Continuous symmetry and shape measures, or how to measure the distance between polyhedra representing molecules

Pere Alemany

Institut de Química Teòrica i Computacional de la Universitat de Barcelona (IQTCUB)

Introduction

One of the most deeply rooted ideas in chemistry is that molecules which are similar should also behave in a similar way, that is, they should exhibit similar chemical and physical properties. This way of reasoning is at the origin of a large number of attempts to rationalize chemical observations, including the type theory developed in the first half of the XIX century, the concept of functional groups in organic chemistry or modern Quantitative Structure Activity Relations (QSAR) methods.

Although it is difficult to define precisely what is understood by similarity in chemistry, it is easy to distinguish two main aspects: chemical similarity arising from the periodic trends in the electronic structure of atoms (alcohols and thiols are similar in this sense since oxygen and sulphur have similar chemical properties) and structural similarity, referring to molecules in which atoms (which may be chemically similar or not) occupy similar positions in space (methane and the phosphate ion would be similar in this sense since both have tetrahedral structures). Intuitively it is easy to conclude that symmetry must play a crucial role in structural similarity, although it is not a trivial one, since symmetry and similarity are usually measured using different approaches: similarity is treated as a continuous (grey) property, allowing the definition of a degree of similarity, while symmetry is described as a discrete (black or white) property which is either present in a molecular structure or not. This gap has, however, been closed in recent years by the proposal of Avnir et al. defining symmetry also as a continuous
property by introducing the formalism of continuous symmetry measures (CSM) and the closely related continuous shape measures (CShM)\(^1\).

Let us start with the conceptually simpler problem that leads to the definition of continuous shape measures. As an example we may consider the phosphate ion (PO\(_4^3\)) present in a large number of inorganic and organic compounds. In classical structural chemistry the phosphate ion is said to have a tetrahedral structure, in the sense that the four oxygen atoms are located around the central phosphorus atom occupying the positions of the vertices of a tetrahedron (figure 1). Although the widespread idea that the phosphate ion is tetrahedral, a detailed analysis of 211 good quality crystal structures of compounds containing PO\(_4^3\) ions revealed, however, that none of the phosphorus atoms in these crystals was really located at a position compatible with full tetrahedral symmetry with all the phosphate ions presenting more or less pronounced distortions from the shape of an ideal tetrahedron.

Figure 1. Left: Crystal structure of Berlite (AlPO\(_4\)) showing PO\(_4\) tetrahedra in red and AlO\(_4\) tetrahedra in blue. Right: Ball and stick representation of one PO\(_4\) tetrahedron with the central phosphorus atom in black and the four oxygen atoms in red.

\(^1\) H. Zabrodsky, S. Peleg, D. Avnir, \textit{J. Am. Chem. Soc.} \textbf{1992}, \textit{114}, 7843-7851. This article introduces actually Continuous Symmetry Measures, not Continuous Shape Measures, although in the particular case of the tetrahedron these two measures can be shown to be equivalent.
A careful analysis reveals that situations like this are not an exception in structural chemistry, but that they are the rule. Several tools have been developed in order to quantify the distortions from ideal polyhedral structures, relying on the analysis of geometrical parameters such as bond distances and angles or on group theoretical techniques applied to a normal mode analysis for a given structure. All this approaches, although useful for the purposes for which they were developed, share a common problem, since they are not easily generalizable and their application is limited in many cases to small distortions for a specific polyhedron.

A general solution to this problem was given by Avnir and coworkers in 1992 by introducing continuous shape measures. The basic idea behind CShM is to define a distance between the real and the ideal polyhedra and to use this distance to gauge the similarity/dissimilarity between both structures.

The mathematical expression of this statement for an arbitrary structure \( Q \), defined as a set of \( N \) vertices with coordinates \( \{q_k\} \), with respect to a reference structure \( P \), defined as another set of \( N \) vertices with coordinates \( \{p_k\} \) is:

\[
S(Q,P) = \min_{k=1}^{N} \frac{\sum_{k=1}^{N} |q_k - p_k|^2}{\sum_{k=1}^{N} |q_k - q_0|^2} \cdot 100
\]

