Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

Probabilistic model-based methodology for the conformational study of cyclic systems: application to copper complexes double-bridged by phosphate and related ligands

M. Kessler,[a]* J. Pérez,[b]* M. C. Bueso,[a] L. García,[b] E. Pérez,[b] J. L. Serrano[b] and R. Carrascosa[b]

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

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

Received 19 June 2007
Accepted 16 October 2007

A methodology for the conformational study of cyclic systems through the statistical analysis of torsion angles is presented. It relies on a combination of different methods based on a probabilistic model which takes into account the topological symmetry of the structures. This methodology is applied to copper complexes double-bridged by phosphate and related ligands. Structures from the Cambridge Structural Database (CSD) are analyzed and the *chair*, *boat-chair* and *boat* conformations are identified as the most frequent conformations. The output of the methodology also provides information about distortions from the ideal conformations, the most frequent being: *chair* ↔ *twist-chair*, *chair* ↔ *twist-boat-chair* and *boat* ↔ *twist-boat*. Molecular mechanics calculations identify these distortions as energetically accessible pathways.

1. Introduction

The conformational features of organic (Allen, 2002) and metallic complexes (Zimmer, 2001) have been extensively studied. The CSD has proved to be a useful tool in this kind of study (Allen & Taylor, 2004; Orpen, 1993). In particular, it is worth mentioning the work of Álvarez and co-workers in several conformational aspects of transition-metal complexes (Cirera *et al.*, 2004; Aullón *et al.*, 1999). The CSD has also been widely applied to life-science research (Taylor, 2002), where it has turned out to be a very significant tool for the rational design of biologically active molecules.

A review of different statistical methods for conformation analysis can be found in Zimmer (2001). These methods take a data-analysis approach where no model is assumed for the data-points generation mechanism, and all the conclusions rely on the correlation structure of the data-points or their similarities. Hierarchical cluster analysis and principal components analysis (PCA) are examples of such methods. Their application to crystallographic data is however made difficult by the existence of topological symmetry in the data-points, owing to the fact that there may be no unique, unambiguous atomic numbering that can be applied to a given fragment. To overcome this difficulty Allen and co-workers (Allen *et al.*, 1991*a*,*b*) suggest three agglomerative procedures for carrying out the conformational clustering of crystallographic data, which are modifications of the single-linkage, the complete-linkage and the Jarvis–Patrick algorithms. Continuous symmetry measures (Zabrodsky *et al.*, 1992, 1993) offer a different approach, which allows the quantitative evaluation of the deviation from any symmetry in a nonsymmetric configuration.

In a series of papers (Kessler *et al.*, 2007; Nolsøe *et al.*, 2005; Pérez, Nolsøe, Kessler, García, Pérez & Serrano, 2005; Pérez, García, Kessler, Nolsøe, Pérez, Serrano, Martínez & Carras-

cosa, 2005) the authors have developed different statistical methods that build upon a probabilistic model for the observed torsion angles: on one hand, in a study by Pérez, Nolsøe, Kessler, García, Pérez & Serrano (2005), a **classification method** and a **full Bayesian approach** to conformational classification were applied to cyclooctane structures. The mathematical details of the full Bayesian approach are described in Nolsøe *et al.* (2005). Its implementation is, however, not straightforward, in particular since it requires Monte Carlo simulations. On the other hand, an easier-to-use clustering procedure consisting of a **modification of recent model-based clustering techniques** is described in Kessler *et al.* (2007), together with a mathematically precise formulation of the topological symmetries of the data-points. In particular, this advanced clustering procedure was applied to a dataset consisting of 204 cyclooctane fragments and related compounds, and it was shown that the resulting conformational information is coherent with the expected preferences for this well studied system.

The main aim of this paper is to propose a methodology for the conformational analysis of rings which relies on a combination of the classification method and the model-based conformational clustering. We illustrate its practical implementation by performing the conformational classification of three datasets containing a total of 161 fragments, which correspond to copper complexes double-bridged by phosphate and related ligands retrieved from the CSD. This system is important in the study of metalloenzymes that catalyze the hydrolysis of phosphate ester bonds in biological molecules. A systematic variation of ligands and coordination environments in these biological analogues has provided information about the structure and function of enzymes. Although zinc is most commonly found in these enzymes, copper(II) complexes have contributed to the general understanding of these processes (Kövári & Krämer, 1996; Fry *et al.*, 2003). Moreover, these complexes exhibit a topological symmetry which makes the conformational analysis more challenging.

In order to confirm the conformational results provided by the statistical methods, molecular mechanics calculations were carried out and are presented in the paper. Notice that the statistical methodology is implemented as a *JAVA* program called *RingConf*, which can be freely downloaded from http://www.dmae.upct.es/ringconf, together with the datasets used in the present work, therefore allowing the reader to reproduce our analysis.

The described methodology allows:

(i) The **model-based conformational clustering** of the fragments, combining a hierarchical model-based algorithm and an expectation–maximization (EM) algorithm to approximate the maximum-likelihood estimators of the parameters of interest (centroids, proportions, standard deviations). As a by-product of the EM algorithm, the Bayesian Information Criterion (BIC) is computed to guide the decision about the number of clusters present in the dataset.

(ii) The **automatic expansion of canonical conformations**: starting from the canonical conformations known when all atoms are the same in the ring, the software performs the expansion and describes the resulting canonical conformations in the case of different atoms. As an example, the ten canonical conformations of cyclooctane expand to 21 if rings corresponding to metal complexes double-bridged by phosphate ligands are considered.

(iii) The **classification of individual structures** with respect to user-defined preferred conformations. These are generally the (expanded or not) canonical conformations. This classification is particularly applied to the centroids of the clusters resulting from the model-based conformational clustering method in order to identify each cluster's conformation type.

We emphasize the fact that the canonical conformations are only used for classification purposes, but are not required to perform the model-based conformational clustering method.

2. The methodology

2.1. Brief mathematical description of the methods

Consider an m-membered ring and denote the associated sequence of torsion angles by $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_m)$. In Kessler *et al.* (2007) an operator $T^{s,d,\delta}$ is introduced which transforms $\boldsymbol{\tau}$, by choosing the starting point s in the sequence, the direction d ($d = 1$ corresponds to reading through $\boldsymbol{\tau}$ clockwise, while $d = -1$ corresponds to reading through $\boldsymbol{\tau}$ counterclockwise), and changing the sign of $\boldsymbol{\tau}$ by multiplying by $\delta = \pm 1$. The symmetries in the data-points are taken into account by identifying a subset S of the set of all pairs (s,d), $s = 1, \ldots, m$ and $d = \pm 1$, for a given sequence of atom symbols, which leads to two sequences of torsion angles $\boldsymbol{\tau}^{(1)}$ and $\boldsymbol{\tau}^{(2)}$ being considered to be equivalent if a pair (s,d) exists in S and $\delta = \pm 1$ such that $T^{s,d,\delta}\,\boldsymbol{\tau}^{(1)} = \boldsymbol{\tau}^{(2)}$. For more details about set S, see Kessler *et al.* (2007). When, for example, the ring members are all C atoms set S contains all the pairs (s,d), with $s = 1, \ldots, m$ and $d = \pm 1$. As an example, for cyclooctane fragments the torsion angles sequences

$$\boldsymbol{\tau}(1) = (88.0, -93.2, 51.9, 44.8, -115.6, 44.8, 51.9, -93.2)$$

and

$$\boldsymbol{\tau}(2) = (51.9, -93.2, 88.0, -93.2, 51.9, 44.8, -115.6, 44.8)$$

are to be considered as equivalent since $T^{3,-1,1}\,\boldsymbol{\tau}^{(1)} = \boldsymbol{\tau}^{(2)}$.

