SIMILARITY ANALYSIS IN TWO AND THREE DIMENSIONS USING LATTICE ANIMALS AND POLYCUBES

Paul G. MEZEY

Mathematical Chemistry Research Unit, Department of Chemistry and Department of Mathematics, University of Saskatchewan, Saskatoon, Canada S7N 0W0

Abstract

The quantification and analysis of molecular similarity are fundamental problems of both theoretical and applied chemistry. The continuum similarity problem of planar domains with Jordan curve boundaries can be discretized and quantified using interior filling animals (square cell configurations). A similar approach is applicable to the continuum similarity problem of formal molecular bodies enclosed by contour surfaces, where interior filling polycubes provide a method for discretization and quantification of molecular similarity in three dimensions. This technique leads to resolution based similarity measures (RBSMs), suitable for automatic, non-visual evaluation of the degree of similarity between shapes of general objects, in particular, of molecular charge distributions, or fused sphere Van der Waals surfaces. Using the framework of the RBSM method, the polycube method of chirality quantification is extended to the quantification of approximate symmetry of molecular electron distributions.

1. Introduction

Square-cell configurations on a planar square lattice (often called lattice animals, or simply animals, if some constraints are satisfied), as well as polycubes of a three-dimensional cubic lattice provide very useful, simple models for the characterization of physical objects and processes [1–8]. In this study, we shall describe some results involving lattice animals and polycubes which are of relevance to the study of molecular similarity.

Similarity of structural properties, in particular, the similarity of shapes of electron distributions of various molecules, has often been invoked in explaining similar chemical or biochemical behavior [9]. Yet, no standard convention exists for the evaluation of similarity of shapes. In applications, such as drug design, the usual approach involves a visual evaluation of similarity of molecular images generated on a computer screen. Such techniques, however, appear both subjective and not well reproducible; consequently, alternative techniques such as *non-visual, algorithmic methods for similarity evaluation* are likely to become more useful in applications where reproducibility and reliability are of importance. In an earlier work, a conceptually simple method has been proposed for the evaluation of a numerical measure of similarity of formal molecular bodies, using the principle of resolution based similarity measures (RBSMs) designed to mimic certain aspects of visual perception on a computer [10]. In what follows below, we shall briefly review the background of the RBSM method, which is the basis of the developments discussed in this work.

The basic idea behind this method is very simple. When two objects of different shapes are observed at a great distance, they are likely to appear nearly identical, as two distant points. At a smaller distance, their shape differences may become apparent, and at a close distance their differences are well established. Observations at great and small distances may be regarded as observations at low and high resolution, respectively. Clearly, the more similar the two objects the higher the resolution needed to detect differences. Hence, a numerical similarity index can be associated with the level of resolution required for the detection of shape differences, leading to resolution-based similarity measures.

Instead of evaluating the shape differences of the original objects directly, one may consider *discrete* approximations of the objects. For example, one may use inscribed polycubes in order to provide such discrete approximations [10]. By polycube, we mean a face-connected family of cubes within the three-dimensional cubic lattice, fulfilling some additional constraints [10]. An *interior filling polycube* is one which fits within the object but no polycube of the same cube size and of more cubes fits within the object. One may use the methods of discrete mathematics to evaluate the similarities of the families of polycubes inscribed into the objects. Hence, the similarity problem of two continua (the two objects) can be approximately represented by a similarity problem of two polycubes with discrete characterization.

Note that we do not distinguish between rotated and translated versions of a polycube, and between versions which differ only in size: any two polycubes P_n and P_n' of n cubes which can be superimposed on one another by scaling, translation, and rotation in 3D space are regarded as identical, $P_n = P_n'$. Consequently, when comparing two polycubes, only the relative, topological arrangements of their cubes are relevant. Since for a finite number of cubes there are only a finite number of topologically distinct cube arrangements, polycubes do, indeed, provide a discretized approximation to the shape description of the original objects.

When using interior filling polycubes, a natural, size-independent level of resolution can be associated with the number of cubes of the inscribed polycube. If the resolution is low, then only a rough approximation of the object is given by a polycube containing only a few cubes; if the resolution is high, then a close approximation of the object is given by an interior filling polycube of a large number of cubes. A size-independent shape-similarity measure is obtained if both small and large objects are described by the *same number* of cubes, and the corresponding interior filling polycubes are compared. For this purpose, each level of resolution is defined by n, the number of cubes of interior filling polycubes, which, of course, depends on the relative size of the objects as compared to the cube size s.

The shape of a molecule with a formal, fixed nuclear geometry K may be represented by isodensity contours. Following the notations of ref. [10], $G(a)$ and

$B(a)$ denote the isodensity contour surface and the formal molecular body enclosed by it, respectively, if the electronic density value along the contour is the constant a. The ith n-cube interior filling polycube of $G(a)$ is denoted by $P_i(G(a), n)$.

Consider two contour surfaces, G_1 and G_2 of two different molecules, or of the same molecule at two different density values a_1 and a_2, and denote by $F(G_1, G_2, n)$ the *family of common n*-cube interior filling polycubes $P_i(G_1, n)$ and $P_j(G_2, n)$. The similarity index $i_0(G_1, G_2)$, the degree of dissimilarity $d(G_1, G_2)$, and the degree of similarity $s(G_1, G_2)$ of the two contour surfaces G_1 and G_2 have been defined [10] as follows.

The *similarity index* $i_0(G_1, G_2)$ is the smallest n_c value at and above which all interior filling polycubes of contour surfaces G_1 and G_2 are different:

$$i_0(G_1,G_2) = \begin{cases} \min\{n_c : F(G_1,G_2,n) \text{ is empty if } n \geq n_c\}, & \text{if the minimum exists,} \\ \infty & \text{otherwise.} \end{cases} \tag{1}$$

If two contour surfaces G_1 and G_2 can be obtained from one another by translation, rotation, and scaling, then their shapes are identical; for G_1 and G_2 of identical shapes, no finite n_c value exists and $i_0(G_1, G_2) = \infty$.