where \( q_0 \) is the geometric centre of \( Q \). Since the symmetry of an object does not depend neither on its size, position or orientation, the measure must be minimized for all possible relative positions and orientations between the structures, as well as with respect to their relative size. In other words, if we consider our problem structure \( Q \) with vertices \( Q = \{q_k\} \) and an ideal structure \( P \) with vertices located initially at \( P^0 = \{p_k^0\} \) the coordinates for the ideal structure \( P = \{p_k\} \) minimizing equation (1) can be expressed as:

\[\text{\footnotesize Ref. 2 M. Pinsky, D. Avnir, Inorg. Chem. 1998, 37, 5575-5582.}\]

\[\text{\footnotesize Ref. 3 For a detailed description of the mathematical background of Continuous Shape Measures see: David Casanova, Mesures de Forma i Simetria: Algorismes i Aplicacions (Doctoral Thesis), Universitat de Barcelona, 2006.}\]
\[ P = A R P^0 + t \]  (2)

where \( A \) is a scaling factor, \( R \) a matrix associated to a unitary transformation, the 3x3 rotation that determines the optimal spatial orientation of \( P \), and \( t \) a translation vector. The problem of minimizing \( S(Q,P) \) to find \( A, R \) and \( t \) has been solved analytically and an efficient algorithm is thus available. To ensure that \( S(Q,P) \) is really a measure of the \( P \)-shape contents of \( Q \), one must perform an additional minimization over all possible pairings between the vertices of \( Q \) and \( P \). Finally, the values of \( S(Q,P) \) are scaled by multiplying them by an arbitrary factor of 100. Adopting this definition for the CShM, an object having the desired \( P \)-shape will have \( S(Q,P) = 0 \), while distortions of this object from the ideal symmetry will lead to higher values of the measure (Figure 2).

**Figure 2.** Continuous Shape Measures for the coordination polyhedron around the cobalt atom in \([\text{Co}\{\text{PC}_6\text{H}_{5}\text{(EtS)}_2\}_2]\) showing that the best description for its geometry is that of a triangular prism.

Since minimization with respect to position, orientation, and scaling can be performed analytically, it is easy to see that the computational bottleneck in practical applications is associated to the minimization over all \( N! \) possible pairings between the vertices of \( Q \) and those of \( P \).

The definition of continuous shape measures is easily extended to the so-called continuous symmetry measures. The question in this case is somewhat different. While
in CShM we compare the shape of a problem structure $Q$ defined by $N$ vertices with that of a reference structure $P$ with the same number of vertices, in CSM we search for the closest structure $T$ with $N$ vertices that belongs to a given point group $G$: 

$$S(Q, G) = \min_{k=1}^{N} \frac{\sum_{k=1}^{N} |q_k - t_k|^2}{\sum_{k=1}^{N} |q_k - q_0|^2} \cdot 100$$

(3)

Although for some special structures such as the tetrahedron, CShM and CSM are coincident since the tetrahedron is the only 4-vertex polyhedron with $T_d$ symmetry, in general, CShMs are more restrictive than CSMs. If we want to evaluate, for instance, the $S(Q, D_{4h})$ value for a given 8-vertex structure $Q$, we need not only to compare $Q$ with a given square prism $P$, but we need to find which is the square prism $T$ that is closest (regardless of position, orientation and size) to the problem structure $Q$. In this example this implies an additional minimization of the length/height ratio for the square prisms, but in the general case, looking for $T$ is not a trivial task. To solve this problem, an efficient numerical algorithm, called the folding-unfolding method has been developed\(^4\), and as in the case of CShMs, the computational bottleneck in the calculation of CSMs is again the minimization over all $N!$ possible vertex pairings between the problem and reference structures $Q$ and $T$.

Lately we have found an alternative way to search for the desired structure $T$ with a given symmetry that is based on the calculation of the so-called symmetry operation measures\(^5\). Given a symmetry point group $G = \{R_i\}$ with $h$ symmetry operations and a problem structure $Q$ we can define for each operation $R_i$ a symmetry operation measure $Z(Q, R_i)$ as:

$$Z(Q, R_i) = \min_{k=1}^{N} \frac{\sum_{k=1}^{N} |q_k - R_i q_k|^2}{4 \sum_{k=1}^{N} |q_k - q_0|^2} \cdot 100$$

(4)

\(^4\) See ref. 1 for a detailed description of this algorithm.