The probabilistic model consists of a mixture of an unknown number of Gaussian multivariate distributions, with diagonal covariance matrices, but with possibly different standard deviations. The mean parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots \boldsymbol{\mu}_G$ of the Gaussian distributions represent the cluster centroids, while the standard deviations $\sigma_1, \sigma_2, \ldots \sigma_G$ are related to the cluster diameters. Moreover, we weight the different Gaussian distributions through the proportions $p_1, p_2, \ldots, p_G$. Finally, the topological symmetries of the data-points are incorporated in the model formulation by considering all equivalent versions of the sequence of torsion angles. To sum up, the probabilistic density associated with the data-points is proportional to

$$f(\boldsymbol{\tau}) = \sum_{k=1\ldots G} p_k \sum_{(s,d)\,\mathrm{in}\,S} \sum_{\delta=-1,1} f_{\mathrm{G}}(T^{s,d,\delta}\boldsymbol{\tau},\ \boldsymbol{\mu},\ \sigma_k^2), \qquad (1)$$

$\boldsymbol{\tau} \rightarrow f_{\mathrm{G}}\ (\boldsymbol{\tau}, \boldsymbol{\mu}, \sigma^2)$ denoting the density of the m-dimensional Gaussian law with mean $\boldsymbol{\mu}$ and diagonal covariance matrix $\sigma_c^2 Id$.

Three statistical methods that build upon the probabilistic model described above are used in this paper.

2.1.1. The classification method. As described in the paper by Pérez, García, Kessler, Nolsøe, Pérez, Serrano, Martínez & Carrascosa (2005), if the user specifies *the 'preferred' conformations* by providing the values of the parameters $\boldsymbol{\mu}_k$ (typically the canonical conformations) and sets the value of σ_k and p_k, it is possible to perform an individual classification of the given structure. Based on the m values of the torsion angles for a structure, we are able, through the Bayes rule, to compute the posterior probability that the structure comes from each of the preferred conformations.

2.1.2. The model-based agglomerative clustering method. An agglomerative hierarchical clustering algorithm is constructed as a stepwise procedure in which 'optimal' pairs of clusters are successively merged. Such an algorithm usually starts with as many clusters as observations and ends when all observations are in a single cluster. Examples of such hierarchical procedures modified for conformational analysis are given in the series of papers by Allen and co-workers (Allen *et al.*, 1991*a*,*b*). In these papers different methods are suggested, relying on different criteria to choose, at a given stage, which clusters should merge. In model-based clustering, the criterion for merging two clusters corresponds to a maximum-likelihood criterion: the dissimilarity between two clusters is computed using the classification likelihood; specifically two clusters are merged if their merging leads to the larger increment of the likelihood, computed on the basis of model (1). The mathematical details and the inclusion of the topological symmetries in the method can be found in the paper by Kessler *et al.* (2007).

2.1.3. The EM algorithm method. In the paper by Kessler *et al.* (2007) a modification of the mclust procedure suggested by Raftery and co-authors (see *e.g.* Fraley & Raftery, 2002) is described. For a given partition resulting from the model-based agglomerative clustering method, the Expectation Maximization (EM) algorithm tries to improve the estimation of the parameters by computing a numerical approximation of the maximum likelihood estimators. Using these numerical approximations the two different partitions can be compared (*e.g.* a five-clusters partition and a six-clusters partition) by computing the Bayesian Information Criterion (BIC) in order to influence the decision regarding the number of clusters present in the dataset. Notice that the EM algorithm method is expected to perform well only when a sufficient number of data-points is available.

2.2. Practical implementation of the methodology: general procedure

The procedure followed in this paper to analyze the datasets essentially consists of three steps:

(i) **First step**: the classification method is performed on the dataset to provide a first insight into the assignment of individual structures to the user-defined *type* conformations (for example, the canonical conformations). In the application of the classification method, one of the parameters to be fixed by the user to compute the assignment probabilities is the common value σ of the standard deviations $\sigma_1, \sigma_2, \ldots \sigma_G$ in model (1). Since model (1) describes a mixture of Gaussian densities with centers given by the user-defined type conformations and standard deviation σ, the assignment probabilities measure the associated probabilistic distance of the considered fragment to the different centers. If the value of σ is higher, the fragment will be 'closer' to more candidate centers and the assignment probabilities are expected to be significant for more type conformations. We usually perform the classification method twice: once using $\sigma = 10^\circ$ and a second time using $\sigma = 20^\circ$, checking whether the assignment probabilities change between the runs. The cases when individual structures are assigned to more than one conformation with significant probabilities provide information about distortions from *type* conformations. The classification of individual structures can also be used to have a first idea about the approximate number of groups present in the dataset.

(ii) **Second step**: The model-based agglomerative clustering method and the EM algorithm are applied: we have to decide about a possible range of values for the number of groups present in the dataset. This is a delicate issue common to all clustering procedures. A first graphical indicator corresponds to the plot of dissimilarity *versus* step number (also known as the merging-cost plot) in the agglomerative clustering method. This is the simple indicator used in particular in the paper by Allen *et al.* (1991*a*,*b*). One of the advantages of formulating a probabilistic model is that, as a by-product of the EM algorithm, we can compute an additional indicator: the Bayesian Information Criterion (BIC), a popular quantity in model selection issues. For a review on BIC and model selection in cluster analysis, see the paper by Fraley & Raftery (2002). For a given partition of the dataset into G groups, we compute the BIC value, which we denote by BIC(G), on the basis of model (1) and using the estimators obtained as a result of the EM algorithm. The difference between the BIC values for a G-groups partition and for a $(G-1)$-groups partition, *i.e.* BIC(G) − B($G-1$), approximates twice the logarithm of the Bayes factor. It is also interpreted as a summary of the evidence provided by the data in favor of the partition into G groups as opposed to the partition into $(G-1)$ groups. The higher the difference BIC(G) − BIC($G-1$), the stronger the evidence in the data in favor of the G groups partition. We therefore plot BIC(G) *versus* the number G of clusters in the partition. We then search ideally for the value of G for which BIC(G) is maximum. Unfortunately, in real data examples we have found that BIC(G) simply increases as G increases, tending to favor more and more groups in the dataset. We therefore check for significant increments in the BIC values and search for changes of slope in the BIC evolution as an indicator of the optimal stop point in the clustering process. This detection can be made easier through a plot of the BIC

successive differences: $\mathrm{BIC}(G) - \mathrm{BIC}(G-1)$ *versus* the number G of groups.

