Journal of Mathematical Chemistry Vol. 30, No. 4, November 2001 (© 2002)

Metric properties of factor space of molecular shapes *

A. Frolov [a], E. Jako [b] and P.G. Mezey [c,**]

[a] *Moscow Power Engineering Institute (Technical University), Chusovskaya, 11-7-8, Moscow 107203, Russia*
E-mail: frolovab@mpei.ru

[b] *Eötvös Loránd University, Department of Plant Taxonomy and Ecology, Tompa u. 17, Budapest 1211, Hungary*
E-mail: jakoeena@hotmail.com

[c] *Department of Chemistry and Department of Mathematics and Statistics, University of Saskatchewan, 110 Science Place, Saskatoon, SK, S7N 5C9 Canada*
E-mail: mezey@sask.usask.ca

Received 26 May 2001

Dedicated to the 80th birthday of Professor Frank Harary

Molecular shape equivalence classes defined with respect to equivalence of geometrical and topological properties are represented by logical models. Consequently, the factor space of molecular shapes is provided by a metric useful in shape comparisons.

KEY WORDS: molecular shape, logical models, Boolean n-cube, pattern recognition

1. Introduction

In the previous article [1] we have discussed the theory of logical models of molecular shape representations. This theory is based on classical and quantum chemistry approaches to modeling of molecular shape [2], on topology and differential geometry [3,4], and on theory of logical diagnostics [5–7]. It explores the fact that the topology of a molecular shape is based on the partitioning of the molecular surface on two-, one-, and zero-dimensional subsets with a finite set of properties. These subsets interrelate with each other through known relations. This allows one implementation of the finite topology principle based on the method of logical modeling introduced in the previous article. The subset of these partitions, as a base, defines a finite topology, containing a lower Boolean sub-lattice of this topology, corresponding to a Boolean n-cube as a domain of the proposed logical model. A logical function can be obtained that reflects the properties of the topological domains as well as the interrelations on the set of do-

* This article, dedicated to the 80th birthday of Professor Frank Harary, has been written as part of an Institute for Advanced Study, Collegium Budapest Fellowships scientific project: P.G. Mezey (1999–2000), E. Jako, A. Frolov (2000–2001).

** Corresponding author.

mains. Based on classical or quantum-chemical representations of molecular shape, as it was shown in the previous article, these models allow one the implementation of methods of logical diagnostics [5–7] in chemistry, and the definition of a metric on the set of molecular shape equivalence classes avoiding numerical characterization and explicit embedding of objects into a vector space. The families of molecular shapes can be considered as sets of logical models or as a united, more complex logical model.

In the present article, we focus on the metric properties of the factor space of shape equivalence classes. In the following section, the new discrete mathematical concepts related to the theory of logical modeling of molecular shapes, proposed in [1] are briefly recalled.

2. Method of metric and semimetric on the set of functional models

2.1. Logical models

A logical model $f(x_1, \ldots, x_n)$ is a function from a finite Cartesian product

$$X_1 \times \cdots \times X_n$$

of variable values to a finite set Y of function values:

$$f : X_1 \times \cdots \times X_n \to Y.$$

Function *variables* correspond to elements of object structure. Function *values* reflect properties and interrelation of these elements.

As an example, logical diagnostics models [5,7] and Boolean models of discrete N-dimensional systems for ecological studies [8] can be considered. In the case of molecular shape, object structure is defined as a set or a subset of the set of two-dimensional domains from shape partition. These domains are considered as structural elements. Their properties defined as local curvature properties, their interrelation depends on neighboring conditions. Domain of a functional model is Boolean n-cube with atoms corresponding to the structural elements.

Notice that as functional models one may consider many earlier discrete mathematical models, for example, shape matrices and shape codes [2]. Indeed, in the case of shape matrix there are two variables that correspond to structural elements (shape domains). Function value reflects curvature indexes (from matrix diagonal elements) or neighboring interrelation. In the case of shape code, there is only one variable that corresponds to positions of shape code elements, the function values correspond to Betty numbers in these positions.

Shape families can be represented as finite functions from a space of parameter values to the set of shape descriptors. For example, (a, b)-maps of the shape groups can be considered as such functions.

2.2. Logical models of topological spaces with finite topology

Recall that logical model of a topological space of a finite topology can be constructed taking into account properties of base sets as well as their interrelations [1].

