Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

Bayesian methods for the conformational classification of eight-membered rings

J. Pérez,[a]* K. Nolsøe,[b] M. Kessler,[b]* L. García,[a] E. Pérez[a] and J. L. Serrano[a]

[a]Departamento de Ingeniería Minera, Geológica y Cartográfica, Área de Química Inorgánica, Universidad Politécnica de Cartagena, Spain, and [b]Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena, Spain

Correspondence e-mail: jose.pperez@upct.es, mathieu.kessler@upct.es

Received 13 May 2005
Accepted 26 July 2005

Two methods for the classification of eight-membered rings based on a Bayesian analysis are presented. The two methods share the same probabilistic model for the measurement of torsion angles, but while the first method uses the canonical forms of cyclooctane and, given an empirical sequence of eight torsion angles, yields the probability that the associated structure corresponds to each of the ten canonical conformations, the second method does not assume previous knowledge of existing conformations and yields a clustering classification of a data set, allowing new conformations to be detected. Both methods have been tested using the conformational classification of Csp^3 eight-membered rings described in the literature. The methods have also been employed to classify the solid-state conformation in Csp^3 eight-membered rings using data retrieved from an updated version of the Cambridge Structural Database (CSD).

1. Introduction

The conformational analysis of organic (Allen & Motherwell, 2002) and metallic complexes (Zimmer, 2001) is an active research area, the CSD being a powerful tool in this kind of study (Allen & Taylor, 2004; Orpen, 1993). Despite the large amount of structural data available, a full understanding of the factors that determine the molecular structure of a particular compound has not yet been achieved. In coordination and organometallic chemistry the manner in which a ligand controls the properties of the complex depends on a combination of steric, electronic and conformational factors. Detailed knowledge of these effects will allow a rational design of complexes with specific and predictable properties (Meyer, 1989).

A review of the different statistical methods for conformation analysis can be found in Zimmer (2001). Reviews such as this generally take a data-analysis approach where no model is assumed for the data generation mechanism, and all the conclusions rely on the correlation structure of the data or their similarities. Cluster analysis and principal-component analysis are examples of such methods. In contrast, an essential step in our approach consists of specifying a probabilistic model for the observed sequences of torsion angles. This model is mainly based on assuming that the sequences of torsion angles are generated, after a perturbation that takes into account measurement errors, from a number k of 'preferred' conformations. Two levels of generality can then be chosen: either the 'preferred' conformations are assumed to be provided, *a priori*, by the user; for example, they could consist of the ten canonical conformations for cyclooctane, as described in Hendrickson (1967) and shown in Fig. 1; or no assumption is made about the 'preferred' conformations nor

about their number. Associated with these two levels of generality, we propose the following two methods:

(i) First level of generality: the 'preferred' conformations are provided by the user: an individual classification of the observed structures is performed. Based on the eight values of the torsion angles for a structure, it is possible, through Bayes' rule, to compute the posterior probability that the structure comes from each of the preferred conformations. These probabilities provide more information than only a classification: their relative order of magnitude indicates, in particular, the strength of the evidence in the data in favour of a given conformation. Likewise, similarities between conformations can be detected. We will refer to this method as the '*Classification method*'.

(ii) Second level of generality: no previous knowledge of the 'preferred' conformations is assumed: Bayesian inference on the number of conformations, the conformations themselves, their frequencies of occurrence, as well as the standard deviations associated to each conformation were determined. As a result of the Bayesian approach, a posterior distribution for each of the parameters of interest can be obtained. The structures in each of the obtained conformations can also be classified. We will refer to this method as the '*Full Bayesian Analysis method*'.

Notice that the Classification method performs the individual classification of structures but requires *a priori* specification of the 'preferred' conformations. The Full Bayesian Analysis applies to a set of structures but allows for the detection of new conformations and is not dependent on theoretical canonical conformations.

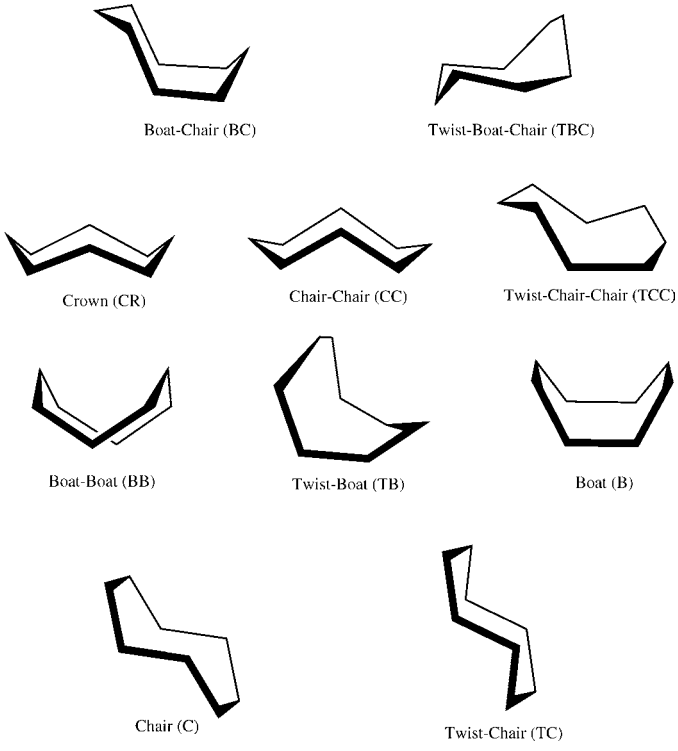

Figure 1
Canonical forms of cyclooctane.

In order to test the methods described in this article we have studied the data of the Csp^3 eight-membered rings analyzed by Allen and co-workers (Allen *et al.*, 1996) using cluster and principal components techniques. We have also studied the data of Csp^3 eight-membered rings extracted from an updated CSD version. It is to be stressed that the proposed methods can easily be extended to rings with differing numbers of atoms.

2. Theory

2.1. The model

The eight torsion angles observed for a given structure retrieved from the CSD are denoted by $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3, \tau_4, \tau_5, \tau_6, \tau_7, \tau_8)$. The model we assume for the data-generation mechanism resulting in a realisation of $\boldsymbol{\tau}$ is built up in three steps:

(i) Randomly choose one of the k 'preferred' conformations, according to the probabilities $p_1, p_2, \ldots, p_k$. These probabilities are unknown parameters that correspond to the natural frequency of occurrence of each 'preferred' conformation. We denote by C the index of the chosen conformation (C thus ranges from 1 to k).