(iii) **Third step**: Once we have decided upon a range of values for the number of clusters in the dataset, we study in detail each corresponding partition. For any of these partitions, *RingConf* provides estimates of the proportion, *i.e.* the fraction of assigned structures, the centroid and the standard deviation corresponding to each cluster of the partition. Notice in particular that the estimates of the standard deviation can be used to assess the homogeneity of the structures in each cluster. Moreover, assignment of the individual fragments to the clusters is given. Finally, it is also possible to perform a classification of the clusters' centroids with respect to user-supplied canonical conformations.

Careful scanning of these outputs together with a chemical interpretation of the formed groups allows the user to extract useful conformational information from the data.

Remark: Notice that one of the clusters might present a high standard deviation which does not reduce significantly as the number of clusters increases. This indicates fragments which may be of the same conformational type, but differ significantly from one another. This was the case for example for the dataset 8C1 of Allen *et al.* (1996, §4.4, p. 888), where seven fragments out of 32 were left apart in the conformational analysis.

3. Experimental

3.1. Structural analysis

The Cambridge Structural Database (CSD, Version 5.28; Allen, 2002) was searched for all the structures containing the fragment shown in Fig. 1 when M = Cu. The search was modified so that only structures which present no errors have a crystallographic $R \leq 0.12$ and are not disordered were retrieved. A total of 91 REFCODES matched the search; the total number of fragments being 161. Distances, bond angles and torsion angles were measured by *CONQUEST*1.8 (Allen, 2002) and transferred to an ASCII text file for statistical and graphical analysis.

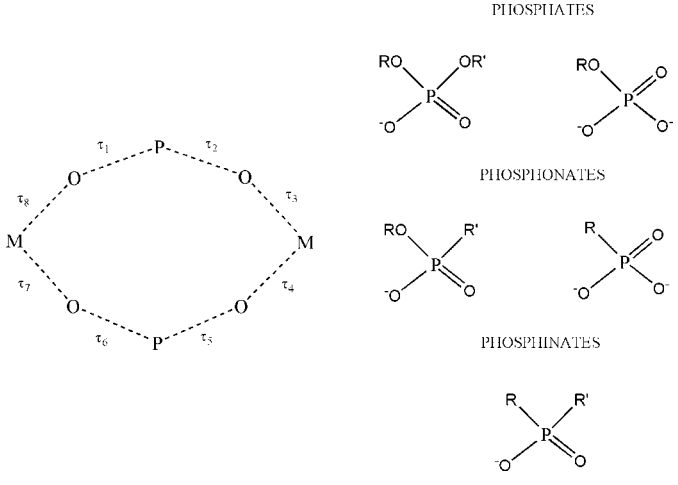


Figure 1
Substructure used in the conformational study. In the CSD search M = Cu was employed.

Out of these 91 structures, 35 structures were found to incorporate several types of phosphate moieties (hereafter referred to as phosphates), 36 bis(phosphonate) or bis(phosphonate ester) complexes (hereafter referred to as phosphonates), and 20 bis(phosphinate) complexes (hereafter referred to as phosphinates). Fig. 1 shows the type of ligands included in these structures. Nearly all the complexes contain Cu^{II}: 35, 29 and 15 structures are phosphates, phosphonates and phosphinates, respectively.

3.2. Molecular mechanics

The calculations were performed with the strain minimization program *MOMEC*97, Version 2.1.3 (Comba *et al.*, 1997). The expressions used to compute the potential energy are given as supporting information[1].

The points-on-a-sphere approach (Hambley *et al.*, 1981) was used, which implies that the O—Cu—O terms are replaced by O···O van der Waals terms. The torsion term O—Cu—O—P was omitted following a general practice in molecular mechanics calculations (Comba, 2001).

In the optimization parameters the r.m.s. (root-mean square) deviation between two structural models was calculated as the square-root of the sum of the squares of the deviations between all n corresponding pairs of atoms divided by the square-root of n. Specifically n was set to 8, since we considered all the atoms in the eight-membered ring (Cu, O and P atoms).

Several algorithms and programs have been suggested in the literature for parameterization (Hunger & Huttner, 1999; Comba & Remenyi, 2003). In our case parameters were scanned systematically around the initial values. First of all, the stretching parameters were optimized and then the bending parameters were corrected and a finer tuning of the stretching parameters was carried out again. Finally, the torsion angle term was optimized. This process was repeated until no significant improvement was obtained.

In the energy calculations along interconversion pathways, constraints were applied to torsion angles and the energy was minimized with *MOMEC*97 by a full-matrix Newton–Raphson algorithm. The selected termination criteria for Cartesian coordinates (r.m.s.) was < 0.001 Å.

4. Results and discussion

4.1. Preliminary theoretical aspects

Ten symmetrical conformations have been established for eight-carbon cyclic fragments: *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). Obviously, the introduction of different types of atoms in the ring decreases the symmetry. We study the core

[1] Supplementary data for this paper are available from the IUCr electronic archives (Reference: BS5051). Services for accessing these data are described at the back of the journal.

shown in Fig. 1 that shows the topological symmetry D_{2h}; notice that a ring built only of one type of atom exhibits the topological symmetry D_{8h}.

A corresponding decrease in the symmetry of the point group for the ten conformations described above is also found. At the same time, new conformations appear because of the lower topological symmetry: the sequence of torsion angles may remain unchanged but the associated sequence of atoms is different. As a consequence, a total of 21 possible conformations appear when the core shown in Fig. 1 is studied, referring to the ten canonical conformations of cyclooctane. Fig. 2 shows these 21 conformations with their corresponding point-group symmetry. In order to obtain a more general analysis, bond distances and angles were not considered in the derivation of these *type* conformations. The *JAVA* program *RingConf* can be used to automatically expand the ten canonical conformations of cyclooctane to the 21 conformations of Fig. 2.

Regarding mechanistic aspects, Hendrickson described three accessible paths for the interconversion of conformations in cyclooctane: in the *symmetrical mode* a plane or axis in the ring remains unchanged (applies to CR ↔ CC ↔ BC, CR ↔ TCC and BC ↔ BB conversions); in the *pseudorotation mode* a continuous change of dihedral angles takes place (applies to BB ↔ TB ↔ B, TC ↔ C, CC ↔ TCC and BC ↔ TBC conversions) and in the *asymmetric mode* only one side of a symmetrical ring is allowed to change as in a wagging of a single atom, which in general will invert the signs of the two dihedral angles directly adjacent to it (it applies to the BC ↔ TC conversion).