where we compare now the problem structure $Q$ with the image obtained from the symmetry operation acting on it. The minimization procedure is somewhat simplified in this case since we need not to minimize $Z(Q,R_i)$ with respect to translations (if $R_i$ is a symmetry operation belonging to a point group it must leave a single point invariant and this point coincides with the geometric centre of $Q$) and the relative size of both structures (since point group symmetry operations do not change the object’s size). The only geometrical minimization that is needed in this case is with respect of the relative orientation of the problem structure and its image, or equivalently to minimize the measure with respect to the spatial orientation of the symmetry element that is associated with the symmetry operation $R_i$. As in the other cases, an optimization over all possible pairings between vertices of the problem and reference structures is also needed, although in this case there is an additional restriction to the possible permutations, since only G-symmetry preserving permutations should be considered.

It can be shown that the value of the CSM for a given group can be easily evaluated using the symmetry operation measures for the $h$ operations of the group:

$$S(Q,G) = \frac{1}{h} \sum_{i=1}^{h} \bar{Z}(Q,R_i)$$

(5)

where the symbol $\bar{Z}(Q,R_i)$ is used to indicate that the minimization procedure to find the symmetry operation measures must be performed simultaneously for all operations in the group while imposing the necessary constraints to keep the relative orientation of the different operations fixed (for a $C_{2h}$ group, for example, one must perform the minimizations necessary to obtain the symmetry operation measures while keeping the $C_2$ axis perpendicular to the mirror plane). In this respect, $\bar{Z}(Q,R_i)$ are not real symmetry operation measures since the value of $Z(Q,R_i)$ may be lower if these constraints are released.

---

Problem

As discussed above, one of the main problems in the application of continuous shape or symmetry measures to structural chemistry is the $N!$ dependence of the algorithms devised for the numerical calculation of CShMs and CSMs. A typical study in structural chemistry starts with a search for a given fragment formed by $N$ atoms in a database such as the Cambridge Structural Database, containing information on the crystal structures for over 400,000 chemical compounds. In this type of search, the desired fragment is usually described using the connectivity between atoms, although additional geometrical restrictions on bond distances and/or angles may be imposed to limit the search. As a result, the user obtains for each fragment satisfying the conditions specified in the search a set of $N$ coordinates, one for each atom in the fragment. CShMs and/or CSMs are calculated afterwards for each fragment and used, for instance, to classify the retrieved fragments according to their coordination geometries. The plot of $S(Q,P)$ vs $S(Q,P')$, where $P$ and $P'$ are two different reference polyhedra is called a shape map and it has been shown to be a very useful tool in detecting structural trends for different families of compounds (Figure 3).

![Shape map](image)

**Figure 3** Shape map for over 13,000 tetracoordinated ML$_4$ transition metal complexes classified according to their tetrahedral (T4) and square-planar (SP4) continuous shape measures.

With our actual programs this process is feasible for sets of around 10,000 fragments with up to approximately 10 vertices. Although we have been able to calculate CSMs for a small set of structures with up to 15 atoms, the study of structures
with more than 20 atoms is nowadays impossible if no additional simplifying assumptions are introduced. It is therefore crucial for the application of CShMs and CSMs in structural chemistry to develop efficient algorithms to deal with the optimum vertex pairing problem.

Although there are some differences in the treatment of the minimization of vertex pairings in the calculation of CShMs and CSMs, the basic problem remains the same:

Is there an efficient way to avoid the sweeping across all N! index permutations to find the optimal vertex pairing between the problem and the reference structures?

In the following we will sketch the actual algorithm implemented in our program as well as some considerations related to the chemical properties of the analyzed structures that may, in some cases, provide additional information useful to avoid a search over all possible permutations.

In order to simplify the procedure, in a first step the size of $Q$ is normalized so that:

$$N = \sum_{k=1}^{N} |q_k - q_0|^2$$

(6)

Omitting the arbitrary factor of 100 that appears in the equations above, the minimization in eq. (1) is actually performed in two steps:

$$S(Q,P) = \frac{1}{N} \min_{A,R,t} \left( \min_{k} \sum_{k=1}^{N} |q_k - p_k|^2 \right) = \frac{1}{N} \min_{A,R,t} S_{N,n,m}(Q,P)$$

(7)

where first we minimize the sum of the square of the distances between vertices with respect to size, orientation, and translation $(A, R, t)$ for a given vertex pairing $(P_m)$ to obtain $S_{N,n,m}(Q,P)$, and afterwards we sweep through all permutations of vertices of the reference shape to find the vertex pairing with the smallest value of $S_{N,n,m}(Q,P)$ where the subindex $N$ is used to highlight that all $N$ vertices of $P$ and $Q$ are being compared.