(ii) Let $\boldsymbol{\mu}(C) = (\mu_{C,1}, \mu_{C,2}, \mu_{C,3}, \mu_{C,4}, \mu_{C,5}, \mu_{C,6}, \mu_{C,7}, \mu_{C,8})$ be the sequence of torsion angles associated with conformation C. The observed values of the torsion angles in τ may correspond to a different starting point in the structure than that in the canonical sequence $\boldsymbol{\mu}(C)$. To take this fact into account, a starting point ν between 1 and 8 was randomly chosen, with equal probabilities and the cyclically translated sequence

$$\mu(C,\nu) = (\mu_{C,\nu}, \mu_{C,((\nu)\bmod 8+1)}, \mu_{C,(\nu+1)\bmod 8+1}, \ldots, \mu_{C,(\nu+6)\bmod 8+1})$$

was constructed, where, for any integer j, j mod 8 denotes j modulo 8, *i.e.* the remainder of the integer division of j by 8. Moreover, the sequence of torsion angles can be read in a clockwise or counter-clockwise manner. The counter-clockwise version of $\boldsymbol{\mu}(\mathrm{C},\nu)$ is readily obtained as

$$\mu(C,\nu) = (\mu_{C,\nu}, \mu_{C,(\nu+6)\bmod 8+1}, \mu_{C,(\nu+5)\bmod 8+1}, \ldots, \mu_{C,(\nu\bmod 8+1)}).$$

Let us now introduce the variable d which takes the values 1 or -1 according to whether the direction of the rotation is clockwise or counter-clockwise. The two previous formulae can now be summarized as

$$\mu(C, \nu, d) = (\mu_{C,\nu,\mu C,(\nu-1+d\times 1)\bmod 8+1}, \mu_{C,(\nu-1+d\times 2)\bmod 8+1}, \ldots, \mu_{C,(\nu-1+d\times 7)\bmod 8+1}).$$

It is also necessary to consider the coordinate inversions from which sequences of torsion angles are readily obtained by a change of sign. We therefore introduce the random variable δ which can take either the value 1 or -1 with equal probabilities and

$$\mu(C, \nu, d, \delta) = \delta \times (\mu_{C,\nu,\mu C,(\nu-1+d\times 1)\bmod 8+1}, \mu_{C,(\nu-1+d\times 2)\bmod 8+1}, \ldots, \mu_{C,(\nu-1+d\times 7)\bmod 8+1}).$$

As an example, consider conformation BC, index C = 1 in *Appendix A*, we have

$$\mu(1, 2, -1, 1) = (44.7, 65.0, -65.0, -44.7, 102.2, -65.0, 65.0, -102.2).$$

(iii) Finally, we consider that the observed sequence is obtained from $\mu(C, \nu, d, \delta)$ after an additive perturbation $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5, \varepsilon_6, \varepsilon_7, \varepsilon_8)$, that is

$$\tau = \mu(C, \nu, d, \delta) + \varepsilon,$$

where the perturbation's components $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5, \varepsilon_6, \varepsilon_7, \varepsilon_8$ are assumed to be independent Gaussian random variables with a zero mean and then unknown variance parameter σ_c^2, which may depend on the conformation C.

As a conclusion, from the relation between $\boldsymbol{\tau}, \boldsymbol{\mu}(C, \nu, \delta)$ and ε, in step (iii) we deduce that the density $(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5, \tau_6, \tau_7, \tau_8) \rightarrow f(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5, \tau_6, \tau_7, \tau_8)$ of the random variable τ is easily computable: it is a mixture of multivariate laws

$$f(\tau) = \sum_{c=1,\ldots,k} p_c f(\tau, c),$$

where $f(\boldsymbol{\tau},\mathbf{c})$ are themselves mixtures of multivariate Gaussian laws

$$f(\tau, c) = \sum_{\nu=1,\ldots,8} \sum_{d=-1,1} \sum_{\delta=-1,1} f_G\left(\tau, \mu(c, \nu, d, \delta), \sigma_c^2\right),$$

$\boldsymbol{\tau} \rightarrow f_G\,(\boldsymbol{\tau}, \boldsymbol{\mu}(c, \nu, d, \delta), \sigma_c^2)$ denoting the density of the eight-dimensional Gaussian law with mean $\mu(c, \nu, d, \delta)$ and diagonal covariance matrix $\sigma_c^2 Id$.

When analysing torsion angle data, the symmetry of the conformation space has to be taken into account. In particular, to be able to apply the principal-component analysis, for example, the initial torsion angle data have to be expanded by symmetry. In our paper, we avoid expanding the data by incorporating the symmetries in the model formulation through the starting point, rotation direction and sign of the torsion angles, as described above.

3. The Classification method

As explained in §1, the Classification method assumes the first level of generality for the model: the 'preferred' conformations are supplied by the user. In the following these will be taken to be the ten canonical conformations crown (D_{4d}), twist-boat (S_4), boat–boat (D_{2d}), boat (D_{2d}), twist-chair–chair (D_2), chair–chair (C_{2v}), chair (C_{2h}), twist-chair (C_{2h}), twist-boat–chair (C_2) and boat–chair (C_s), as described in Allen *et al.* (1996), for example. The table in *Appendix A* contains the torsion angles corresponding to these canonical conformations.

Using Bayes' rule we are able to compute, given an observed sequence $(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5, \tau_6, \tau_7, \tau_8)$, the probability that it was generated from the conformation c:

$$P(C = c|\tau = (\tau1, \tau2, \tau3, \tau4, \tau5, \tau6, \tau7, \tau8)) = pc\, f(\tau, c) \Big/ \left(\sum_{c'=1,\ldots,10} pc\, f(\tau, c') \right).$$

The computation of these probabilities requires the specification of a prior distribution both for $(p1, p2, p3, p4, p5, p6, p7, p8, p9, p10)$ and for σ_c, the standard deviation of the perturbations ε, which are to reflect our prior knowledge, if any, about these quantities. As for the proportions, we make the choice $pc = 1/10$, for $c = 1, \ldots, 10$, which means that we do not favour *a priori* any particular canonical conformation. For σ_c, if we choose $\sigma_c = 10^\circ$, from the well known property of the Gaussian law, 95% of the values taken by the perturbations will lie between -20 and 20°, which seems a reasonable range of values. We recommend repeating the analysis for different values of σ_c, in order to check that the classification results are not too sensitive to changes in values of σ_c.

For an illustration, we finish this section by applying the Classification method to the YOPPIC structure (Villar *et al.*, 1995), retrieved from the CSD containing two molecules in the asymmetric unit with torsion angles: $\boldsymbol{\tau}(a)$ = (44.93, −47.95, 98.13, −53.31, −57.89, 81.61, 13.77, −72.05) and $\boldsymbol{\tau}(b)$ = (4.58, 21.13, −89.86, 55.43, 55.75, −89.68, 15.40, 18.44). Fig. 2 illustrates the posterior probabilities of each of the 10 canonical conformations in the case of $\sigma = 10^\circ$. For both molecules the boat–chair (BC) conformation is the most likely one. In the case of YOPPIC1, the posterior probability of the BC conformation is almost one, while for YOPPIC2 a posterior probability of 0.21 is assigned to the boat–boat (BB) conformation and a posterior probability of approximately 0.79 to the BC conformation, indicating that the structure is intermediate between the BB and BC conformations.