Suppose that each base set has one of the finite number of properties $\{0, 1, 2, \ldots, K-1\}$. Moreover, assume, that the base sets participate in the symmetric relations

$$\rho_0^{m_1}, \ldots, \rho_i^{m_i}, \ldots, \rho_{S-1}^{m_{s-1}}$$

where a relation, $\rho_i^{m_i}$, is a symmetric relation if, together with any m_i-ordered set (m_i-tuple), it contains any of its permutations. (We assume that the system of these relations is orthogonal, i.e., each m_i-tuple participate in only one m_i-dimensional relation). In addition, it should be noted that the mentioned properties and relations should be defined based on the real properties of the geometrical or discrete structures of the modeled objects. Also, the properties of all the elements from the base set of the topology, as well as the relations on the set of base elements, are numbered $0, \ldots, K-1$ and $0, 1, \ldots, S-1$, respectively.

The information about geometry, as well as about the properties and interrelations of the elements from the base of topology can be represented as a logical function $f(x_1, \ldots, x_N)$ or as an equivalent system of Boolean functions.

We use the Boolean N-cube B_N isomorphic to lower Boolean sublattice o a finite topology as a field of definition (domain) of a logical function $f(x_1, \ldots, x_N)$, $f : B_N \to \{0, 1, \ldots, k-1\}$, $k = \max(K, S)$. This function maps the elements from the base of the topology to the values defining their properties from the set $\{0, 1, \ldots, K-1\}$ of numbers of the properties. Later, if the sets $A_1, \ldots, A_t$ from the base of the topology belong to relation

$$\rho_t^{m_t},$$

then the function takes the topology element corresponding to the union of these base sets to the value t from the set $\{0, 1, \ldots, S-1\}$. Finally, the function f takes all the remaining elements of Boolean N-cube B_N to the zero value.

The corresponding examples were given in previous paper [1].

2.3. Logical equivalence relation on topological spaces

Let us consider two finite partitions Π_1, Π_2 of the sets X_1, X_2. We assume that these partitions are chosen taking into account some natural properties of real objects, denoted as sets X_1, X_2. Consider that the equivalence relations corresponding to these partitions are isomorphic under a given isomorphism $\varphi : X_1 \to X_2$. Hence, there exists a bijection $\psi_\varphi : \Pi_1 \to \Pi_2$ such that for all elements x of the subset $A \in \Pi_1 \to \varphi(x) \in \psi_\varphi(\Pi_1)$. Let T_1, T_2 be topologies corresponding to isomorphic (with respect to isomorphism ψ_φ) subsets of two partitions Π_1, Π_2. These topological spaces are referred to as *isomorphic topological spaces*. Furthermore, we assume that the elements of bases B_i, $B_1 \subseteq T_1$, $B_2 \subseteq T_2$ possess some properties and participate in symmetrical relations reflecting the natural properties and interrelations of corresponding parts of the sets X_1, X_2. If these properties and relations are isomorphic under the same isomorphism ψ_φ, assuming that isomorphic properties and relations possesses the same physical sense, we refer the isomorphic topological spaces (X_1, T_1) and (X_2, T_2)

with finite topologies (as well as the corresponding natural objects) as *equivalent with respect to system of properties and relations under isomorphism* φ.

Two topological spaces (X_1, T_1) and (X_2, T_2) with finite topologies are called equivalent with respect to a given system of properties and relations (they are logically equivalent) if they are *equivalent with respect to the system of these properties and relations under some isomorphism* φ.

By definition, the following statement is valid:

Statement 2.1. Two objects X_1 and X_2 are equivalent with respect to the system of properties and symmetric relations on the topology base under given isomorphism $\varphi : X_1 \to X_2$ if and only if two logical functions defined using this system coincide.

2.4. Test sets and logical diagnoses

Let as consider N distinct vectors $(v_1, \ldots, v_j, \ldots, v_n)$ describing feature values of objects $o_1, \ldots, o_i, \ldots, o_m$. We assume, that these values are taken from finite sets V_j, $j = 1, \ldots, n$ of possible feature values. The diagnostics problem can be formulated as follows: define a minimal subset $\{j_1, \ldots, j_s\}$ of the set $\{1, \ldots, j, \ldots, n\}$ of features that allows distinguishing the objects and describe their mutual differences. Notice that a solution is ambiguous because, in general case, there exist a number of minimal sets. Nevertheless, this problem is most important for practical applications in different fields such as medicine, logic design, technical diagnostics etc.

Here we consider the general approach to solution of the diagnostics problem following the ideas proposed by Chegis and Yablonsky [5].

Let the vectors of feature values be considered as a functions f_i, $i = 1, \ldots, n$ from the set $\{1, \ldots, j, \ldots, n\}$ of features to a finite set $V = \bigcup_{j=1}^{n} V_j$ of possible feature values ($f_i\ (j) \in V_j$). These functions we consider as functional descriptors of objects $o_1, \ldots, o_i, \ldots, o_m$.

