. Org. Chem., 6 (1968) 103–184.
17 Poland, D. and Scheraga, H.A., Biochemistry, 6 (1967) 3791–3800.
18 Duran, B.S. and Odell, P.L., Cluster Analysis: A Survey, Lecture Notes in Economics and Mathematical Systems Springer-Verlag, New York, NY, 1974.
19 Massart, D.L. and Kaufman, L., The Interpretation of Analytical Chemistry Data by the Use of Cluster Analysis (Chemical Analysis Monographs), Wiley, New York, NY, 1983.
20 Duda, R.O. and Hart, P.E., Pattern Classification and Scene Analysis, Wiley, New York, NY, 1973, Ch. 6.
21 SAS Institute Inc., Box 8000, Cary, NC 27511.
22 Ward, J.H., J. Am. Statist. Assoc., 58 (1963) 236–244.
23 Sarle, W.S., SAS Technical Report A-108, SAS Institute Inc., Box 8000, Cary, NC 27511, 1983.