The *degree of dissimilarity* $d(G_1, G_2)$ is defined as

$$d(G_1,G_2) = 1/(i_0(G_1,G_2) - 2). \tag{2}$$

For both cube numbers $n = 1$ and $n = 2$, the polycubes are unique and on these levels of resolution no dissimilarity exists; hence, $i_0(G_1, G_2) > 2$ is always valid. This is why the number two appears in the denominator; $d(G_1, G_2)$ takes values from the [0, 1] interval.

The *degree of similarity* $s(G_1, G_2)$ of two contour surfaces G_1 and G_2 is defined as

$$s(G_1, G_2) = 1 - d(G_1, G_2). \tag{3}$$

If the two contour surfaces G_1 and G_2 have identical shapes, then their degree of similarity $s(G_1, G_2) = 1$, otherwise $s(G_1, G_2)$ is a smaller positive number.

Although for molecular shape analysis the three-dimensional case is of the most relevance, the two-dimensional case of shape characterization of planar continua is of special importance. Hence, we shall describe the analogous methods involving lattice animals. By contrast to earlier studies on similarity [10], as well as on chirality [11,12], here we shall not restrict ourselves to simply connected planar domains; multiply connected square cell configurations will also be regarded as formal lattice animals. We shall have the following requirements:

A lattice animal A is a connected arrangement of a finite number n of impenetrable squares c (called cells) of uniform size s in the plane, if

(i) only two types of contacts between cells are allowed: a common edge or a common vertex;

(ii) if $n > 1$, then each cell of the animal A must have an edge contact with another cell of A;

(iii) if there is a vertex contact between two cells c and c' of A, then there must also be either an edge contact between them or there must exist a cell c'' with edge contact to both c and c';

(iv) the animal, as a planar set, is topologically equivalent to the planar continuum it represents. In the most common case, this continuum is topologically equivalent to a disk; however, more complicated topologies are also of importance.

An n-cell *interior filling animal* $A_i(C, n)$ of a planar continuum T with a boundary C of finite length is an animal which fits within C, but no animal of the same cell size s and more that n cells can be inscribed in C. Note that this definition is broader than that used in earlier studies [10–12], where no multiply connected square-cell configurations were considered and requirement (iv) was not specified.

Consider two planar sets T_1 and T_2, their boundaries C_1 and C_2, respectively, and denote by $F(C_1, C_2, n)$ the *family of common* n-cell interior filling animals $A_i(C_1, n)$ and $A_j(C_2, n)$. Following the definition in ref. [10] for the simpler case, the *similarity index* $i_0(C_1, C_2)$, the *degree of dissimilarity* $d(C_1, C_2)$, and the *degree of similarity* $s(C_1, C_2)$ of the two planar sets C_1 and C_2 can be defined as follows:

The *similarity index* $i_0(C_1, C_2)$ is the smallest n_c value at and above which all interior filling animals of the planar set with boundaries C_1 and C_2 are different:

$$i_0(C_1,C_2) = \begin{cases} \min\{n_c : F(C_1,C_2,n) \text{ is empty if } n \geq n_c\}, & \text{if the minimum exists,} \\ \infty & \text{otherwise.} \end{cases} \tag{4}$$

If two boundaries C_1 and C_2 can be obtained from one another by translation, rotation, and scaling, then their shapes are regarded as identical. For C_1 and C_2 of identical shapes, no finite n_c value exists and we obtain $i_0(C_1, C_2) = \infty$.

We define the *degree of dissimilarity* $d(C_1, C_2)$ as

$$d(C_1, C_2) = 1/(i_0(C_1, C_2) - 2). \tag{5}$$

This is similar to the three-dimensional case, since for both cell numbers $n = 1$ and $n = 2$ the animals are unique; hence, on these levels of resolution, no dissimilarity may exist. This fact is taken into account by the inclusion of the number two in the denominator; $d(C_1, C_2)$ takes values from the [0, 1] interval.

The *degree of similarity* $s(C_1, C_2)$ of two boundary lines C_1 and C_2 is defined as

$$s(C_1, C_2) = 1 - d(C_1, C_2). \tag{6}$$

If the two boundary lines C_1 and C_2 have identical shapes, then their degree of similarity $s(C_1, C_2) = 1$, otherwise $s(C_1, C_2)$ is a smaller positive number.

The detection of the presence or the lack of chirality is also resolution dependent. Although chirality is an absolute property, this resolution dependence allows one to introduce a formal scale for chirality. Clearly, if chirality is already detectable at a low level of resolution, one may regard the object "more chiral" than another object that reveals its chirality only at a much higher level of resolution. For both the two- and three-dimensional cases of chirality, a formal degree of chirality has been introduced [11,12] based on a discretization of shape features using lattice animals and polycubes. Following the original definition given for interiors of Jordan curves in the plane [11], here we shall give a definition for a more general boundary line C in the plane. This definition is based on chiral animals.

An animal A is achiral if and only if A can be superimposed on its mirror image $A^{\Diamond}$ by translation and rotation within the plane:

$$A = A^{\Diamond}. \tag{7}$$

Otherwise, the animal A is chiral.

For each boundary C, we shall consider a chirality index, defined as a critical cell number $n_{\chi}(C)$, at and above which all interior filling animals $A_i(C, n)$ are chiral.

We say that C is chiral at and above cell number n_{χ} if each $A_i(C, n)$ is chiral if $n \geq n_{\chi}$. The *chirality index* $n_{\chi}(C)$ is the smallest n_{χ} value above which all interior filling animals $A_i(C, n)$ are chiral,

$$n_{\chi}(C) = \begin{cases} \min\{n_{\chi} : A_i(C,n) \text{ is chiral if } n \geq n_{\chi}\}, & \text{if the minimum exists,} \\ \infty & \text{otherwise.} \end{cases} \tag{8}$$

Since the smallest chiral lattice animals have four cells [11], the minimum possible value for the chirality index is $n_{\chi}(J) = 4$. The *degree of chirality* $\chi(C)$ of a boundary curve C of a planar continuum T is defined as

$$\chi(C) = 1/(n_{\chi}(C) - 3). \tag{9}$$

This measure of chirality gives the value 1 for "very chiral" curves and 0 for achiral ones.

A similar treatment has been applied for the three-dimensional case [12]. A polycube P_n is achiral if and only if P_n can be superimposed on its mirror image $P_n^{\Diamond}$ by translation and rotation:

$$P_n = P_n^{\Diamond}. \tag{10}$$

Otherwise, the polycube P_n is chiral.