The above mechanisms can be applied to the 21 possible conformations in the system analyzed in this paper. It is thus possible to establish accessible paths for the interconversion of conformations in the complexes studied (Fig. 3). By a *symmetrical mode* the conversions $CR_1 \leftrightarrow CC_n$ (n = 1, 2, 3), $BB_2 \leftrightarrow BC_n$ (n = 1, 2) *etc.* are possible. In Fig. 3 the symbol *s* is used for interconversions by a *symmetrical mode*. A *pseu-*

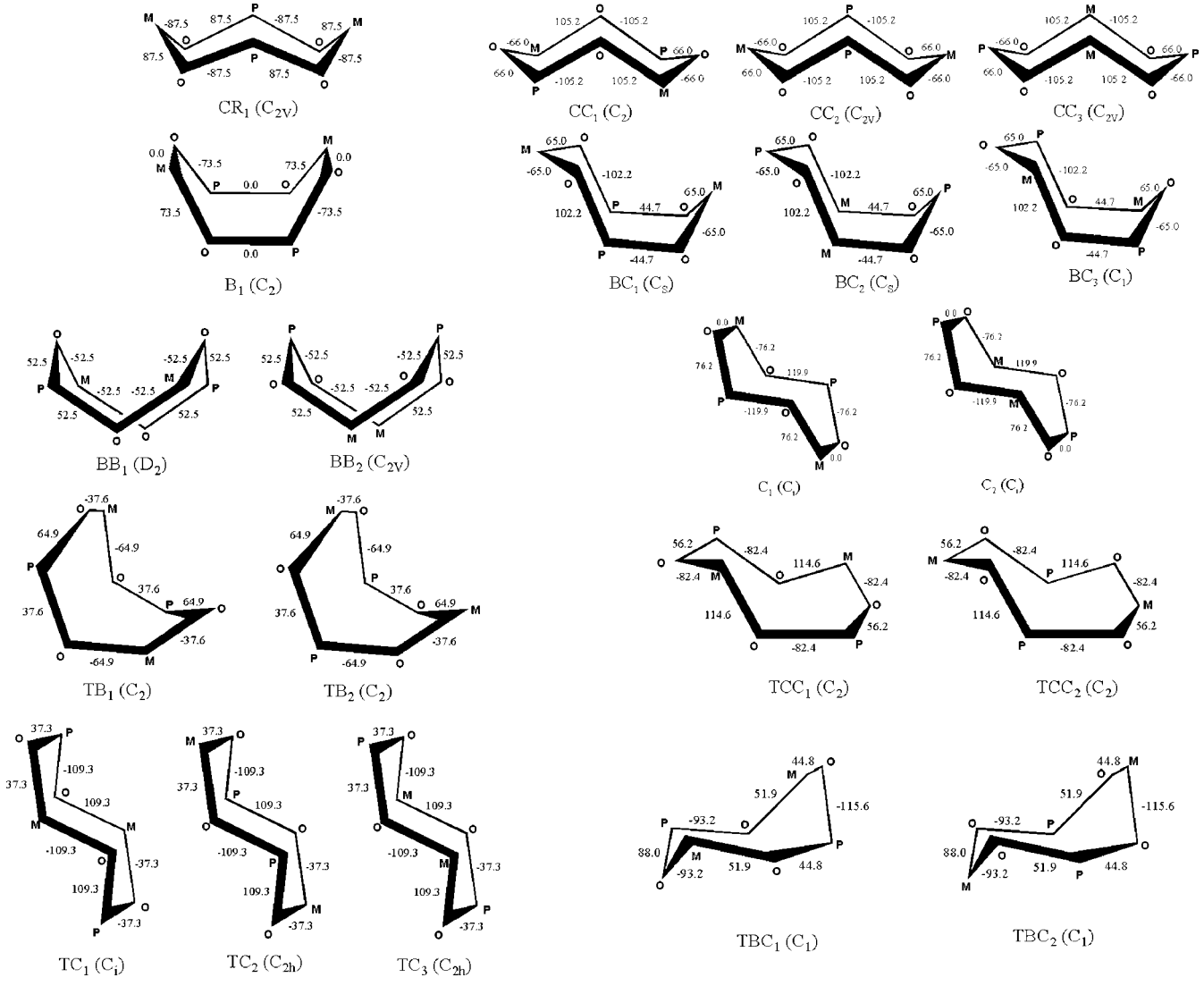


Figure 2
Conformation type for {Cu(μ-OPO)}$_2$.

dorotation mode is also possible, for example $TC_2 \leftrightarrow C_1 \leftrightarrow TC_1 \leftrightarrow C_2 \leftrightarrow TC_3$ or $BB_1 \leftrightarrow TB_1 \leftrightarrow B_1 \leftrightarrow TB_2 \leftrightarrow BB_2$ conversions (ψ symbol in Fig. 3). Finally, by an *asymmetric mode* the conversions $TC_1 \leftrightarrow BC_n$ (n = 1, 2) and $BC_3 \leftrightarrow TC_n$ (n = 2, 3) are possible (represented by *a* in Fig. 3).

4.2. Conformational analysis of copper complexes double-bridged by phosphate ligands

First step: the classification method showed the most frequent conformations in these compounds to be C_1, B_1 and BC_1. Fig. 4 shows the output of *RingConf*, consisting of a summary histogram of the frequencies of occurrence of each canonical conformation, deduced from the assignment of each fragment to its most likely canonical conformation. An individual checking of the posterior probabilities assigned to each canonical conformation for a given fragment is also possible. In particular, the appearance of a significant probability for more than one ideal conformation can be used to identify a structure as intermediate between two or more theoretical conformations (Pérez, García, Kessler, Nolsøe, Pérez, Serrano, Martínez & Carrascosa, 2005). When setting $\sigma = 20^\circ$ in the classification method, the most frequent conformation (C_1) turned out to show three kinds of distortions: towards TBC_1, TC_1 or TC_2 conformations. A table containing the probabilities for the most likely conformations (σ = 10 and 20°) for all the analyzed phosphate complexes (35 structures, 90 fragments) is provided as supporting information. As an example of the observed distortion towards TBC_1, we can mention the complex bis(μ_2-cytidine-5′-phosphato-*O*,*O*′)-(2,2′-dipyridylamine-*N*,*N*′)-aqua-copper(II) (Aoki, 1979; refcode CUCMPA), for which the classification method using $\sigma = 20^\circ$ yielded probabilities of 0.999 and 0.001 for C_1 and TBC_1 conformations, respectively. The distortion towards the TC_1 conformation can be illustrated with bis(μ_3-adenosinetriphosphate)tetrakis(2,2′-bipyridyl)tetracopper(II) (Kato & Tanase, 2005; refcode FEYBER): the probabilities obtained were 0.969 (C_1) and 0.031 (TC_1). Finally, an example of distortion towards the TC_2 conformation is bis(μ_2-guanosine-3′-monophosphate)-bis(aqua-1,10-phenanthroline)-copper heptahydrate (Wei *et al.*, 1978), refcode GMPCUP. In all the above examples when a deviation of $\sigma = 10^\circ$ is used, a probability of 1.000 is obtained for the C_1 conformation. As can be seen in Fig. 3 distortions towards TC_1 or TC_2 can be easily justified by a pseudorotation mechanism. Distortion from C_1 towards TBC_1 will be analyzed in the molecular mechanics section below.

Second step: We applied the hierarchical agglomerative classification and then computed the maximum likelihood estimators using the EM algorithm when the dataset is assumed to be partitioned in k groups, k ranging from 1–15 groups. For each of these possible sizes of the partition, the BIC criterion was computed. In Fig. 5 the evolution of the BIC and the corresponding successive differences are plotted against the considered number of groups (output of *RingConf*). We detect a change point for G = 6 and for G = 9 or 10, which indicates a change in the slope of the BIC values. These change points are confirmed by a plot of the dissimilarity *versus* the step number in the agglomerative clustering process. We therefore decide to explore a classification of the fragments into 5–8 or 9 groups.