Let N be the set of all function pairs (f_p, f_t), $p \neq t$, $p, t \in \{1, \ldots, m\}$. A minimal subset $T = \{j_1, \ldots, j_s\} \subseteq \{1, \ldots, n\}$ of these features is called a *minimal test set* for the set $\{f_1, \ldots, f_i, \ldots, f_m\}$ of functions if for each pair (f_p, f_t), $p \neq t$, $p, t \in \{1, \ldots, m\}$ there exists a feature $j \in T$ such that $f_p(j) \neq f_t(j)$. The set of all pairs $(f_p(j) f_t(j))$ $f_p(j) \neq f_t(j)$, $j \in T$ is called a *diagnosis* corresponding to the test set T. Such a diagnosis can be interpreted in terms of feature values. To calculate the minimal test sets, it is convenient to calculate the discriminate functions $f_{p,t}$ such that

$$f_{p,t}(j) = \begin{cases} 1 & \text{if } f_p(j) \neq f_t(j), \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Making use of the above discriminate functions it is possible to construct a binary table $t = (t_{j,k})$ with a size $n \times ((m-1)(m-2)/2)$, where $t_{j,k} = f_{p,t}(j)$, k is the number of pairs (p, t) under lexicographic (alphbetical) ordering. The minimum test sets correspond to minimum coverings of the binary table by rows. (Recall that a row j *covers* a column s if $f_{j,s} = 1$. The subset T of rows covers the binary table if each column is covered

Table 1
Functions $f_a(x_1, x_2, x_3)$, $f_b(x_1, x_2, x_3)$, and $f_c(x_1, x_2, x_3)$, corresponding to discriminate functions $f_{ab}(x_1, x_2, x_3)$, $f_{ac}(x_1, x_2, x_3)$, $f_{ac}(x_1, x_2, x_3)$ and a metric w on the set of functions.

J	f_a	f_b	f_c	f_{ab}	f_{ac}	f_{bc}
0	0	0	0	0	0	0
1	3	3	3	0	0	0
2	3	3	3	0	0	0
3	0	0	1	0	1	1
4	3	3	3	0	0	0
5	1	1	1	0	0	0
6	1	1	1	0	0	0
7	0	1	2	1	1	1
w				1	2	2

by some row from the set T.) For calculation of minimum covering, one may use the known logical methods [3,5–7,9,10].

Example 2.1. Three functions f_a, f_b, f_c from the set $\{0, 1, 2, 3, 4, 5, 6, 7\}$ of features to the set $V = \{0, 1, 2, 3\}$ of feature values are given in table 1 with corresponding discriminate functions f_{ab}, f_{ac}, and f_{bc}. The minimum covering $\{8\}$ consists of a unique row, that is the test set $T = \{9\}$, and the diagnosis can be represented then as the set of pairs

$$\big(f_a(7), f_b(7)\big) = (0, 1),$$
$$\big(f_a(7), f_c(7)\big) = (0, 3),$$
$$\big(f_b(7), f_c(7)\big) = (2, 3).$$

Notice that as logical models of objects, this approach allows us to implement the functions of more than one variables. In this case, the sets of elementary feature values are considered as structural feature values. For example, instead of feature values $1, 2, \ldots, 9$ from example 2.1. one may consider structural feature values (0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0), (1,1,1) (table 2) and corresponding functions of three variables.

We can conclude that to distinguishing the structural peculiarities of the considered type of objects, one should:

(1) Implement the above modelling method to construct the logical models $f_i(x_1, \ldots, x_n)$, $i = 1, 2, \ldots, N$, for all the study objects under consideration;

(2) Define the discriminate functions $f_{i,j}$, $i, j \in N$, $i \neq j$;

(3) Construct the binary table where the rows correspond to the set of binary n-tuples $(a_1, \ldots, a_n)$ and at most $N(N-1)/2$ columns represent nonzero discriminate functions;

Table 2
VDWS functions $f_a(x_1, x_2, x_3)$, $f_b(x_1, x_2, x_3)$, $f_c(x_1, x_2, x_3)$ and VDWS discriminate functions $f_{ab}(x_1, x_2, x_3)$, $f_{ac}(x_1, x_2, x_3)$, $f_{ac}(x_1, x_2, x_3)$ for the VDWSs in figure 6(a,b,c).

	f_a	f_b	f_c	f_{ab}	f_{ac}	f_{bc}
000	0	0	0	0	0	0
001	3	3	3	0	0	0
010	3	3	3	0	0	0
100	0	0	1	0	1	1
100	3	3	3	0	0	0
101	1	1	1	0	0	0
110	1	1	1	0	0	0
111	0	1	2	1	1	1
w				1	2	2

(4) Derive the minimum coverings of this table by choosing the minimum number of rows such that any column will contain the value 1 in at least one selected row;[1]

(5) Interpret these coverings in terms of the objects under consideration.