An isodensity contour surface $G(a)$ is chiral at and above cube number n_{χ} if each interior filling polycube $P_n(G(a))$ is chiral if $n \geq n_{\chi}$. The *chirality index* $n_{\chi}(G(a))$ is the smallest n_{χ} value above which all interior filling polycubes $P_n(G(a))$ are chiral:

$$n_{\chi}(G(a)) = \begin{cases} \min\{n_{\chi} : P_n(G(a)) \text{ is chiral if } n \geq n_{\chi}\}, & \text{if the minimum exists,} \\ \infty & \text{otherwise.} \end{cases} \tag{11}$$

The degree of chirality $\chi(G(a))$ of a molecular contour surface $G(a)$ is

$$\chi(G(a)) = 1/(n_\chi(G(a)) - 3). \tag{12}$$

The smallest chiral polycube has four cubes, and the number three in the denominator ensures that the degree of chirality is a number from the [0, 1] interval.

The above resolution based chirality measures using interior filling lattice animals [11] and polycubes [12] may be regarded as extensions of the general RBSM principle [10]. They also provide alternatives to

(i) earlier chirality measures based on area and overlap, described by Kitaigorodskii, Gilat, Schulman, Mislow, Buda, and Auf der Heyde [13–19];

(ii) descriptions in terms of reference objects using the Haussdorf distance of point sets for characterization by Rassat [20];

(iii) fuzzy set representation [21] of chirality, using fuzziness in an epistemological sense by Mislow and Bickart [22]; and to

(iv) the principle of energy-weighted fuzzy achirality resemblance of Mezey [12], based on the syntopy model developed by Mezey and Maruani [23].

2. Chirality measures based on the degree of similarity

The RBSM method is applicable for the introduction of two new approaches towards the quantification of chirality. The first such approach is based on the degree of similarity $s(G, G^\lozenge)$ or $s(C, C^\lozenge)$ between the object G or C and its mirror image $G^\lozenge$ or $C^\lozenge$, in the three- or two-dimensional cases, respectively.

The *similarity based measures* $\alpha_s(G)$ and $\alpha_s(C)$ *of achirality* are defined as

$$\alpha_s(G) = 2\, s(G, G^\lozenge) - 1 \tag{13}$$

and

$$\alpha_s(C) = 2\, s(C, C^\lozenge) - 1, \tag{14}$$

respectively. Note that in the general case, the smallest possible value of both the two- and three-dimensional similarity indices is 3. However, for mirror images, the smallest possible value for both similarity indices is 4, obtained for the enantiomeric pair of one of the smallest chiral animals (called "Tippy", see, e.g. ref. [11]), and for the enantiomeric pair of the smallest chiral polycube, respectively. Consequently, a rescaling of the $s(G, G^\lozenge)$ and $s(C, C^\lozenge)$ measures is required, as given in eqs. (13) and (14). The coefficient 2 and the term -1 in the above expressions ensure that these achirality measures are taking values from the [0, 1] interval. Objects which are "fully achiral", that is, objects which appear achiral at any level $n \geq 4$ of resolution, have an achirality measure of $\alpha_s = 1$, whereas objects showing the greatest dissimilarity with their mirror images have an achirality measure of $\alpha_s = 0$.

In turn, the *similarity based measures* $\chi_s(G)$ and $\chi_s(C)$ *of chirality* are defined as

$$\chi_s(G) = 1 - \alpha_s(G) \tag{15}$$

and

$$\chi_s(C) = 1 - \alpha_s(C), \tag{16}$$

respectively. Objects with prominent chirality have χ_s measures close to 1, whereas achiral objects have a χ_s measure equal to 0.

In an indirect way, the above similarity based measures of chirality rely on a reference to animals and polycubes of *maximal* chirality by the earlier criterion, since these are the same animals of polycubes having maximum chirality of 1 by the new measure. The two chiral four-cell animals are A' = "Tippy" and A'' = "Elly", both of a 2×3 mesh and of binary codes c(Tippy) = 110011 and c(Elly) = 111100, respectively [24]. (We recall that the specification of the mesh and the binary code provides a complete characterization of animals [24].) Note that Elly, A'' has achiral interior filling animals of more than four cells, as opposed to Tippy, A', which has not. Consequently, according to our definition, Tippy, A' is more chiral than Elly, and we shall not use Elly as a reference. The smallest chiral polycube is the four-cube screw P_4', where the four cubes are arranged as in condition (ii) of ref. [12].

This measure of chirality is different from the earlier measure [11,12]. It is possible that for some curve C only chiral interior filling animals occur at some level n_c and above, yet the *same* chiral interior filling animal A of n_c cells and its mirror image $A^\Diamond$ may occur for both C and its mirror image $C^\Diamond$. Hence, at this level, the test of full dissimilarity fails, whereas the earlier test of chirality already gives $n_\chi \leq n_c$. In such a case, the new measure provides more discrimination than the earlier one [11].

The testing of interior filling animals and the determination of suitable upper bounds n_c and n_χ for the calculation of similarity and chirality indices and the respective measures involve several computational steps discussed in the appendix. It is possible, however, to choose a given level or a given, limited range of resolution, suitable for the chemical problem at hand, that leads to resolution-dependent similarity and chirality measures. These measures are discussed in section 3.

3. Similarity and chirality measures for limited resolutions

Considering both limited and infinite resolutions, an interesting connection can be established between the discretized, lattice animal and polycube measures of chirality and the measures based on the overlapping area and volume of mirror images, developed by Kitaigorodskii, Gilat, Shulman, Mislow, Buda, and Auf der Heyde [13–19]. In the 2D case, we first consider a fixed number n of cells, a boundary C, its mirror image $C^\Diamond$ and another boundary C'. At level n of resolution, we define two quantities:

$d(C, C', n)$ = the minimum number of cells needed to move in order to turn an interior filling animal $A_i(C, n)$ of C into an interior filling animal $A_j(C', n)$ of C', (17)

$k(C, n)$ = the minimum number of cells needed to move in order to turn an interior filling animal $A_i(C, n)$ of C into an achiral animal. (18)

A resolution-dependent dissimilarity measure is given by

$$D(C, C', n) = d(C, C', n)/n, \tag{19}$$

and two resolution-dependent chirality measures are

$$K(C, n) = k(C, n)/n \tag{20}$$

and

$$K(C, C^{\Diamond}, n) = D(C, C^{\Diamond}, n). \tag{21}$$

Note that if either C or C' has no n-cell interior filling animal, then neither $d(C, C', n)$ nor $D(C, C', n)$ is defined. If C has no n-cell interior filling animal, then no quantities $k(C, n)$, $K(C, n)$ and $K(C, C^{\Diamond}, n)$ are defined.