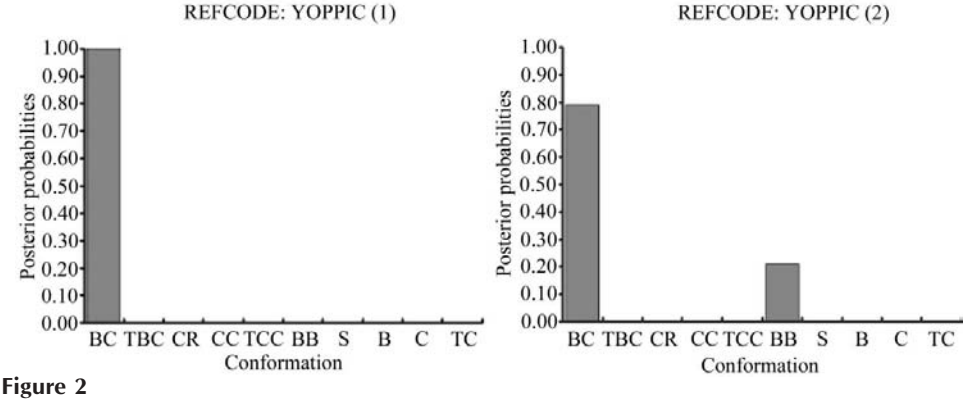

Figure 2
Posterior probabilities for the two molecules in the asymmetric unit of YOPPIC.

4. The Full Bayesian Analysis method

For this method no previous knowledge of the 'preferred' conformations is assumed, in particular, the number of these 'preferred' conformations is unknown. The parameters of the statistical model we consider are therefore: the number k of 'preferred' conformations, the k

Table 1
Conformational analysis of Csp^3 eight-membered rings in Allen *et al.* (1996) according to the methods described in this paper.

	Classification Method		Full Bayesian method
	$\sigma = 10°$	$\sigma = 20°$	
AMCOCA	1.00 BC	1.00 BC	$\mu(1)$
BAGPII	1.00 BC	1.00 BC	$\mu(1)$
BCOCTB	1.00 BC	1.00 BC	$\mu(1)$
CLCOCT	1.00 BC	1.00 BC	$\mu(1)$
COCOAC	1.00 BC	1.00 BC	$\mu(1)$
COCOXA10	1.00 BC	1.00 BC	$\mu(1)$
COVLUU	1.00 BC	0.99 BC; 0.01 TBC	$\mu(2)$
CURBIA	1.00 BC	1.00 BC	$\mu(1)$
CUVZEY	1.00 BC	1.00 BC	$\mu(1)$
CYOCDL	1.00 BC	1.00 BC	$\mu(1)$
DEZPUT	0.73 CR; 0.03 CC; 0.24 TCC	0.67 CR; 0.15 CC; 0.19 TCC	$\mu(4)$
ECOTDA	0.90 BC; 0.10 TBC	0.51 BC; 0.49 TBC	$\mu(3)$
EOCNON10	1.00 TCC	0.06 CC; 0.94 TCC	$\mu(5)$
GATRAU	1.00 BC	1.00 BC	$\mu(1)$
GIVBAO	1.00 BC	1.00 BC	$\mu(1)$
HOXTHD	1.00 BC	0.98 BC; 0.02 TBC	$\mu(2)$
HUMULB10	0.12 BC; 0.88 TBC	0.27 BC; 0.73 TBC	$\mu(3)$
KESVIN	1.00 BC	0.98 BC; 0.02 TBC	$\mu(2)$
OCSHYD	1.00 BC	1.00 BC	$\mu(1)$
PCDODO	1.00 BC	0.78 BC; 0.22 TBC	$\mu(3)$
SATKIH (1)	0.41 CC; 0.59 TCC	0.47 CC; 0.52 TCC	$\mu(6)$
SATKIH (2)	0.08 CC; 0.92 TCC	0.35 CC; 0.65 TCC	$\mu(6)$
SATKIH (3)	0.95 CC; 0.05 TCC	0.01 CR; 0.65 CC; 0.34 TCC	$\mu(6)$
SATKIH (4)	0.21 CR; 0.04 CC; 0.75 TCC	0.55 CR; 0.18 CC; 0.27 TCC	$\mu(4)$
SEJFIW (1)	0.89 BC; 0.08 TBC; 0.03 TC	0.38 BC; 0.33 TBC; 0.29 TC	$\mu(7)$
SEJFIW (2)	0.79 BC; 0.19 TBC; 0.02 TC	0.35 BC; 0.35 TBC; 0.30 TC	$\mu(7)$
SPOCTC10	1.00 BC	1.00 BC	$\mu(1)$
SPTZBN	0.88 BC; 0.12 TBC	0.49 BC; 0.51 TBC	$\mu(3)$
VALGOE (1)	1.00 BC	0.99 BC; 0.01 TBC	$\mu(1)$
VALGOE (2)	1.00 BC	0.99 BC; 0.01 TBC	$\mu(1)$
VASWOB	1.00 BC	0.97 BC; 0.03 TBC	$\mu(2)$

sequences of torsion angles corresponding to these conformations, $\boldsymbol{\mu}(1), \ldots, \boldsymbol{\mu}(k)$, their corresponding frequencies of occurrence $p1, \ldots, pk$, and the associated standard deviations $\sigma_1, \ldots, \sigma_k$. The Bayesian approach consists of first providing *prior distributions* for these parameters and updating these distributions with the information provided by the observed data through Bayes' rule, to finally compute the *posterior distributions* of the parameters. As mentioned in §2.1 we describe the data by a multivariate mixture model. The problem of Bayesian inference in mixture models has been extensively studied in the last decade. Useful reviews can be found in Robert & Casella (2005) or Robert (1996). As described in these reviews, no analytically tractable form for the posterior distributions of the parameters is available. It is, however, possible to simulate as many samples of these posterior distributions as desired using a Markov Chain Monte Carlo algorithm, for a simple introduction see Robert & Casella (2005) and for a description of the application to mixture models see Robert (1996). From the samples it is then possible to obtain approximations of any quantity of interest related to the posterior distribution: a histogram of the draws for example can provide an approximation to the density. The case when the number of components in the mixture is itself a parameter requires a somehow more sophisticated Markov Chain Monte Carlo algorithm called Reversible-Jump algorithm, which has been described in Richardson & Green (1997). A detailed description of the mathematical aspects of the 'Full Bayesian Analysis' method we have implemented can be found in Nolsøe *et al.* (2005), which is available from the authors upon request.

The output of the 'Full Bayesian Analysis' method consists of, on one hand, a histogram of the number of 'preferred' conformations after the structures have been analyzed, providing probabilities that the data have been generated from a one-, two-, three- *etc* conformations mixture. For each of the possible estimated number of conformations, one obtains as well histograms of the posterior distributions of the 'preferred' conformations torsion angles, of their frequencies and their associated standard deviations.

5. Experimental

5.1. Structural analysis

The Cambridge Structural Database (Allen, 2002), Version 5.25, was searched for all the structures containing Csp^3 eight-membered rings in organic structures. A total of 95 refcodes matched the search, the total number of fragments was 115. Torsion angles were tabulated and transferred to Excel for statistical analysis, these data are included in the supplementary material.[1]

The Classification method is very simple to implement and was programmed in Visual Basic and incorporated as a macro in *Excel*. The Full Bayesian Analysis method is not so straightforward to implement and was programmed in *Java*.