In the actual algorithm for calculating CShMs, if the number of vertices is large enough (polyhedra with more than 7-8 vertices) the set of $N$ vertices is divided into $n_L$ subsets of $L < N$ vertices, and, if necessary, an additional set with $N$-$n_L L$ vertices. Given a fixed vertex pairing, it is possible to define $n_L$ functions $S_{L,n,m}(Q,P)$ analogous to $S_{N,n,m}$
in which only the vertices of a given subset are compared and the optimal values for \( A, R, t \) for each subset are calculated. Since all these functions are positive, for a given vertex pairing \( (P_m) \) the following inequality must hold:

\[
S_{N,m} \geq \frac{L}{N} \sum_{i=1}^{2^L - 1} S_{L,m,i}
\] (8)

This expression shows that if any of the individual \( S_{L,m} \) values is found to be larger than a previously stored value of \( S_{N,m} \), then it is not necessary to calculate values of \( S_{N,m} \) for all \((N-L)!\) permutations with the same vertices in this partition. Tests using \( L = 4 \) have shown that for structures with more that \( N = 8 \) vertices the algorithm based on partitioning the set of vertices in different subsets is, in general, more efficient than sweeping over all \( N! \) permutations since for most of the analyzed structures a large amount of vertex pairings can be discarded. This algorithm has allowed us to reach the calculation of CShMs for structures with up to around 20 vertices without imposing any additional approximation to discard vertex pairings.

Besides this algorithm, the program allows to introduce additional restrictions in order to reduce the number of permutations that must be analyzed. The simplest one is to restrict the analyzed permutations just to a given set that is provided in the input by the user. This is an extremely fast solution that is mostly applicable in the cases in which the problem and reference structures are closely related and a visual inspection is able to provide unambiguously the optimum vertex pairing (or a small set of candidates for it). Although extremely efficient, this solution is not applicable to the automatic computation of CSMs for large sets of data since it requires a previous visual inspection to decide the permutations that will be analyzed, and it is also susceptible of important errors since it depends on the subjective choice of a limited set of vertex pairings, a task that is not always easy.

Another approximation that may be useful is to divide the vertices in a number of subsets and to perform all permutations only within each of these subsets. This approximation can be applied if chemical information is available (we can, for instance, in the structure of an hydrocarbon \( C_nH_m \) molecule restrict the number of permutations to those which interchange only carbon atoms with other carbon atoms and hydrogen atoms with other hydrogen atoms). Although useful, this approximation is not very
efficient since the limit of feasible permutations within each subset is easily achieved. Besides this, it is often desirable to compare the structures of molecules built from different atoms in order to highlight structural similarities that go beyond the chemical nature of the molecules, a situation where it is difficult to define the different sets of atoms unambiguously. This technique has also been applied to molecular clusters that are built by concentric layers of atoms. In this case it is sometimes possible to restrict permutations to atoms within each layer (or at least between atoms in a given layer or in the two closest ones).

Related to this approach, it is also possible to use often information related to the graph representing the chemical bonds in a molecule. For the case of CSMs we have devised an algorithm that classifies the different vertices into different subsets according to their connectivity in the molecular graph, and restricting the permutations only between atoms that have the same connectivity (Figure 3). Although this algorithm saves a large amount of time for some cases, we are aware that it does not exploit the full symmetry of the molecular graph, a direction that is probably worth exploring.

![Figure 3](image-url)

**Figure 3.** Ball and stick representation of diphenyl-chlorarsine showing the partitioning of the 12 carbon atoms into 4 different sets A, B, C, and D according to their connectivity in the molecular graph.