In the limit of large n, $K(C, C^{\Diamond}, n)$ converges to the ratio of the area of the part of the interior of C not covered by the interior of $C^{\Diamond}$, and the area of the interior of C, assuming maximal overlap between C and $C^{\Diamond}$. Let us denote the area of a set of boundary C by $a(C)$, and the area of maximum area intersection of the interiors of C and C' by $a(C \wedge C')$. Then,

$$\lim_{n\to\infty} K(C, C^{\Diamond}, n) = (a(C) - a(C \wedge C^{\Diamond})) / a(C). \tag{22}$$

It is easily seen that this limit is equal to the overlap measure $\chi(T)$ of chirality as used by Mislow et al. [16–19]:

$$\lim_{n\to\infty} K(C, C^{\Diamond}, n) = \chi(T), \tag{23}$$

with $T = T(C)$, the planar set of boundary C.

Consider two n-cell animals A_1 and A_2. Their distance is defined as

$d(A_1, A_2)$ = the minimum number of cells of A_1 must be moved in order to turn A_1 into A_2. (24)

Clearly,

$$d(A_1, A_2) = d(A_2, A_1). \tag{25}$$

If (but not only if) both C and C' have only one n-cell interior filling animal each, $A_1(C, n)$ and $A_1(C', n)$, respectively, then

$$d(C, C', n) = d(A_1(C, n), A_1(C', n)), \tag{26}$$

and

$$D(C, C', n) = d(A_1(C, n), A_1(C', n))/n. \tag{27}$$

Furthermore, if $C' = C^{\Diamond}$, then

$$K(C, C^{\Diamond}, n) = d(A_1(C, n), A_1(C^{\Diamond}, n))/n = d(A_1(C, n), A_1^{\Diamond}(C, n))/n. \tag{28}$$

For an n-cell animal A, we define

$$k(A) = \text{the minimum number of cells must be moved in order to turn } A \text{ into an achiral animal.} \tag{29}$$

If (but not only if) C has only one interior filling animal $A_1(C, n)$, then

$$k(C, n) = k(A_1(C, n)) \tag{30}$$

and

$$K(C, n) = k(A_1(C, n))/n. \tag{31}$$

Consider two animals A_1 and A_2, of n_1 and n_2 cells, respectively. Without loss of generality, we may assume that $n_1 \leq n_2$. A *common parasite animal*

$$P(A_1, A_2) \tag{32}$$

of A_1 and A_2 is an animal that is contained in both A_1 and A_2, whereas a *common host animal*

$$H(A_1, A_2) \tag{33}$$

of A_1 and A_2 is an animal that contains both A_1 and A_2. A *maximal common parasite animal*

$$P_{\mathrm{m}}(A_1, A_2) \tag{34}$$

of A_1 and A_2 is a parasite of A_1 and A_2 of largest number of cells, whereas a *minimal common host animal*

$$H_{\mathrm{m}}(A_1, A_2) \tag{35}$$

of A_1 and A_2 is a host of A_1 and A_2 of smallest number of cells.

The *shrinking coincidence index*

$$\mathrm{sc}(A_1, A_2) = n_2 - n(P_{\mathrm{m}}(A_1, A_2)) \tag{36}$$

of A_1 and A_2 is the difference between the cell numbers of the larger of the two animals and a maximal parasite. Intuitively, if both animals are losing cells, then the shrinking coincidence index is the minimum number of cells the larger animal must lose before it can become a parasite common for both animals.

The *growing coincidence index*

$$\mathrm{gc}(A_1, A_2) = n(H_\mathrm{m}(A_1, A_2)) - n_1 \tag{37}$$

of A_1 and A_2 is the difference between the cell numbers of a minimal host and the smaller of the two animals. Intuitively, if both animals are growing cells, then the growing coincidence index is the minimum number of cells the smaller animal must grow before it can become a host common to both animals.

The above two indices define two additional measures of chirality of animals. The first of these is based on the minimum number of cells that need to be removed in order to turn an animal A into an achiral animal. This measure is given as

$$\chi_{\mathrm{sc}}(A) = \mathrm{sc}(A, A^{\Diamond})/n. \tag{38}$$

The second measure is based on the minimum number of cells that need to be added in order to turn an animal A into an achiral animal. This measure is given as

$$\chi_{\mathrm{gc}}(A) = \mathrm{gc}(A, A^{\Diamond})/n. \tag{39}$$

All definitions and results of this section can be generalized for the three-dimensional case of polycubes, or for the case of abstract υ-dimensional hypercubes, by replacing the terms and symbols of animal, A, boundary, C, and cell with polycube, P, contour surface, G, and cube, or with polyhypercube, P, contour hypersurface, G, and hypercube, respectively.

4. Resolution based symmetry deficiency measures

Chirality can be regarded as the lack of certain symmetry elements, and chirality measures and measures of symmetry deficiency. In three-dimensional chirality, special symmetry elements are involved: the presence of a reflection plane σ or any one of the rotation-reflections S_{2n} of even indices renders a set achiral. By analogy with chirality, symmetry deficiency and various measures of symmetry deficiency can be defined more generally, for an arbitrary collection of point symmetry elements.

Consider a family $\mathrm{R} = \{R_1, R_2, \ldots, R_m\}$ of point symmetry elements. We shall use the term R-set for a set U of an Euclidean space E^n if set U has all point symmetry elements of family R. Set V of an Euclidean space E^n is an R-deficient subset of E^n if V has none of the point symmetry elements of family R. Some basic properties of R-sets and R-deficient sets are listed in the appendix.