6. Results and discussion

6.1. Conformational analysis of Csp^3 eight-membered rings in Allen *et al.* (1996).

6.1.1. Classification method: the preferred conformations are assumed to be the canonical conformations. In order to test the conformational classification method described above we employed Csp^3 eight-membered rings. We chose this system because the conformations of cyclooctane and related eight-membered rings have been widely studied, both theoretically (Hendrickson, 1967) and experimentally (Allen *et al.*, 1996; Evans & Boeyens, 1988). Table 1 shows the most likely canonical conformations, together with the associated probabilities, that were deduced from the computation of the posterior probabilities with the Classification method, for the

[1] Supplementary data for this paper are available from the IUCr electronic archives (Reference: BS5019). Services for accessing these data are described at the back of the journal.

rings with Csp^3 atoms analyzed by Allen and co-workers (Allen *et al.*, 1996).

As can be seen in Table 1 the most frequent conformation is boat–chair (BC), this is the most likely conformation in 23 of the 31 data analyzed. The twist-boat–chair (TBC) and twist-chair–chair (TCC) conformations often exhibited a significant probability in the structures reviewed in Table 1. These conformations (BC, TBC and TCC) in cyclooctane have been identified as energy minima with respect to all the small distortions (Anet & Krane, 1973), with BC being the most stable conformation.

In Table 1, 13 structures have a probability of 1 for the BC conformation with σ = 10 or 20°. These structures exhibit torsion angles similar to those of the ideal BC conformation and can be identified as free (non-fused) cyclooctane rings in accordance with the results of Allen and co-workers (Allen *et al.*, 1996). Interestingly, the appearance of a non-zero probability for more than one ideal conformation can be used to identify a structure as intermediate between two or more theoretical conformations. Thus, some structures in Table 1 show a small probability (0.01–0.02) for the TBC conformation, but only in the case of σ = 20°. All the structures show a similar trend in the deviation from the ideal BC conformation: the ideal BC angles of 44.7 and −102° (see *Appendix A*) show absolute value ranges of 19–30 and 88–94°. These structures correspond to the cluster BC/TBC described by Allen and co-workers (Allen *et al.*, 1996). Significantly, a value of σ = 20° was necessary to detect this conformational feature; the advantage of using σ = 20° rather than σ = 10° in order to detect the intermediate character of a conformation can be observed throughout Table 1.

In some cases the deviation from the BC conformation is larger (ECOTDA, HUMULB10, SPTZBN) and a significant probability for TBC conformation appears even with σ = 10°. This corresponds to the second BC/TBC cluster identified by Allen and co-workers (Allen *et al.*, 1996). The remaining structures are distorted conformations with significant probabilities for CR/CC/TCC or BC/TBC/TC conformations.

6.1.2. Full Bayesian Analysis method: no previous knowledge of preferred conformations assumed. We have also studied the set of data analyzed by Allen and co-workers (Allen *et al.*, 1996) by the Full Bayesian method without *a priori* knowledge of the ideal conformations described previously. A histogram for the posterior distribution of k, the number of detected conformations, is given in Fig. 3.

Six or seven clusters are found to be most likely. In Fig. 4, box-plot-like diagrams are presented for the posterior distributions of the detected conformational sequences of torsion angles $\boldsymbol{\mu}(1), \ldots, \boldsymbol{\mu}(7)$, their corresponding frequencies of occurrence $p1, \ldots, p7$, and the associated standard deviations $\sigma_1, \ldots, \sigma_7$, when seven clusters are chosen. Eight horizontal lines were drawn for each box, representing the percentiles 10, 15, 25, 50, 75, 85 and 90% of the distribution. Moreover, in the case of the conformations $\boldsymbol{\mu}(1), \ldots, \boldsymbol{\mu}(5)$, a wider horizontal black line was drawn which represents the torsion angles of the representative member of clusters detected by Allen *et al.* (1996). The torsion angles of the centroids for the seven clusters are presented in Table 2.

Table 2
Torsion angles (°) for the centroids of the clusters obtained by the Full Bayesian Analysis method applied to the Csp^3 eight-membered rings of Allen *et al.* (1996).

Cluster	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	Cluster type	Data
$\mu(1)$	−98.9	39.3	67.8	−63.2	−46.0	100.8	−66.0	66.7	BC	15
$\mu(2)$	−97.5	54.9	53.6	−82.4	−8.1	81.1	−71.3	66.8	BC/TBC	4
$\mu(3)$	−82.5	5.1	85.8	−53.9	−57.7	100.8	−69.6	72.3	BC/TBC	4
$\mu(4)$	73.5	−94.4	92.6	−70.2	69.9	−87.2	87.7	−68.7	–	–
$\mu(5)$	68.6	−101.0	87.5	−49.1	54.0	−93.7	89.1	−57.4	–	–
$\mu(6)$	1.7	61.9	101.8	−67.3	60.3	101.0	83.8	−5.8	–	–
$\mu(7)$	88.5	3.6	−74.5	89.8	−88.1	70.8	−4.9	−88.5	–	–

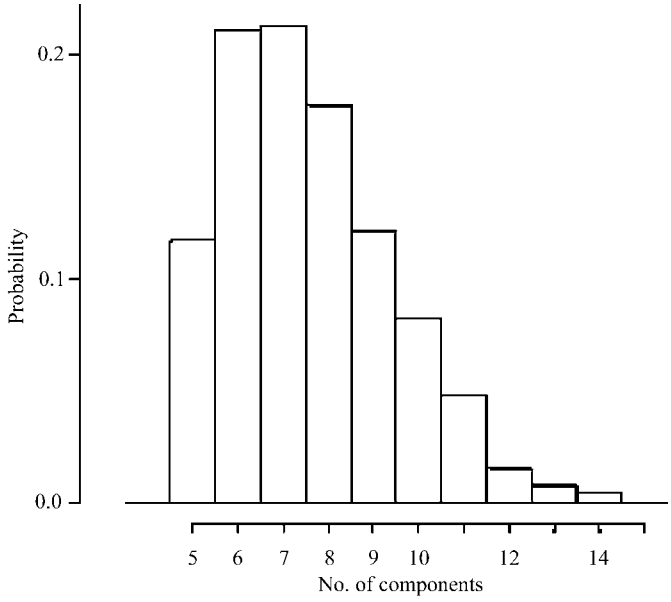

Figure 3
Probabilities for the number of conformations.

In Table 1 the last column indicates to which of the detected clusters is each individual structure in the dataset assigned. The detected clusters are interpreted as follows.

Cluster $\mu(1)$: This is the most populated cluster, it includes 15 data. The torsion angles of the centroids of the clusters (Table 2) are close to those of the boat–chair conformation. This cluster essentially agreres with the BC cluster reported by Allen and co-workers (Allen *et al.*, 1996). The structures correspond to free (non-fused) cyclooctane rings and it is the expected conformation according to energy (Hendrickson, 1967; Anet & Krane, 1973).

Cluster $\mu(2)$: As can be seen in Table 1 this cluster includes 4 data. The torsion angles in this structures are close to those of the ideal boat–chair (BC), but there are some differences: the ideal BC angles of 44.7 and −102° (see the supplementary material) show absolute value ranges of 19–30 and 88–94°; this means a flattening of the BC structure towards the twist-boat–chair (TBC) conformation. In addition, BC and TBC confor-

mations can interconvert by a pseudorotation pathway (Allen *et al.*, 1996).