Although the use of the molecular graph may be useful in limiting the number of vertex pairings to be analyzed, its use is restricted for various reasons. Two different questions arise when trying to use molecular graphs in order to limit the number of vertex pairings to be analyzed. Both problems are related to the possibility of defining the molecular graph unambiguously.
A molecular structure is a graphical representation of the average position of the atomic nuclei in a given molecule. These nuclei are in reality in constant movement around their so-called equilibrium positions. If a nucleus is separated from its equilibrium position, the molecular energy is increased. For a polyatomic molecule with N atoms, the potential energy surface (a function that gives the energy of the molecule given the position of the nuclei) shows, in general, more than one equilibrium nuclear configurations in which the nuclei are arranged in different positions. Each of these nuclear configurations gives a different molecular graph in which the connections between atoms correspond to chemical bonds. The criteria to decide if there is a bond or not between two atoms is, however, not unambiguous. The potential energy surface for a diatomic molecule is a simple function of the internuclear distance that can be conveniently described by the Morse function (Figure 4, in blue).

**Figure 4.** Potential energy curves for a diatomic molecule. The blue curve corresponds to the Morse function while the green curve corresponds to the harmonic approximation, used to study small deviations of the nuclei from their equilibrium positions.

The Morse function is characterized by two parameters, the equilibrium distance \( r_e \) and the dissociation energy \( D_e \). The first one corresponds to the internuclear distance for which the energy is minimal, while \( D_e \) indicates the energy that is needed to separate the two atoms (to break the bond). Although in a polyatomic molecule it is not possible to write a simple mathematical expression for the potential energy surface, it is often
possible to analyze how the energy changes when a pair of atoms is separated and these energy changes can be described also by a Morse function. Analyzing different molecular structures it is easy to find similar $r_e$ and $D_e$ values for a given pair of atoms and this allows chemists to visualize a molecule as a set of atoms linked by a number of chemical bonds. In hydrocarbons, for example, for C-C pairs typical $r_e$ distances around 1.5Å are found and programs used to visualize molecular structures will plot a bond between carbon atoms at a distance shorter than approximately 1.6 – 1.7Å. If the minimum in the Morse curve between two atoms is narrow, the energy rises very steeply with the distance and there is a small variation in the bond distances between these two atoms in different molecules. This is, however, not always the case. There are situations where there are “weak bonding” interactions between atoms, giving a broad Morse curve for which variations around $r_e$ represent only small energy changes. This is translated in a large dispersion of values in the internuclear distances in different molecules containing this pair of atoms and the difficulty in establishing a clear criterion of what should be considered a bond and, hence, in defining the molecular graph. This situation is found, for instance, in metal clusters like Au$_{28}$ (figure 5), where there are a total of 94 Au-Au distances between 2.7 and 3.4Å with an average Au-Au distance around 2.9Å.

![Figure 5](image.png)

**Figure 5.** Ball and stick representation of an Au$_{28}$ cluster plotting bonds for distances below 2.9Å (left) or 3.4Å (right).

This example shows the first limitation of the application of the molecular graph to the computation of CSMs and CShMs. Changing the distance to define an Au-Au bond between 2.7 and 3.4Å gives different molecular graphs that lead to different values for the symmetry or shape measures when using the graph to limit the number of vertex pairings to be searched, or even to cases where these measures can not be calculated.
because the possible the number of vertex pairings within each of the vertex subsets is too high.

The other problem is related to the application of molecular graphs to continuous shape measures. While in continuous symmetry measures the problem structure and the reference structure (the image of the problem structure under the application of a symmetry operation) share the same molecular graph, this is not necessarily true for CShM. Although it is possible to establish a standard molecular graph for the reference structure, this graph will, in general, be only preserved for small structural distortions. For larger distortions chemical bonds may be broken or formed and the graph of the problem structure will change as shown in the example in figure 6. When analyzing large sets of fragments retrieved from a structural database it is very difficult to detect these changes since the definition of the molecular graph depends, as explained above, on the chemical nature of the atoms and it is quite difficult to develop an algorithm to establish unambiguously the equivalence of a given graph with that of the reference structure.

![Figure 6. Ball and stick representation of two deformations of a tetrahedral fragment with different molecular graphs.](image)

As a conclusion, let us restate our problem: Is there an efficient way to avoid the sweeping across all N! index permutations to find the optimal vertex pairing between the problem and the reference structures? As we have shown, the problem can be addressed in several ways, depending on the additional information that we have at hand for each case. The question remains, however, open for difficult cases like the Au_{28} cluster shown in figure 5, where the molecular graph is of little aid and the number of permutations to high to be treated in our actual computers.