The R-*deficiency index* $i_0(C, \mathrm{R})$ of a contour C is the smallest n_R number of cells at and above which all interior filling animals $A_i(C, n)$ of C are R-deficient:

$$i_0(C,\mathrm{R}) = \begin{cases} \min\{n_\mathrm{R} : \text{all } A_i(C,n) \text{ are R-deficient if } n \geq n_\mathrm{c}\}, & \text{if the minimum exists,} \\ \infty & \text{otherwise.} \end{cases} \tag{40}$$

The *degree of* R-*deficiency* $d(C, \mathrm{R})$ is defined as

$$d(C, \mathrm{R}) = 1/(i_0(C, \mathrm{R}) - 1). \tag{41}$$

For cell number $n = 1$, all possible symmetry elements of animals are present, hence $i_0(C, \mathrm{R}) > 1$ is always valid for all feasible choices of set R. This justifies the inclusion of the number 1 in the denominator.

At level n of resolution, we define three quantities:

$m(C, \mathrm{R}, n)$ = the minimum number of cells that must be moved in order to turn an interior filling animal $A_i(C, n)$ of C into an animal which is an R-set, (42)

$r(C, \mathrm{R}, n)$ = the minimum number of cells that must be removed in order to turn an interior filling animal $A_i(C, n)$ of C into an animal which is an R-set, (43)

$a(C, \mathrm{R}, n)$ = the minimum number of cells that must be added in order to turn an interior filling animal $A_i(C, n)$ of C into an animal which is an R-set. (44)

Three resolution-dependent R-imperfection measures are given by

$$\mathrm{im}(C, \mathrm{R}, n) = m(C, \mathrm{R}, n)/n, \tag{45}$$

$$\mathrm{ir}(C, \mathrm{R}, n) = r(C, \mathrm{R}, n)/n \tag{46}$$

and

$$\mathrm{ia}(C, \mathrm{R}, n) = a(C, \mathrm{R}, n)/n. \tag{47}$$

Note that if C has no n-cell interior filling animal, then the above six quantities $m(C; \mathrm{R}, n)$, $r(C, \mathrm{R}, n), \ldots, \mathrm{ia}(C, \mathrm{R}, n)$ are not defined.

Taking limits as the number of cells grows to infinity, one obtains the infinite resolution ("resolution-independent") R-imperfection measures. These measures are related to various areas.

If $a(C)$ is the area enclosed by the contour C and $C'(C, \mathrm{R})$ is an R-set of area equal to that of C, obtained from C by minimal deformation as defined by the maximal overlap measure between C and $C'(C, \mathrm{R})$, then

$$\lim_{n\to\infty} \mathrm{im}(C,\mathrm{R},n) = \big(a(C) - a(C \cap C'(C,\mathrm{R}))\big) / a(C). \tag{48}$$

If $M(C)$ and $N(C)$ are maximal area R-subset and minimal area R-superset, respectively, of the set of contour C, then

$$\lim_{n\to\infty} \mathrm{ir}(C,\mathrm{R},n) = \big(a(C) - a(M(C))\big) / a(C), \tag{49}$$

$$\lim_{n\to\infty} \mathrm{ia}(C,\mathrm{R},n) = \big(a(N(C)) - a(C)\big) / a(C). \tag{50}$$

The R-deficiency index $i_0(G, \mathrm{R})$ of a contour C and the degree of R-deficiency $d(C, \mathrm{R})$, as well as the definitions of the resolution-dependent and resolution-independent R-imperfection measures can be generalized for the three-dimensional case of bodies and polycubes, or for the case of abstract, υ-dimensional bodies and polyhypercubes, by replacing the terms and symbols of animal, A, boundary, C, and cell with polycube, P, contour surface, G, and cube, or with polyhypercube, P, contour hypersurface, G, and hypercube, respectively.

Appendix

Take a set T of the plane such that

(i) T is simply connected,

(ii) T has a finite area $a(T)$,

(iii) T has a perimeter P of finite length $p(T)$,

(iv) $P = P(t)$ is a parametric curve with $0 \le t \le 1$ such that for $t < t'$, $P(t) = P(t')$ if and only if $t = 0$ and $t' = 1$. (T is nowhere "infinitely thin". For practical purposes, we shall assume that there are no "bottlenecks" in T narrower than $2\sqrt{2}\, s'$ for a small constant s').

DEFINITION A1

M' is a maximal achiral subset of T if M' is achiral, $M' \subset T$ and if no achiral set M'' exists such that $M' \subset M''$, $M' \neq M''$, and $M'' \subset T$.

Note that M' is not necessarily unique for a given set T.

DEFINITION A2

M is a maximal area achiral subset of T if M is achiral, $M \subset T$ and if for all maximal achiral subsets M' of T, $a(M') \le a(M)$.

Note that for a given set T, set M is not necessarily unique either; however, the area $a(M)$ is a unique number for each T. Evidently, if T is achiral, then M is unique and $M = T$.

DEFINITION A3

N' is a minimal achiral superset of T if N' is achiral, $T \subset N'$ and if no achiral set N'' exists such that $N'' \subset N'$, $N' \neq N''$, and $T \subset N''$.

Note that N' is not necessarily unique for a given set T.

DEFINITION A4

N is a minimal area achiral superset of T if N is achiral, $T \subset N$ and if for all minimal achiral supersets N' of T, $a(N) \le a(N')$.

Note that, for a given set T, set N is not necessarily unique either; however, the area $a(N)$ is a unique number for each T. Evidently, if T is achiral, then N is unique and $N = T$.

The actual determination of a set M for some chiral set T and the calculation of area $a(T)$ are rather difficult problems (see some relevant comments in refs. [16–19]). However for the determination of an RBSM and the analogous chirality measures given in terms of a discretization procedure using lattice animals and polycubes, there is no need to determine a maximal area achiral subset M and to calculate its exact area $a(M)$. We shall show that the determination of any upper bound a such that

$$a(M) \leq a < a(T) \tag{A.1}$$

is sufficient for our purposes. We may think of the number a as the area $a = a(M_c)$ of a chiral subset M_c of T which contains M.