Cluster μ(3): This cluster includes 4 data. When the torsion angles for the centroid of this cluster are analyzed by the Classification method significant probabilities for the BC and TBC conformations are obtained. In fact, this corresponds to the BC/TBC cluster proposed by Allen and co-workers (Allen *et al.*, 1996).

The remaining clusters are less populated (Table 1) and they correspond to distorted structures between the ideal conformations. The torsion angles of the centroids are shown in Table 2. Notice that the distribution for the corresponding

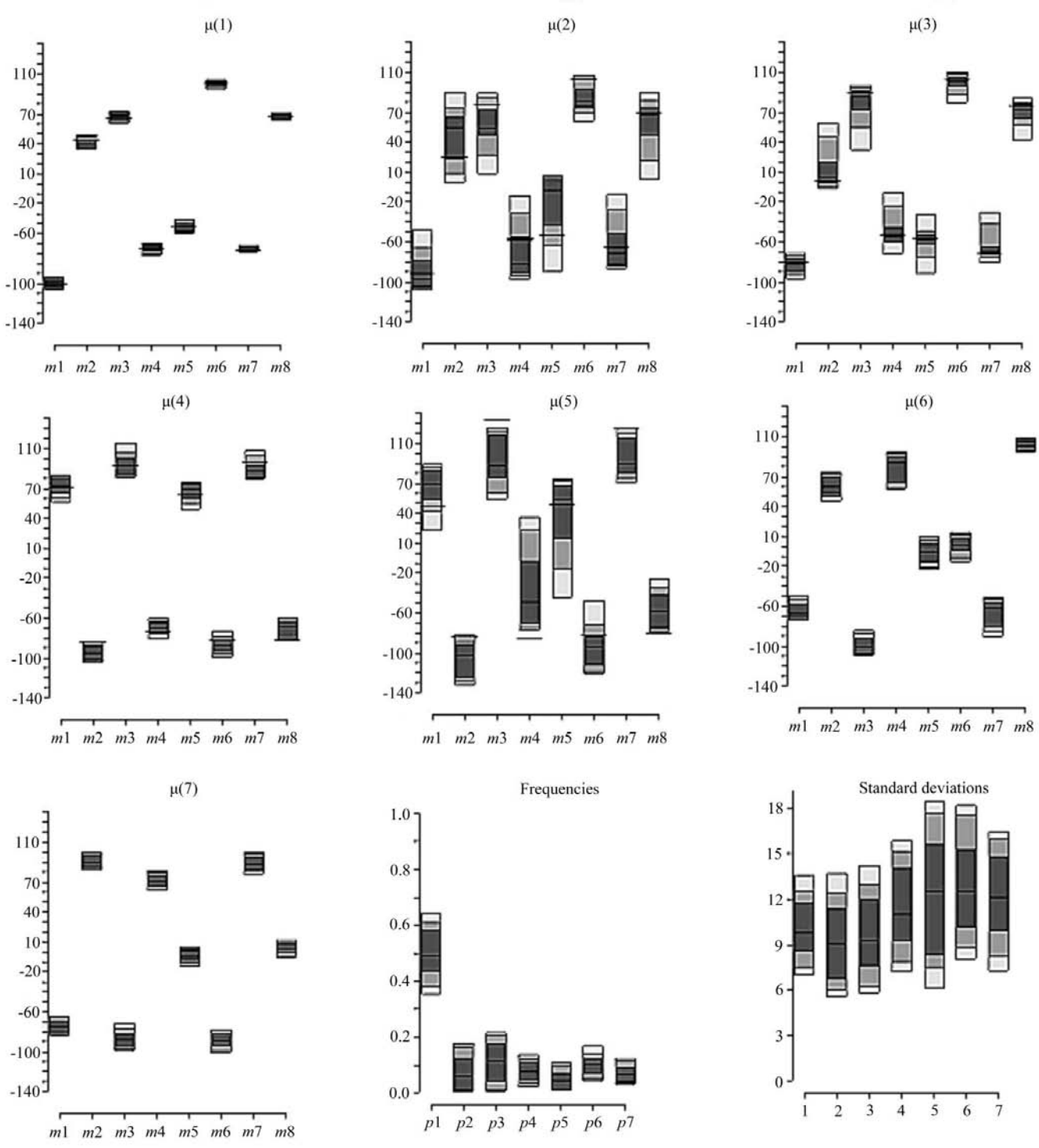

Figure 4
Box-plot type representations of the posterior distributions of the conformations detected, frequencies and standard deviations of Csp^3 eight-membered rings in Allen *et al.* (1996). Eight horizontal lines were drawn for each box, representing the 10, 15, 25, 50, 75, 85 and 90% percentiles of the distribution.

Table 3
Conformational classification of Csp^3 eight-membered rings in the CSD, Version 5.25, according to the methods described in this paper.