Third step: Careful scanning of the partitions. Consider a partition of the dataset into five clusters: the most populated cluster (29 data-points) shows a low standard deviation (11.56) and the assignment of the fragments to this cluster as given by *RingConf* confirms they were identified as C_1 by the classification method. There are two small clusters (six data-points each) with a very low standard deviation (2.0) with CC_2/TCC_2 and BC_1 structures, respectively. The fourth cluster contains 21

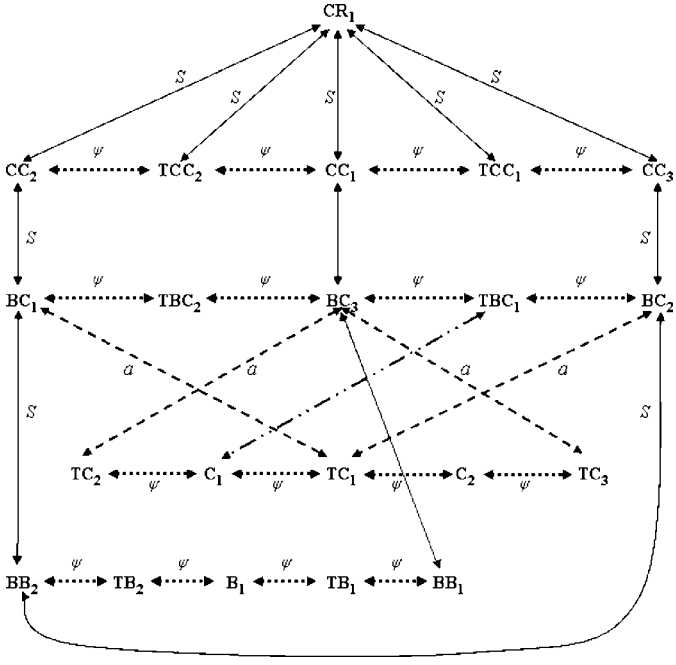


Figure 3
Interconversions of the {Cu(μ-OPO)}$_2$ core.

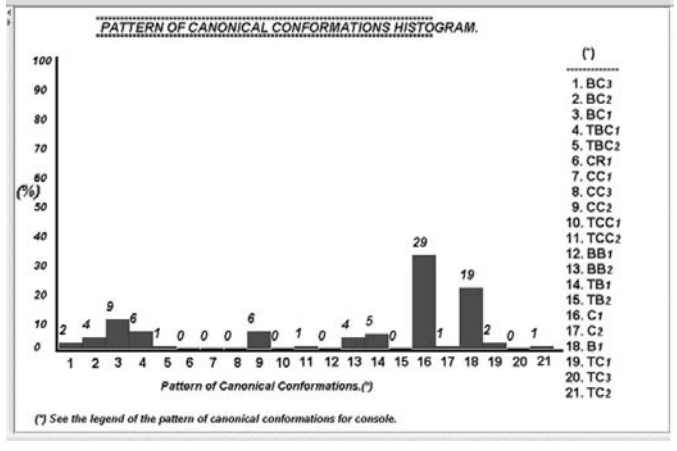


Figure 4
Frequencies of the most likely conformations in phosphate complexes obtained with the Classification method (RingConf output).

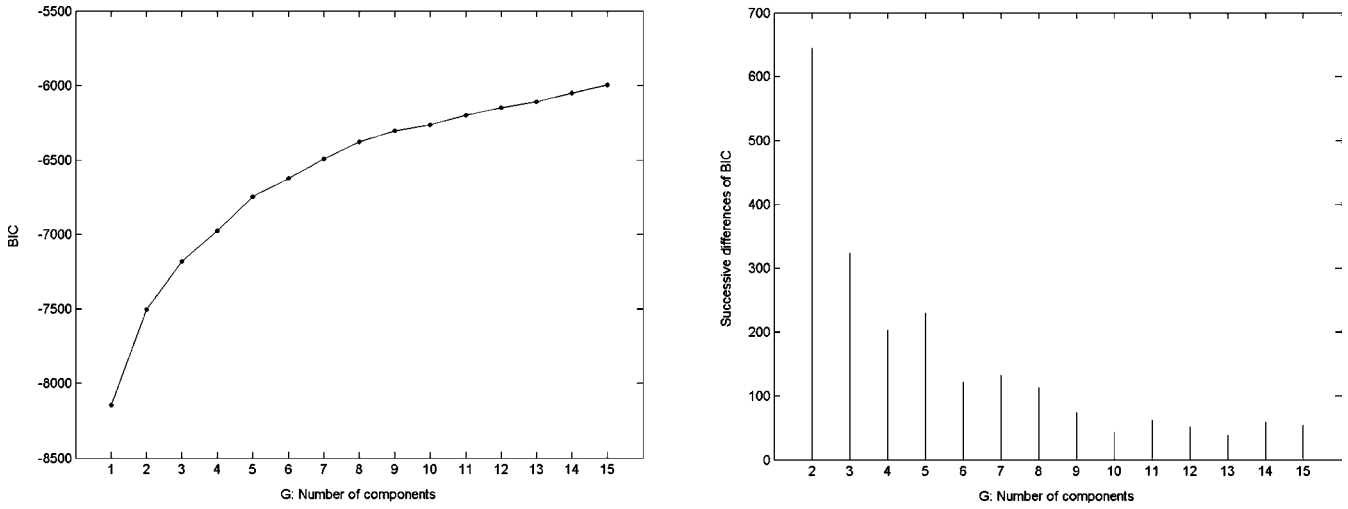


Figure 5
BIC and BIC successive differences for phosphate complexes (*RingConf* output).

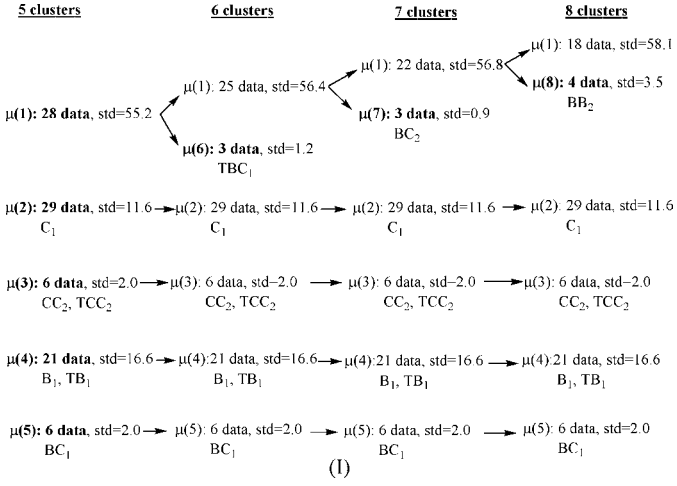


Figure 6
Phosphate structures: evolution of the partitions as the partition size grows.