As a consequence of restriction (iv), for any chiral T, the relations $a(M) < a(T)$ and $a(M_c) < a(T)$ hold. That is, the differences

$$a' = a(T) - a(M) > 0 \tag{A.2}$$

and

$$a'' = a(T) - a > 0 \tag{A.3}$$

are positive.

Clearly, any interior filling animal $A_i(T, n)$ of T is chiral if

$$a(A_i(T, n)) > a. \tag{A.4}$$

We shall find a large enough n at and beyond which this is guaranteed.

Define a cell size (length of the side of the square) s' as

$$s' = a''/(bp(T)), \tag{A.5}$$

where $p(T)$ is the finite length of the perimeter P of T and $b \geq 2\sqrt{2}$. If within T there are "bottlenecks" narrower than $2\sqrt{2}\, s'$, then we replace s' by $s'/2$ and repeat the test of bottlenecks.

For any chiral set T, this cell size is positive, $s' > 0$. We shall investigate how well the set M_c is covered by an interior filling animal A of cell size s'. The local misalignment of the perimeter P and the lattice directions, the actual placement of A within T, as well as narrow foldings of the perimeter may render some areas of T near the perimeter inaccessible to animal cells. All these areas are necessarily contained within a belt of width $s'2\sqrt{2}$ along the perimeter, and their total area is less than the area of the belt, hence the inaccessible area has an upper bound of $p(T)s'2\sqrt{2}$. Consequently,

$$a(A) > a(T) - p(T)s'2\sqrt{2} = a(T) - a''2\sqrt{2}\,/\,b \geq a(T) - a'' = a(M_c) \geq a(M) \tag{A.6}$$

and any interior filling animal A of cell size s' or less is chiral.

This result evidently provides a constraint in terms of cell numbers. Take

$$n' = \mathrm{int}(a(T)/(s')^2) + 1, \tag{A.7}$$

then any n-cell interior filling animal $A_i(T, n)$ of T is chiral if $n \geq n'$.

Consequently, for any chiral set T of properties (i)–(iv) there exist finite n_c numbers at and above which all interior filling animals are chiral, hence there exists a unique chirality index $n_\chi(T)$ that is the smallest such n_c number. Since for an achiral set T one has $M = T$, no finite n_c and $n_\chi(T)$ numbers exist for achiral sets T.

The above results provide explicit proof of an assertion in ref. [11]. They also form the basis of a procedure for the determination of the chirality index $n_\chi(T)$ of actual sets T. Note an important advantage of the *discrete* technique of representing the set T by n-cell interior filling animals of *finite* n: for each n value, *there are only a finite number* m_n *of animals to be tested* for chirality and for their occurrence as interior filling animals for the given set T.

Since $a(M_c) \geq a(M)$, hence $n' \geq n_\chi(T)$. Consequently, it is possible that with a poor choice for n', more animals must be tested than in the optimal case of $n' = n_\chi(T)$. Note, however, that the possibility of extra tests is more than compensated for by the advantage that there is *no need for determining* M *and* $a(M)$ *precisely*, and as long as a suitable estimate for $a(M_c)$ is available, the lower bound s' for cell size and the upper bound n' of (A.7) for the number of cells are valid. By testing at most a finite number

$$m(n') = \sum_{n=4,n'} m_n \tag{A.8}$$

of animals for chirality and for their occurrence as interior filling animals for the given set T, the actual chirality index $n_\chi(T)$ can be determined.

Whether an animal A is chiral or not can be easily deduced from its matrix representations, as given in ref. [11]. To determine the chirality index $n_\chi(T)$, one may follow the procedure below.

(i) Initialization by setting $n = n' + 1$.

(ii) Set $n = n - 1$. For each of the m_n n-cell animals $A_i(n)$, determine the maximum cell size $s_i(T, n)$ compatible with fitting $A_i(n)$ within T. Denote the maximum of these cell sizes by $s(T, n)$:

$$s(T,n) = \max_{i=1,m_n} \{s_i(T,n)\}. \tag{A.9}$$

If $n = n'$, then return to the beginning of step (ii). Animal $A_i(n)$ is an n-cell interior filling animal $A_j(T, n)$ of set T if and only if $s_i(T, n) > s(T, n + 1)$.

(iii) Test each of the obtained n-cell interior filling animals for chirality. If there is an achiral animal among them, then the chirality index is found,

$$n_c(T) = n + 1, \tag{A.10}$$

and the procedure is completed. Otherwise, return to step (ii).

In practice, in order to determine whether a given animal A is an interior filling animal, the fitting of each animal $A_i(n)$ within T is carried out by some approximate method. One may choose a positive grid size factor g, $0 < g < 1$, a positive angle increment α, and a translation increment f. Choose a large enough initial grid size s such that the animal $A_i(n)$ certainly does not fit within T. For fitting $A_i(n)$ within T, one may follow the steps below:

(i) Reduce the current grid size by multiplying it by g: set $s = gs$. Set the initial rotation angle as $\beta = -\alpha$.

(ii) Set $\beta = \beta + \alpha$. If $\beta \geq 2\pi$, then return to step (i). Rotate T on the square grid by angle β.

(iii) Generate a sub-grid of sub-cell size f within a cell size s, and a series of translation vectors v_k from a chosen vertex of the cell to each subgrid point. Apply each translation vector v_k to T and test for each whether animal $A_i(n)$ occurs within T. If yes, then we have an approximation

$$s_i(T, n) = s \tag{A.11}$$

and the procedure is completed. Otherwise, return to step (ii).

By choosing factor g closer to 1, and both α and f closer to 0, one may improve the approximation (A.11) as desired, at the expense of carrying out a larger number of tests.

In some cases (e.g. in the case of molecules in an external field), the relative orientation of the objects to one another or to some external direction plays an important role. The relative orientations of the Jordan curves or contours with respect to the grid can be fixed, leading to *oriented similarity measures* and *oriented symmetry deficiency measures*. In such cases, the reorientation step (ii) in the above optimization procedure is omitted.

The analogous 3D method applies to polycubes. A similar treatment applies for more general symmetry deficiency problems, and below we list some relevant definitions. The generalization of these definitions and properties to any finite n-dimensional symmetry problems in an Euclidean space E^n is straightforward, by considering the symmetry elements in E^n and simply replacing area with the n-dimensional volume. We shall describe in detail the two-dimensional case, although some of the notations and concepts (e.g. the point symmetry elements S_{2n}) will refer to the three-dimensional case.