	Classification method		Full Bayesian method
	$\sigma = 10^\circ$	$\sigma = 20^\circ$	
BAJQIN	1.00 BC	0.90 BC; 0.10 TBC	$\mu(2)$
BAJQOT	1.00 BC	0.77 BC; 0.23 TBC	$\mu(2)$
BAJQUZ	1.00 BC	1.00 BC	$\mu(1)$
BEYPAW	1.00 BC	1.00 BC	$\mu(1)$
DUVFUV03 (1)	1.00 BC	1.00 BC	$\mu(1)$
DUVFUV03 (2)	1.00 BC	1.00 BC	$\mu(1)$
FADTAG	1.00 BC	1.00 BC	$\mu(1)$
GIPPAW	1.00 B	1.00 B	$\mu(6)$
GIQRUT	1.00 BC	1.00 BC	$\mu(1)$
HABXEN (1)	1.00 TCC	0.07 CR; 0.17 CC; 0.76 TCC	$\mu(4)$
HABXEN (2)	1.00 BC	1.00 BC	$\mu(1)$
HAVGOA	1.00 C	1.00 C	$\mu(5)$
HAVGUG	1.00 C	1.00 C	$\mu(5)$
HEMTIC (1)	1.00 BC	0.70 BC; 0.30 TBC	$\mu(2)$
HEMTIC (2)	1.00 BC	0.72 BC; 0.28 TBC	$\mu(2)$
HEMTIC (3)	0.93 BC; 0.07 TBC	0.53 BC; 0.47 TBC	$\mu(2)$
HEMTIC (4)	1.00 BC	0.81 BC; 0.19 TBC	$\mu(2)$
IFOKOD	1.00 BC	1.00 BC	$\mu(1)$
IGARUD	1.00 BC	0.98 BC; 0.02 TBC	$\mu(2)$
JAGQIR	0.08 CR; 0.12 CC; 0.80 TCC	0.47 CR; 0.22 CC; 0.31 TCC	$\mu(4)$
JITNIJ	1.00 C	1.00 C	$\mu(5)$
JIWWUH (1)	1.00 BC	1.00 BC	$\mu(1)$
JIWWUH (2)	1.00 BC	1.00 BC	$\mu(1)$
JOQMIL (1)	0.98 BC; 0.02 TBC	0.61 BC; 0.39 TBC	$\mu(2)$
JOQMIL (2)	1.00 BC	0.48 BC; 0.52 TBC	$\mu(2)$
JOQMIL01 (1)	1.00 BC	0.59 BC; 0.41 TBC	$\mu(2)$
JOQMIL01 (2)	0.85 BC; 0.15 TBC	0.48 BC; 0.52 TBC	$\mu(2)$
KIKWAC	1.00 C	1.00 C	$\mu(5)$
KOJDIW	0.75 BC; 0.25 TBC	0.44 BC; 0.56 TBC	$\mu(2)$
KOJDOC	1.00 BC	0.97 BC; 0.03 TBC	$\mu(2)$
KOJNUS	1.00 B	1.00 B	$\mu(6)$
KOJPAA (1)	1.00 C	1.00 C	$\mu(5)$
KOJPAA (2)	1.00 C	1.00 C	$\mu(5)$
MOZSAV	1.00 C	1.00 C	$\mu(5)$
NACGAZ	1.00 BC	1.00 BC	$\mu(1)$
NADCOK	1.00 S	0.20 BB; 0.78 S; 0.02 B	$\mu(6)$
NADNAH	0.48 BC; 0.52 TBC	0.37 BC; 0.63 TBC	$\mu(2)$
NUJHUV	1.00 BC	1.00 BC	$\mu(1)$
NUWMIB	1.00 BC	0.86 BC; 0.14 TBC	$\mu(2)$
NUZDER	1.00 BC	0.97 BC; 0.03 TBC	$\mu(2)$
PAPWAE (1)	1.00 C	1.00 C	$\mu(5)$
PAPWAE (2)	1.00 C	1.00 C	$\mu(5)$
PAZJEF	1.00 BC	1.00 BC	$\mu(1)$
PETFOJ	0.04 BC; 0.96 TBC	0.21 BC; 0.79 TBC	$\mu(3)$
PIVWEW	1.00 BC	1.00 BC	$\mu(1)$
POYJOC (1)	1.00 BC	1.00 BC	$\mu(1)$
POYJOC (2)	1.00 BC	1.00 BC	$\mu(1)$
QADNEP	1.00 BC	0.97 BC; 0.03 TBC	$\mu(2)$
QETVUG	1.00 TBC	1.00 TBC	$\mu(3)$
QOTGEL	1.00 BC	1.00 BC	$\mu(1)$
QOTGEL01	1.00 BC	0.99 BC; 0.01 TBC	$\mu(2)$
RECFOU	1.00 BC	1.00 BC	$\mu(1)$
RECPUK	1.00 BC	1.00 BC	$\mu(1)$
RILWIS (1)	1.00 BC	0.75 BC; 0.25 TBC	$\mu(2)$
RILWIS (2)	1.00 BC	0.72 BC; 0.28 TBC	$\mu(2)$
RIZCUY	0.96 BC; 0.04 TBC	0.56 BC; 0.44 TBC	$\mu(2)$
RULMAM	0.91 BC; 0.09 TBC	0.52 BC; 0.48 TBC	$\mu(2)$
SORDAE	1.00 C	1.00 C	$\mu(5)$
VIDNAX	1.00 C	1.00 C	$\mu(5)$
VIDNAX01 (1)	0.01 S; 0.99 B	0.14 S; 0.86 B	$\mu(6)$
VIDNAX01 (2)	0.27 S; 0.73 B	0.32 S; 0.68 B	$\mu(6)$
VIDNEB	1.00 C	1.00 C	$\mu(5)$
WAHRAY	1.00 B	1.00 B	$\mu(6)$
WIDSEH	0.02 BC; 0.98 TBC	0.19 BC; 0.81 TBC	$\mu(3)$
WIRPAO	1.00 BC	1.00 BC	$\mu(3)$
WOQKUI	1.00 BC	0.97 BC; 0.03 TBC	$\mu(2)$
XENREN	0.97 BC; 0.03 TBC	0.58 BC; 0.42 TBC	$\mu(2)$
XEPWUK	1.00 BC	1.00 BC	$\mu(1)$

standard deviations presents a higher dispersion (Fig. 4), which can be explained by the fact that only very few observations belong to these clusters.

6.2. Conformational analysis of Csp^3 eight-membered rings in the CSD

In order to complete the conformational analysis described above we have studied the Csp^3 eight-membered rings included in the CSD, Version 5.25. The refcodes of the structures and the results obtained by the Classification method and the full Bayesian Analysis method are shown in Table 3 (the results of the refcodes analyzed in §6.1 have been omitted for clarity).

The BC/TBC pseudorotation pathway is clearly visible from the results in Table 3 for the Classification method: 30 data show a probability of 1.00 for BC, 19 data exhibit a small distortion from the BC to the TBC conformation (a significant probability appears only when $\sigma = 20^\circ$), 12 data show a larger deviation from BC and two structures even show a probability of 1.00 for the TBC conformation.

An accessible deformation of the BC conformation towards the TBC conformation can also be inferred from the probabilities of the structures QOTGEL, HEMTIC or JOQMIL, where chemically indistinguishable fragments show different probabilities for BC and TBC conformations.

In Table 3, 12 data sets are assigned to the chair (C) conformation, but all of them have fused rings in the 1,2 and 5,6 positions, most being chemically very similar. There are also 3 data with a probability of 1 (using both $\sigma =$ 10 or 20°) for the boat (B) conformation having a variable number of fused rings to the Csp^3 eight-membered ring.

When data are analyzed by the Full Bayesian method the histogram of the posterior distribution for k indicates that seven clusters are found to be most likely. In Fig. 5, box-plot-like diagrams are presented for the

Table 3 (continued)

	Classification method		Full Bayesian method
	$\sigma = 10^\circ$	$\sigma = 20^\circ$	
XOLYEC	0.99 BC; 0.01 TBC	0.67 BC; 0.33 TBC	$\mu(2)$
XUDWEY	1.00 BC	1.00 BC	$\mu(1)$
XUDWIC	1.00 BC	1.00 BC	$\mu(1)$
XUDWOI	1.00 BC	1.00 BC	$\mu(1)$
XULROL	1.00 TBC	1.00 TBC	$\mu(5)$
YAFLAS	1.00 BC	1.00 BC	$\mu(1)$
YOPPEY (1)	1.00 BC	1.00 BC	$\mu(1)$
YOPPEY (2)	1.00 BC	1.00 BC	$\mu(1)$
YOPPIC (1)	1.00 BC	0.98 BC; 0.02 TBC	$\mu(2)$
YOPPIC (2)	0.79 BC; 0.21 BB	0.18 BC; 0.08 TBC; 0.60 BB; 0.15 S	$\mu(2)$
ZAVRET	1.00 BC	1.00 BC	$\mu(1)$
ZAYPEU	1.00 BC	0.95 BC; 0.05 TBC	$\mu(2)$
ZAYPIY	1.00 BC	0.98 BC; 0.02 TBC	$\mu(1)$
AHOQOD	1.00 BC	1.00 BC	$\mu(1)$
BEHNEI	1.00 BC	1.00 BC	$\mu(1)$
UMIJIJ	0.86 BC; 0.14 TBC	0.49 BC; 0.51 TBC	$\mu(3)$

Table 4
Torsion angles (°) for the centroids of the clusters obtained by the Full Bayesian Analysis method applied to Csp^3 eight-membered in the CSD, Version 5.25.