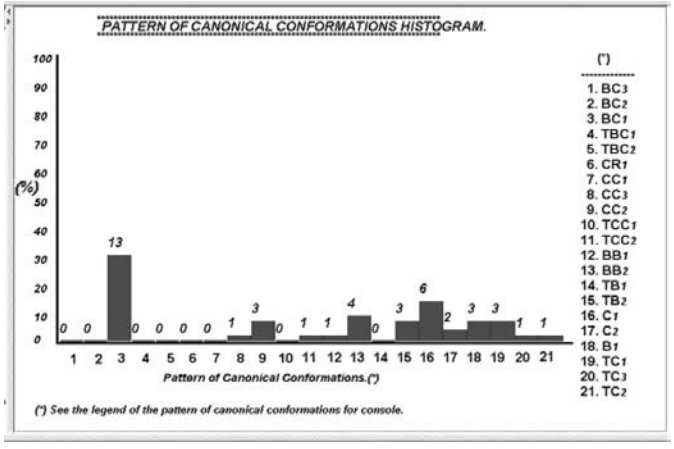


Figure 7
Frequencies of the most likely conformations in phosphonate complexes obtained with the Classification method (*RingConf* output).

data-points with a standard deviation of 16.6; it contains mostly B_1 structures and all the TB_1 ones (as can be seen in Fig. 3, B_1 and TB_1 structures are closely related by a pseudorotation interconversion mechanism). The low standard deviations for these clusters indicate that they consist of conformationally homogeneous fragments and they remain essentially unchanged when the chosen numbers of clusters varies from 5 to 8. The remaining cluster contains 28 data-points, and shows a high standard deviation of 55.24, generating small clusters with very low standard deviations with TBC_1, BC_2 or BB_2 conformations when the number of clusters is raised to 6, 7 or 8, respectively. A cluster with a high standard deviation remains which contains conformers which differ significantly from one another. These results are summarized in Fig. 6.

4.3. Conformational analysis of copper complexes double-bridged by phosphonate ligands

The classification method showed the most frequent conformation in these compounds to be BC_1 (Fig. 7). In all of them the P atoms appear to be connected by a bridging C atom, this P—C—P connection probably constrains the conformation of the eight-membered ring as reported in other conformational studies (Norenberg *et al.*, 1997). This is confirmed by the fact that for 11 data-points out of a total of 13 assigned to the BC1 conformation, the classification probability is 1.000 even when $\sigma = 20^\circ$ is used. An example of these complexes is *catena*-butane-1,4-diammonium bis(μ_4-1-hydroxyethylidenediphosphonato)tricopper(II) dihydrate (Zheng *et al.*, 1999), refcode COYLOR. The C_1 conformation (Fig. 7) is also present, see for example bis(μ_2-phosphonomethylcarboxylato-O,O′,O″)-bis(1,10-phenantroline)copper(II) (Zhao *et al.*, 2006), refcode TECNIZ. The probabilities for the most likely conformations (σ = 10 and 20°) for all the analyzed phosphonate complexes (35 structures, 42 fragments) are provided as supplementary information.

Table 1
Statistical analysis of selected bond parameters (Å, °).

Parameter	Mean value	Range
Cu—O	2.005	1.825–2.989
P—O	1.518	1.417–1.599
Cu···Cu	4.612	2.908–5.510
P···P	3.687	2.831–4.890
O···O†	3.757	3.030–4.759
O—Cu—O	93.140	78.40–154.46
Cu—O—P	129.497	74.51–163.65
O—P—O	113.455	107.60–125.84

† O atoms in relative positions 1,5.

When the hierarchical agglomerative classification and the EM algorithm are used, a four-group partition seems to be reasonable (the BIC values are provided as supplementary information): cluster 1 contains 11 data-points and shows a low standard deviation of 6.6, in fact it contains only structures which turn out, as confirmed by the classification method, to exhibit a classification probability of 1.000 for the BC_1 conformation, with either $\sigma = 10$ or 20°. There is also a cluster with 6 data-points and a low standard deviation of 8.6, which contains all the C_1-classified structures. An additional cluster with 4 data-points (standard deviation 8.6) is present. The corresponding structures are classified as TC_1 or C_2 conformations, which are related by a pseudorotation interconversion pathway. Finally, a cluster with 21 data-points and high standard deviation of 50.7 is found. This cluster contains fragments which differ significantly from one another and remain essentially unchanged when the number of clusters in the partition increases.

4.4. Conformational analysis of copper complexes double-bridged by phosphinate ligands

The most frequent conformation in these compounds is C_1 (Fig. 8), 13 data-points out of a total number of 29 are assigned to this conformation, see for example bis(μ_2-ammoniomethylphosphinato-O,O′)diaquatetrachlorodicopper(II) (Sawka-Dobrowolska & Głowiak, 1983), refcode BOXBIZ or bis(μ_2-hypophosphito)-bis(2,2′-bipyridine)copper(II) nitrate (Breneman *et al.*, 2002), refcode LOXSEW. Conformations TB_2 and TC_1 are also exhibited by a significant number of structures (Fig. 8). For this dataset, the probabilities for the most likely conformations ($\sigma = 10$ and 20°) corresponding to all the analyzed phosphinate complexes (20 structures, 29 fragments) can be found in the supporting information.

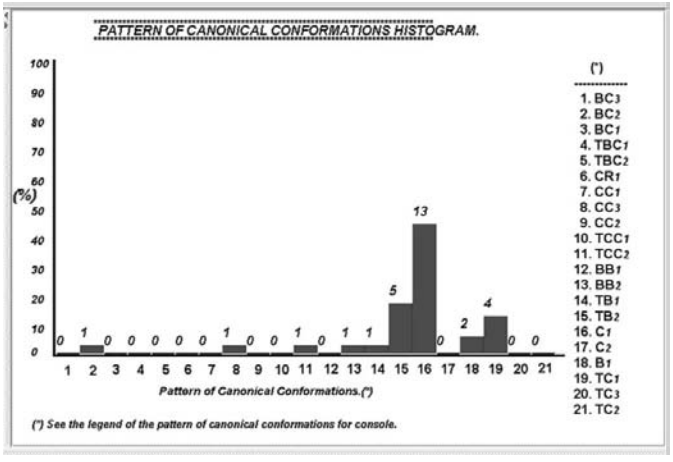


Figure 8
Frequencies of the most likely conformations in phosphinate complexes obtained with the Classification method (*RingConf* output).

The BIC values tend to indicate that partitions consisting of 3–7 clusters could be considered. (The BIC values are provided as supporting information.) When the data-points are grouped into three clusters by the hierarchical agglomerative classification and the EM algorithm, a first cluster of 13 data-points, all classified as C_1, is found. A second cluster which only contains two data-points (TC_1 conformation) is formed, and finally a cluster with 14 data-points and very high standard deviation contains all the remaining fragments. When four clusters are selected the C_1 cluster described above splits into two different clusters, the smaller one containing structures with two torsion angles close to zero.

When the partition size is five, six or seven the cluster with high standard deviation successively splits, separating TB_2, TCC_2 and TC_1 structures, respectively. The evolution of the partitions as the partition size grows can be found in the supporting information.

4.5. Geometrical features

In order to complete the conformational study and to optimize parameters for molecular mechanics calculations we performed the statistical analysis of some distances and angles, the results being summarized in Table 1. No significant influence of the kind of bridging ligand was observed. The range of Cu—O distances is wide; however, 75% of the values are lower than 2.000 Å, the highest values being related to a Jahn–Teller distortion. The P—O distances present a narrow range and similar values in phosphate, phosphonate and phosphinate complexes. The Cu···Cu distances present a large variation, which may be related to the fact that, as can be seen in Fig. 2, the Cu···Cu distance changes drastically according to the atom position on the ring. Finally, the range of P···P distances is large (Table 1), the lowest values corresponding to phosphonate structures with P—C—P bridges. Regarding the coordination number, penta-coordination is by far the most frequent in phosphates (30 from 35 structures). In phosphinates, values of four or five for the coordination number are the most frequent (11 and 6 cases, respectively, out of 20 phosphinate structures). Phosphonates show a wide variety of coordination numbers: 15 penta-coordinated complexes, four hexa-coordinated, three tetra-coordinated complexes and some with mixed coordination number (^{4}Cu/^{5}Cu, ^{4}Cu/^{6}Cu or ^{5}Cu/^{6}Cu) out of a total of 36 structures were identified.