Consider a family $\mathrm{R} = \{R_1, R_2, \ldots, R_m\}$ of point symmetry elements. We recall from the general text of this study that an R-set of an Euclidean space E^n is a set that has all point symmetry elements of family R, whereas an R-deficient set of an Euclidean space E^n is a set that has none of the point symmetry elements of family R.

DEFINITION A5

Set M' is a maximal R-subset of T if M' is an R-set, $M' \subset T$ and if no R-set M'' exists such that $M' \subset M''$, $M' \neq M''$, and $M'' \subset T$.

Note that M' is not necessarily unique for a given set T.

DEFINITION A6

Set M is a maximal area R-subset of T if M is an R-set, $M \subset T$ and if for all maximal R-subsets M' of T, $a(M') \leq a(M)$.

Note that M is not necessarily unique for a given set T; however, the area $a(M)$ is a unique number for each T. Evidently, if T is an R-set, then M is unique and $M = T$.

DEFINITION A7

Set N' is a minimal R-superset of T if N' is an R-set, $T \subset N'$ and if no R-set N'' exists such that $N'' \subset N'$, $N' \neq N''$, and $T \subset N''$.

Note that N' is not necessarily unique for a given set T.

DEFINITION A8

Set N is a minimal area R-superset of T if N is an R-set, $T \subset N$ and if for all minimal R-supersets N' of T, $a(N) \leq a(N')$.

Note that N is not necessarily unique for a given set T; however, the area $a(N)$ is a unique number for each T. Evidently, if T is an R-set, then N is unique and $N = T$.

DEFINITION A9

Set M' is a maximal R-deficient subset of T if M' is an R-deficient set, $M' \subset T$ and if no R-deficient set M'' exists such that $M' \subset M''$, $M' \neq M''$, and $M'' \subset T$.

Set M' is not necessarily unique for a given set T.

DEFINITION A10

Set M is a maximal area R-deficient subset of T if M is an R-deficient set, $M \subset T$ and if for all maximal R-deficient subsets M' of T, the relation $a(M') \leq a(M)$ holds.

Set M is not necessarily unique for a given set T; however, the area $a(M)$ is a unique number for each T. If T is an R-deficient set, then M is unique and $M = T$.

DEFINITION A11

Set N' is a minimal R-deficient superset of T if N' is an R-deficient set, $T \subset N'$ and if no R-deficient set N'' exists such that $N'' \subset N'$, $N' \neq N''$, and $T \subset N''$.

Set N' is not necessarily unique for a given set T.

DEFINITION A12

Set N is a minimal area R-deficient superset of T if N is an R-deficient set, $T \subset N$ and if for all minimal R-deficient supersets N' of T, the relation $a(N) \leq a(N')$ holds.

Set N is not necessarily unique for a given set T; however, the area $a(N)$ is a unique number for each T. If T is an R-deficient set, then N is unique and $N = T$.

Note that without further restrictions such as those of polycubes, R-deficiency can be achieved by infinitesimal changes.

For any pair of R-subset M' and R-superset N' of any set T, the relation

$$a(M') \leq a(N') \tag{A.12}$$

holds. Furthermore,

$$a(N) - a(M) \leq a(N') - a(M') \tag{A.13}$$

for any maximal area R-subset M, minimal area R-superset N, R-subset M', and R-superset N' of any set T.

In the 3D case, using the customary notation S_{2n} for rotation–reflection, special considerations apply. Note that if the family R contains a symmetry element of relection σ or one of the rotation-reflections S_{2n} of even indices, then the R-sets are achiral sets. The various extremal achiral sets can be generated by special R-sets which are extremal over all choices of families R containing at least one of the above point symmetry elements. If one uses subscripts α and R in order to distinguish achiral sets and R-sets, then for maximum achiral subsets M'_α and minimum achiral supersets N'_α of any given set T, the following holds:

Set M'_α is a maximal achiral subset of T if $M'_\alpha \subset M'_\mathrm{R}$, $M'_\mathrm{R} \subset T$, and $\sigma \in \mathrm{R}$ or $S_{2n} \in \mathrm{R}$ for some $n > 0$ and for some maximal R-subset M'_R imply that $M'_\alpha = M'_\mathrm{R}$.

Set N'_α is a minimal achiral superset of T if $N'_\mathrm{R} \subset N'_\alpha$, $T \subset N'_\mathrm{R}$, and $\sigma \in \mathrm{R}$ or $S_{2n} \in \mathrm{R}$ for some $n > 0$ and for some minimal R-superset N'_R imply that $N'_\alpha = N'_\mathrm{R}$.

If $M_\alpha, M_\mathrm{R}, N_\alpha$ and N_r are maximal area achiral subset, maximal area R-subset, minimal area achiral superset and minimal area R-superset of a set T, respectively, then

$$a(M_\alpha) = \max_\mathrm{R}\{a(M_\mathrm{R}) : M_\mathrm{R} \subset T, \sigma \in \mathrm{R} \text{ or } S_{2n} \in \mathrm{R} \text{ for } n > 0\} \tag{A.14}$$

and

$$a(N_\alpha) = \min_\mathrm{R}\{a(N_\mathrm{R}) : T \subset N_\mathrm{R}, \sigma \in \mathrm{R} \text{ or } S_{2n} \in \mathrm{R} \text{ for } n > 0\}. \tag{A.15}$$

In fact, the extremum properties may be stated in terms of general R-subsets M''_R and R-supersets N''_R:

$$a(M_\alpha) = \max_{\mathrm{R},M''_\mathrm{R}}\{a(M''_\mathrm{R}) : M''_\mathrm{R} \subset T, \sigma \in \mathrm{R} \text{ or } S_{2n} \in \mathrm{R} \text{ for } n > 0\} \tag{A.16}$$

and

$$a(N_\alpha) = \min_{\mathrm{R},N''_\mathrm{R}}\{a(N''_\mathrm{R}) : T \subset N''_\mathrm{R}, \sigma \in \mathrm{R} \text{ or } S_{2n} \in \mathrm{R} \text{ for } n > 0\}. \tag{A.17}$$

All the definitions, concepts and procedures listed in this appendix have straightforward generalizations for any finite dimension n. In this context, we must emphasize that chirality is obviously dimension dependent. If a given object is achiral when embedded in a space of k-dimensions, it may be chiral if embedded in a space of some different dimensions. We shall use the following notations: E^{n+1} is an $(n+1)$-dimensional Euclidean space and E^n is an n-dimensional subspace of E^{n+1}. If we refer to the n-dimensional chirality of an object A, then we consider its embedding in an Euclidean space E^n and reflections as well as all motions are restricted to this space. Here, we present a simple proof of the following result:

Any object A that is chiral in n-dimensions is achiral in $(n+1)$-dimensions and in any higher dimensions. Chirality may occur only if the lowest dimension A is embeddable.