Cluster	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	Cluster type	Data
$\mu(1)$	−97.5	38.9	65.6	−63.4	−45.5	98.7	−66.6	67.8	BC	47
$\mu(2)$	−90.3	48.9	55.8	−81.0	−10.1	78.3	−72.6	69.1	BC/TBC	34
$\mu(3)$	−75.3	−0.5	89.3	−56.6	−49.6	83.8	−75.9	77.2	TBC/BC	7
$\mu(4)$	75.5	−67.0	97.3	−94.9	61.8	−56.7	84.6	100.0	CC/TCC	5
$\mu(5)$	115.9	−79.7	4.7	76.7	−120.0	67.8	5.2	−74.3	C	13
$\mu(6)$	−65.1	7.4	75.8	−11.7	−73.5	17.2	68.5	−11.1	B	6
$\mu(7)$	−96.3	64.8	−65.3	94.3	−75.1	2.3	3.1	74.7	–	–

posterior distributions of the conformational sequences of torsion angles detected, their corresponding frequencies of occurrence and the associated standard deviations. The torsion angles of the centroids are presented in Table 4. In clusters $\mu(1)$, $\mu(2)$ and $\mu(3)$ the torsion angles for the centroids are similar to those found in the three most populated clusters obtained in §6.1, corresponding to the BC or BC/TBC conformations.

The most significant differences with data analyzed in §6.1 are clusters $\mu(5)$ and $\mu(6)$. Cluster $\mu(5)$ includes structures having the chair (C) conformation. For $\mu(5)$ in Fig. 5 a wider horizontal black line is drawn which represents the torsion angles of the ideal chair conformation. According to the torsion angles of its centroid (Table 4) cluster $\mu(6)$ corresponds to structures having a boat (B) conformation, which is in agreement with the results obtained by the Classification method.

6.3. Features of the methods described and differences with PCA

When studying the conformations of ring systems using familiar intramolecular parameters such as torsion angles, the exploration and classification of the data are made difficult by the large number of variables for each observed structure: for an eight-membered ring, for example, the data observed lie in an eight-dimensional space, even if it is well known that there are only five degrees of conformational freedom. One way around the problem is to achieve a dimension reduction in order to be able to display the data in a two- or three-dimensional space and visually detect groups of molecules. This is the method used in the principal component analysis: for example, the direction given by the first principal component (pc1) is the direction along which the data present more variability and therefore the possible groups of structures can be distinguished.

The two methods described in this paper do not perform any kind of dimension reduction, but are designed to deal directly with grouping in multivariate data. The purpose and the output of the 'Classification method' and the 'Full Bayesian Analysis method' are different but, in our experience, both provide useful information for the conformational analysis of structures. We would like to emphasize the following facts:

(i) Both methods share the same underlying probabilistic model to explain the observed torsion angles. In particular, treatment of the symmetry of the parameter space, the enantiomers *etc.* is taken into account in the modelling step, which, in contrast to the pca method, allows the symmetry expansion of the initial set of torsion angles to be avoided.

(ii) The 'similarity' between two structures can be understood in terms of the underlying probabilistic model: two structures will be found to be close if their merging in the same group yields a high corresponding likelihood.

(iii) The 'Classification method' requires as input prior knowledge of the preferred conformations. This is, of course, a real restriction to its applicability, since for many practical cases no sound energy-based starting point may be available. However, we would like to stress its simplicity: it is a straightforward sub-product of the model formulation; its implementation only requires a few lines of code and it is able to perform the classification of an individual structure into one of any collection of preferred conformations of interest to the user. Moreover, as illustrated in the experimental study, significant classification probabilities for several preferred conformations give hints about the interconversion pathways that connect some of these conformations. This is also one of the acceptable by-products of the pca method.

(iv) The 'Full Bayesian Analysis' method detects groups without requiring either knowledge of the preferred confor-

mations or of their number. It consists of an implementation of the Bayesian paradigm for the case where the number of preferred conformations, the preferred conformations themselves, their relative frequencies of occurrence and the standard deviations of the perturbations are the unknown parameters of interest. Since the number of preferred conformations is itself a parameter to be inferred, the output of the method includes the posterior probability of each possible conformation. The implementation of the 'Full Bayesian Analysis' method was carried out using the Reversible-Jump Markov Chain Monte Carlo (MCMC) algorithm. We developed our code based on the available code as described in Cappé *et al.* (2003). As with any MCMC algorithm, it is on one hand computationally demanding and on the other hand requires tuning of the parameters to ensure the convergence of the algorithm. In that sense, some experience is needed to run the code.

(v) Finally, even if we have chosen to present the methods using eight-membered rings, they extend to *m*-membered rings in a straightforward manner. The only difference lies in