The O—Cu—O, Cu—O—P and O—P—O angles were also analyzed; the mean values and ranges are shown in Table 1; no significant influence of the type of bridging ligand was observed.

Out of a total number of 91 structures, 11 structures were found to have ^{6}Cu. The corresponding double bridge involves an axial and an equatorial position only for five out of these 11 structures, presenting a difference in the Cu—O distance around 0.3 Å. This small number does not allow any correlation to be established between the Jahn–Teller effect and the conformation of the eight-membered ring. Moreover, these five structures were assigned to different conformations, namely TC_1, TB_2 or C_1. In tetrahedral complexes no significant differences in distances were observed.

4.6. Molecular mechanics calculations

It was established in §4.5 that no significant differences in bond distances and angles are observed in complexes of phosphates, phosphonates and phosphinates. Moreover, no correlation between the conformation of the eight-membered ring and the coordination number around copper was detected. For this reason and also for simplicity, a hypothetical fragment $\{Cu(\mu\text{-OPO})\}_2$, see Fig. 1 with M = Cu, was chosen to carry out the calculations. The use of such a hypothetical fragment has proved to be useful to study conformational changes in coordination compounds (Pérez *et al.*, 2004).

The use of tuned molecular mechanics parameters has proved to be crucial in the statistical studies of crystallographic data (Hocking & Hambley, 2002). In our case, *MOMEC*97 does not include suitable parameters for the fragment studied. It was therefore necessary to start the tuning procedure with initial parameters taken either from *MOMEC*97, from the literature or computed as an average obtained from selected crystal structures. These selected structures were chosen to fulfill the following conditions: representative structures with phosphates, phosphonates and phosphinate ligands, representative structures presenting different conformations in the eight-membered ring, no disordered structures, and Cu^{II} and R factor lower than 10%. In order to avoid the use of several parameters for the Cu—O stretching owing to the Jahn–Teller distortion, the selected structures were chosen to have a Cu—O distance lower than 2.40 Å. All the selected REFCODES are provided as supporting information. The agreement is between 0.167 and 0.382 Å; the r.m.s. values for all the structures are provided as supporting information, together with the optimized parameters. Notice that the selected set of structures turns out to be very heterogeneous, and presents significant differences, which may explain the high r.m.s. values.

In previous sections, the chair was established to be the most frequent conformation, which distorts towards the twist-chair conformations. This corresponds to the *pseudorotation* path $TC_2 \leftrightarrow C_1 \leftrightarrow TC_1 \leftrightarrow C_2 \leftrightarrow TC_3$ (see Fig. 3). The torsion angles of the fragment $\{Cu(\mu\text{-OPO})\}_2$ were set to vary along this path and the strain energy was minimized at each point; the results are shown in Fig. 9. As can be seen, the most significant contribution corresponds to the torsional component. The C_1 conformation is situated in a valley of the energy. The minimum energy is at the TC_3 conformation but, as can be seen in Fig. 2, this conformation would require a drastic steric accommodation of ligands around the Cu atoms.

The $C_1 \leftrightarrow TBC_1$ interconversion was also studied using the values shown in Fig. 10 for torsion angles. There is no significant variation in strain energy along this path. This is not a pseudorotation path, but it is an energetically very accessible distortion path, which can justify the high frequency of the C_1/TBC_1 distortion.

The $BB_2 \leftrightarrow TB_2 \leftrightarrow B_1 \leftrightarrow TB_1 \leftrightarrow BB_1$ pseudorotation path was also analyzed and the maximum energy was found for the B_1 conformation. The appearance of this conformation can be attributed to the linkage in the ring rather than to the strain energy.

We conclude that the results shown in Figs. 9 and 10 are in agreement with the structural results obtained by the statistical analysis: the C_1 conformation is an energetically favourable conformation and distortions towards twist-chair and

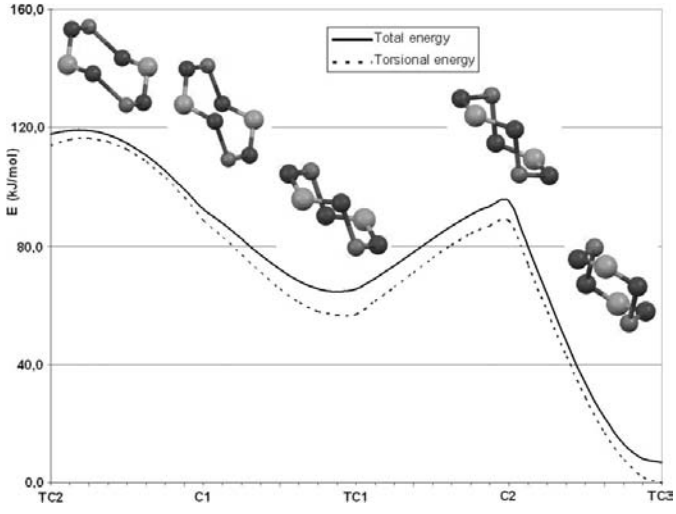
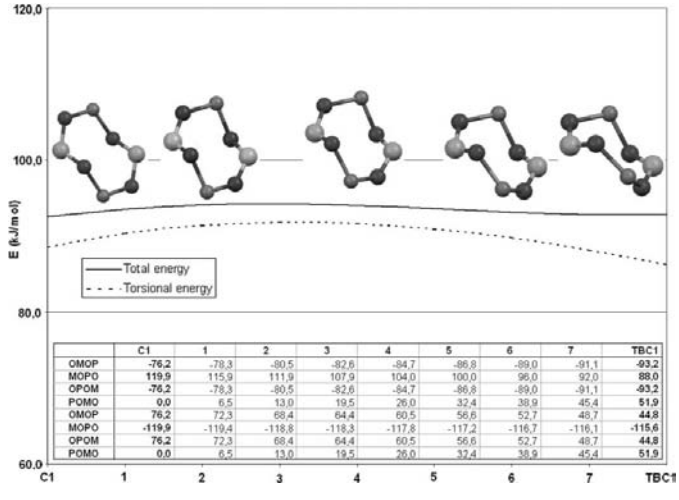


Figure 9
Strain energy for the C/TC *pseudorotation* path.