Proof

Object A is chiral in n-dimensions (that is, when embedded in E^n). Let us denote, the mirror image of A by $A^\lozenge$ and the corresponding mirror image of point $p \in A$ by $p^\lozenge$. By translations and rotation, we can always arrange A and $A^\lozenge$ in E^n so that for all their point pairs p and $p^\lozenge$ their coordinates fulfill the relations

$$p_1^\lozenge = -p_1, \tag{A.18}$$

$$p_i^\lozenge = p_i \qquad (i = 2, 3, \dots, n). \tag{A.19}$$

For this arrangement, the $(n-1)$-dimensional reflection hyperplane E^{n-1} in E^n is defined by

$$x_1 = 0, \tag{A.20}$$

where x_1 is the first coordinate of a point $x \in E^n$.

Consider now the same arrangement of A and $A^\lozenge$ embedded in E^{n+1}, by regarding E^n as a subspace of E^{n+1}. A two-dimensional rotation in E^{n+1} is defined by its $(n-1)$-dimensional axis and by the angle of rotation in the remaining two dimensions. Note that in a k-dimensional space, the axis of rotation is $(k-2)$-dimensional. Choose the rotation axis in E^{n+1} as the $(n-1)$-dimensional subset defined as the reflection hyperplane E^{n-1} of condition $x_1 = 0$ in E^n. With respect to this axis, a rotation of angle $\alpha = \pi$ in the two-dimensional plane spanned by coordinates (x_1, x_{n+1}) superimposes A on $A^\lozenge$ in $(n+1)$-dimensions. Consequently, the object A is achiral in $(n+1)$-dimensions (that is, when embedded in E^{n+1}). Furthermore, the superimposition of mirror images performed in E^{n+1} is a possible motion in any Euclidean space E^{n+k}, $k > 1$, of which E^{n+1} is a subspace, hence A is achiral in any higher dimensions. Consequently, chirality may occur only in the lowest dimension where A is embeddable. □

Acknowledgements

The author is grateful to Professor Frank Harary, Professor Stuart Whittington, and Professor Chris Soteros for stimulating conversations on lattice animals and polycubes. Special thanks are due to Professor Kurt Mislow for illuminating discussions and correspondence on chirality measures and for pointing out an error in a table of a manuscript. The hospitality of Professor Jacques-Emile Dubois of the Institute of Topology and System Dynamics, Paris, France, and of Dr. Jean Maruani, SMOA, CNRS, Paris, France, as well as stimulating discussions with Professor A. Rassat, ENS, Paris, France, during a sabbatical visit of the author are gratefully acknowledged. This work was supported in part by the Computational Chemistry Unit of the Upjohn Laboratories, Kalamazoo, Michigan, and by both strategic and operating research grants from the Natural Sciences and Engineering Research Council of Canada.

References

[1] S. Golomb, *Polyominoes* (Scribner's, New York, 1965).

[2] F. Harary, The cell growth problem and its attempted solutions, Beitrage zur Graphentheorie, Teubner, Leipzig (1968)49.

[3] G. Exoo and F. Harary, Indian Nat. Acad. Sci. Lett. 10(1987)67.

[4] F. Harary and M. Lewinter, Int. J. Comput. Math. 25(1988)1.

[5] C. Soteros and S.G. Whittington, J. Phys. A21(1988)2187.

[6] N. Madras, C. Soteros and S.G. Whittington, J. Phys. A21(1988)4617.

[7] F. Harary and E.M. Palmer, *Graphical Enumeration* (Academic Press, New York, 1973).

[8] M. Gardner, Sci. Amer. 240(1979)18.

[9] M.A. Johnson and G.M. Maggiora (eds.), *Concepts and Applications of Molecular Similarity* (Wiley, New York, 1990).

[10] P.G. Mezey, J. Math. Chem. 7(1991)39.

[11] F. Harary and P.G. Mezey, in: *New Developments in Molecular Chirality*, ed. P.G. Mezey (Kluwer Academic, Dordrecht, 1991), pp. 241–256.

[12] P.G. Mezey, in: *New Developments in Molecular Chirality*, ed. P.G. Mezey (Kluwer Academic, Dordrecht, 1991), pp. 257–289.

[13] A.I. Kitaigorodskii, *Organic Chemical Crystallography* (Consultants Bureau, New York, 1961), p. 230.

[14] G. Gilat and L.S. Schulman, Chem. Phys. Lett. 121(1985)13.

[15] G. Gilat, J. Phys. A22(1989)L545.

[16] A.B. Buda and K. Mislow, Elem. Math. 46(1991)65.

[17] A.B. Buda and K. Mislow, J. Mol. Struct. (THEOCHEM) 232(1991)1.

[18] A.B. Buda, T.P.E. Auf der Heyde and K. Mislow, J. Math. Chem. 6(1991)243.

[19] T.P.E. Auf der Heyde, A.B. Buda and K. Mislow, J. Math. Chem. 6(1991)255.

[20] A. Rassat, Compt. Rend. Acad. Sci. (Paris), Ser. II, 299(1984)53.

[21] B.R. Gaines, Int. J. Man–Mach. Stud. 8(1976)623.

[22] K. Mislow and P. Bickart, Isr. J. Chem. 15(1976)1.

[23] P.G. Mezey and J. Maruani, Mol. Phys. 69(1990)97.

[24] F. Harary and P.G. Mezey, Theor. Chim. Acta 79(1991)379.