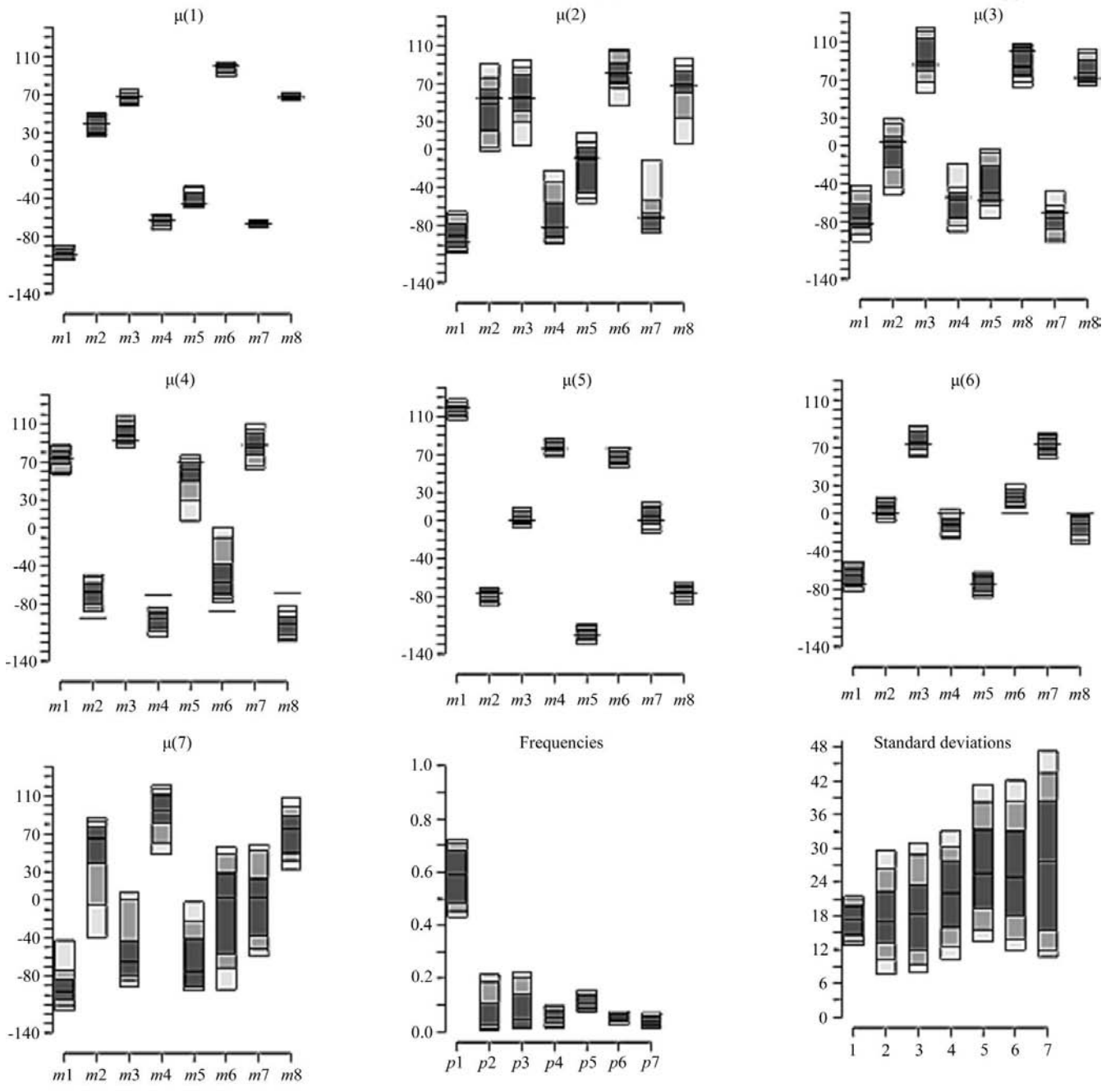

Figure 5
Posterior distribution of the detected conformations, frequencies and standard deviations of Csp^3 eight-membered rings in the CSD, Version 5.25. Eight horizontal lines were drawn for each box, representing the 10, 15, 25, 50, 75, 85 and 90% percentiles of the distribution.

the formulation of the model: the expression of the density functions $f(\tau,c)$ in §2.1 is now

$$f(\tau, c) = \sum_{\nu=1,\ldots,m} \sum_{d=-1,1} \sum_{\delta=-1,1} f_G\big(\tau, \mu(c, \nu, d, \delta), \sigma_c^2\big),$$

with $\boldsymbol{\tau} \rightarrow f_G\,(\boldsymbol{\tau}, \boldsymbol{\mu}(c, \nu, d, \delta), \sigma_c^2)$ denoting the density of the m-dimensional Gaussian law with the mean $\mu(c, \nu, d, \delta)$ given by

$$\delta\times(\mu C, \nu, \mu_{C,(\nu-1+d\times1)\bmod m+1}, \mu_{C,(\nu-1+d\times2)\bmod m+1}, \ldots, \mu_{C,(\nu-1+d\times(m-1))\bmod m+1}).$$

7. Conclusions

(i) The Classification method that uses the canonical conformations described in this article allows the closest ideal conformation for an individual eight-membered ring to be established using the eight torsion angles in the ring. Moreover, the method can easily be extended to rings with different numbers of atoms or when choosing different canonical conformations.

(ii) In the output of the Classification method for an individual structure, the appearance of non-zero probabilities for different theoretical conformations indicates that the structure is an intermediate between theoretical conformations. In addition, the relative value of the probability indicates the proximity to the ideal conformation.

(iii) It is convenient to allow large values for the deviations from the ideal torsion angles (*e.g.* $\sigma = 20^\circ$) in order to detect small deviations from the ideal conformations. We recommend checking the sensitivity of the results to different values of σ.

(iv) The Full Bayesian method does not assume any previous knowledge on the preferred conformations. It allows on one hand a decision about the most likely number of clusters and, on the other hand, provides details of the centroids of the clusters, their frequencies and the estimated standard deviations.

(v) The combined use of both methods draws significant chemical conclusions.

APPENDIX *A*

Torsion angles (°) for the canonical conformations of cyclooctane, *i.e.* the 'preferred' conformations for the Classification method are shown in the table below.

Conf.	C	$\mu_{C,1}$	$\mu_{C,2}$	$\mu_{C,3}$	$\mu_{C,4}$	$\mu_{C,5}$	$\mu_{C,6}$	$\mu_{C,7}$	$\mu_{C,8}$
BC	1	65.0	44.7	−102.2	65.0	−65.0	102.2	−44.7	−65.0
TBC	2	88.0	−93.2	51.9	44.8	−115.6	44.8	51.9	−93.2
CR	3	87.5	−87.5	87.5	−87.5	87.5	−87.5	87.5	−87.5
CC	4	66.0	−105.2	105.2	−66.0	66.0	−105.2	105.2	−66.0
TCC	5	56.2	−82.4	114.6	−82.4	56.2	−82.4	114.6	−82.4
BB	6	52.5	52.5	−52.5	−52.5	52.5	52.5	−52.5	−52.5
S	7	64.9	37.6	−64.9	−37.6	64.9	37.6	−64.9	−37.6
C	8	119.9	−76.2	0.0	76.2	−119.9	76.2	0.0	−76.2
B	9	−73.5	0.0	73.5	0.0	−73.5	0.0	73.5	0.0
TC	10	37.3	−109.3	109.3	−37.3	−37.3	109.3	−109.3	37.3

This work was supported in part by the European Community's Human Potential Programme under contract HPRN-CT-2000-00100, DYNSTOCH, and was developed using the computing facilities of the SAIT, Universidad Politécnica de Cartagena.

References

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
Allen, F. H., Howard, J. A. K. & Pitchford, N. A. (1996). *Acta Cryst.* B**52**, 882–891.
Allen, F. H. & Motherwell, W. D. S. (2002). *Acta Cryst.* B**58**, 407–422.
Allen, F. H. & Taylor, R. (2004). *Chem. Soc. Rev.* **33**, 463–475.
Anet, F. A. L. & Krane, J. (1973). *Tetrahedron Lett.* **50**, 5029–5032.
Cappé, O., Robert, C. P. & Rydén, T. (2003). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65**, 679–700.
Evans, D. G. & Boeyens, J. C. A. (1988). *Acta Cryst.* B**44**, 663–671.
Hendrickson, J. B. (1967). *J. Am. Chem. Soc.* **89**, 7047–7061.
Meyer, T. J. (1989). *Acc. Chem. Res.* **22**, 163.
Nolsøe, K., Kessler, M., Pérez, J. & Madsen, H. (2005). Technical report.
Orpen, A. G. (1993). *Chem. Soc. Rev.* pp. 191–197.
Richardson, S. & Green, P. J. (1997). *J. R. Stat. Soc. Ser. B*, **59**, 731–792.
Robert, C. (1996). *Interdisciplinary Statistics*, pp. 441–464. London: Chapman and Hall.
Robert, C. P. & Casella, C. (2005). *Monte Carlo Statistical Methods*, 2nd ed. Berlin: Springer Verlag.
Villar, J. M., Delgado, A., Llebaria, A. & Moreto, J. M. (1995). *Tetrahedron Asymm.* **6**, 665–668.
Zimmer, M. (2001). *Coord. Chem. Rev.* **212**, 133–163.