	C1	1	2	3	4	5	6	7	TBC1
OMOP	-76,2	-78,3	-80,5	-82,6	-84,7	-86,8	-89,0	-91,1	-93,2
MOPO	119,9	115,9	111,9	107,9	104,0	100,0	96,0	92,0	88,0
OPOM	-76,2	-78,3	-80,5	-82,6	-84,7	-86,8	-89,0	-91,1	-93,2
POMO	0,0	6,5	13,0	19,5	26,0	32,4	38,9	45,4	51,9
OMOP	76,2	72,3	68,4	64,4	60,5	56,6	52,7	48,7	44,8
MOPO	-119,9	-119,4	-118,8	-118,3	-117,8	-117,2	-116,7	-116,1	-115,6
OPOM	76,2	72,3	68,4	64,4	60,5	56,6	52,7	48,7	44,8
POMO	0,0	6,5	13,0	19,5	26,0	32,4	38,9	45,4	51,9

Figure 10
Strain energy for the C_1/TBC_1 path.

twist-boat-chair conformations are energetically accessible pathways.

5. Conclusions

The methodological framework we suggest in this paper consists of a combination of three statistical methods which, in contrast to previously reported statistical methods for conformational analysis, build upon a probabilistic model. We see several advantages in using these model-based methods:

(i) On one hand, the classification method is a mathematically simple by-product of the model formulation, and allows the classification of individual structures with respect to user-defined type conformations, providing the first insight into the conformational preferences of the dataset.

(ii) On the other hand, the agglomerative procedure we have developed uses the probabilistic model to compute the merging criterion. It relies on a modification of recent model-based clustering methods which have proved to perform well in a variety of clustering situations, avoiding for example the chaining problem known to be associated with simpler clustering methods.

(iii) The choice of the optimal stop point, as a consequence of the partition size, is a delicate issue common to all clustering methods. The advantage of basing our analysis on a probabilistic model is that we can compute an additional indicator to asses the number of groups in the dataset: the BIC quantity. However, we stress, following Allen and co-authors, that 'optimal clustering is an essentially subjective judgement to be made primarily on the grounds of chemical sensibility' (Allen *et al.*, 1991*b*, p. 48).

(iv) The model was formulated taking into account the topological symmetry of the data-points, thus avoiding their expansion, which is necessary for example in the PCA.

Moreover these methods, in contrast to the continuous symmetry measure approach, use a reduced number of structural parameters, namely eight torsion angles in the system we have analyzed.

We have illustrated the implementation of the methodology analyzing copper dinuclear complexes double bridged by phosphate and related ligands. The obtained results show that:

(i) For structures with no internal bridge in the eight-membered ring, the preferred conformations are C_1 and B_1, regardless of the type of bridge.

(ii) Crown, chair-chair and twist-chair-chair structures are very scarce.

Notice that the last two points may no longer be valid for rings with internal bridges.

(iii) The observed distortions basically consist of a twist from ideal conformations, *e.g.* C_1 distorts towards TBC_1, TC_1 or TC_2. The molecular mechanics calculations we have carried out confirm that these paths are energetically accessible.

As a conclusion, we think that we have illustrated in this paper how the combination of the three model-based methods, implemented in the easy-to-use *RingConf* software, may be a powerful tool for conformational study.

Financial support from the Comunidad Autónoma de la Región de Murcia (Project 03010/PI/05), Spain, is gratefully acknowledged. The authors thank an anonymous referee for detailed comments which have significantly improved the initial manuscript.

References

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
Allen, F. H., Doyle, M. J. & Taylor, R. (1991*a*). *Acta Cryst.* B**47**, 29–40.
Allen, F. H., Doyle, M. J. & Taylor, R. (1991*b*). *Acta Cryst.* B**47**, 41–49.
Allen, F. H., Howard, J. A. K. & Pitchford, N. A. (1996). *Acta Cryst.* B**52**, 882–891.
Allen, F. H. & Taylor, R. (2004). *Chem. Soc. Rev.* **33**, 463–475.
Aoki, K. (1979). *Chem. Commun.* pp. 589–591.
Aullón, G., Ujaque, G., Lledos, A. & Álvarez, S. (1999). *Chem. Eur. J.* **5**, 1391–1410.
Breneman, G. L., Fields, M. & Parker, O. J. (2002). *Acta Cryst.* E**58**, m262–m264.
Cirera, J., Alemany, P. & Álvarez, S. (2004). *Chem. Eur. J.* **10**, 190–207.
Comba, P. (2001). *Molecular Modeling of Inorganic Compounds*, 2nd ed. Wiley.
Comba, P., Hambley, T. W., Lauer, G. & Okon, N. (1997). *MOMEC.* Heidelberg, Germany.
Comba, P. & Remenyi, R. (2003). *Coord. Chem. Rev.* **238**, 9–20.
Fraley, C. & Raftery, A. E. (2002). *J. Am. Stat. Assoc.* **97**, 611–631.
Fry, F. H., Jensen, P., Kepert, C. M. & Spiccia, L. (2003). *Inorg. Chem.* **42**, 5637–5644.
Hambley, T. W., Hawkins, C. J., Palmer, J. A. & Snow, M. R. (1981). *Aust. J. Chem.* **34**, 45–56.
Hocking, R. K. & Hambley, T. W. (2002). *Inorg. Chem.* **41**, 2660–2666.
Hunger, J. & Huttner, G. (1999). *J. Comput. Chem.* **20**, 455–471.
Kato, M. & Tanase, T. (2005). *Inorg. Chem.* **44**, 8–10.
Kessler, M., Bueso, M. C. & Pérez, J. (2007). *J. Chemometr.* **21**, 53–64.
Kövári, E. & Krämer, R. (1996). *J. Am. Chem. Soc.* **118**, 12704–12709.
Nolsøe, K., Kessler, M., Pérez, J. & Madsen, H. (2005). *J. Chemometr.* **19**, 412–426.
Norenberg, K. M., Shoemaker, C. M. & Zimmer, M. (1997). *Dalton Trans.* pp. 1521–1526.
Orpen, A. G. (1993). *Chem. Soc. Rev.* **22**, 191–197.
Pérez, J., García, L., Kessler, M., Nolsøe, K., Pérez, E., Serrano, J. L., Martínez, J. F. & Carrascosa, R. (2005). *Inorg. Chim. Acta*, **358**, 2432–2436.
Pérez, J., Martínez, J. F., García, L., Pérez, E., Serrano, J. L. & Sánchez, G. (2004). *Inorg. Chim. Acta*, **357**, 3588–3594.
Pérez, J., Nolsøe, K., Kessler, M., García, L., Pérez, E. & Serrano, J. L. (2005). *Acta Cryst.* B**61**, 585–594.
Sawka-Dobrowolska, W. & Głowiak, T. (1983). *Acta Cryst.* C**39**, 345–347.
Taylor, R. (2002). *Acta Cryst.* D**58**, 879–888.
Wei, C.-Y., Fischer, B. E. & Bau, R. (1978). *Chem. Commun.* pp. 1053–1054.
Zabrodsky, H., Peleg, S. & Avnir, D. (1992). *J. Am. Chem. Soc.* **114**, 7843–7851.
Zabrodsky, H., Peleg, S. & Avnir, D. (1993). *J. Am. Chem. Soc.* **115**, 8278–8289.
Zhao, Q.-H., Du, L. & Fang, R.-B. (2006). *Acta Cryst.* E**62**, m219–m221.
Zheng, L.-M., Song, H.-H., Duan, C.-Y. & Xin, X.-Q. (1999). *Inorg. Chem.* **38**, 5061–5066.
Zimmer, M. (2001). *Coord. Chem. Rev.* **212**, 133–163.