2.5. A metric and semimetric on the set of objects

Defining positive numerical functional on the set of discriminate functions, one may define metric or semimetric on the set of original objects. For example, such a functional can be introduced as a weight function

$$w(f) = \sum_{j=1}^{n} f(j) \tag{2.2}$$

or, in the multivariable binary case,

$$w_{i.j=} \sum_{(a_1,\ldots,a_n)\in\{0,1\}^n} f_{i,j}(a_1, \ldots, a_n) \tag{2.2a}$$

providing definition of a metric on the set objects (table 1).

The *weight* or other positive functional on the set of discriminate functions can be considered as a *distance* between logical functions and, as a corollary, as a distance between corresponding real objects. It is easy to see that the function $d(x, y)$ defined on the set of functional models $f_i(x_1, \ldots, x_n)$ and possessing values

$$\begin{aligned} &d(f_i, f_j) = w_{i,j} \text{ satisfies the metric properties (metric axioms)} \\ &d(f_i, f_i) \geqslant 0, \quad d(f_i, f_i) = 0 \Leftrightarrow (f_i = f_i), \end{aligned} \tag{2.2}$$

[1] This step correspond to the NP complete *binary table covering* problem [10,11]. Therefore a practical implementation of the proposed method should make use of approximate algorithms.

$$d(f_i, f_j) = d(f_j, f_i), \tag{2.3}$$

$$d(f_i, f_j) + d(f_j, f_k) \leqslant d(f_i, f_k). \tag{2.4}$$

The properties (2.2) and (2.3) are called the semimetric axioms. Implementing other functional on the set of discriminate functions, one may lose the triangle property (2.4). If all metric axioms are satisfied then a metric is defined on the set of objects, if axiom (2.4) fails, then a semimetric is defined.

The proposed logical models allow one the investigation, comparison and diagnostics of objects with geometrical and topological properties that can be described as a system of symmetrical relations on a topology base for an appropriate topological space (X, T) with finite topology T.

For this purpose, one should be able to define the set X and its finite partition Π, and be able to choose the topology base sets from Π and denote, using numbers, their properties and interrelations reflecting essential natural peculiarities. From this data, the logical description of the corresponding real object can be obtained. These descriptions can be used providing implementation of known methods of data analysis and logical diagnostics. It allows the definition of the metric or semimetric on the set of logical models that can be used as a metric or semimetric on the set of real object equivalence classes.

Notice, that for definition of the logical model of an object, we should take into account the concrete bijection between the base of topology and the set of variables. Hence, comparing real objects, we assume that their logical models are defined implementing isomorphic (with respect to concrete isomorphism) topological spaces that can differ only in properties and relations on topology bases. At first glance, other real objects may seem to be incomparable. Nevertheless, by introducing fictitious variables, it is possible to make them comparable. Thus, if we consider two topological spaces (X_1, T_1) and (X_2, T_2) with finite topologies and their logically equivalent subspaces (X_1', T_1') and (X_2', T_2'), we can denote $f_1(x_1, \ldots, x_i, \ldots, x_j)$, $f_2(x_i, \ldots, x_j, \ldots, x_n)$, $f_1'(x_i, \ldots, x_j) = f_2'(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_n)$, as logical functions corresponding to properties and relations on the bases of topologies T_1, T_2, T_1', T_2'. With comparable functional descriptions of topological spaces (X_1, T_1) and (X_2, T_2) $f_1''(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_n)$, $f_2''(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_n)$ we can introduce fictitious variables $x_{j+1}, \ldots, x_n$ for the function $f(x_1, \ldots, x_i, \ldots, x_j)$, and fictitious variables $x_1, \ldots, x_{i-1}$ for the function $f_2(x_i, \ldots, x_j, \ldots, x_n)$. Notice, that the only situation where topological spaces (X_1, T_1) and (X_2, T_2), with nonisomorphic topology bases, are equivalent is when $f_1(x_1, \ldots, x_i, \ldots, x_j) = f_2(x_i, \ldots, x_j, \ldots, x_n) = c$, where c is a constant. Taking into account that all the functions take the n-tuple $(0, \ldots, 0)$ to the value 0, we conclude that $c = 0$. According to the proposed method of logical modeling, all objects corresponding to this trivial situation should be considered as logically equivalent objects. In order to make these functions mutually comparable, and to obtain a metric on the whole family, all the functions should have the same set of variables. Methods of reducing functions with the same set of variables are known. Notice that the distance $\mathrm{d}(f_1, f_2)$ allows the to find the number $\mathrm{d}(f_1, f_2)/(2_k)$ of binary

$(n - k)$-tuples which functions f_1 and f_2 take to distinct values (k is the number of variables that are fictitious variables for both functions) [9,10].

The proposed in this section method of logical modelling and definition of a metric or semimetric on the set of objects we call as *the finite topology principle.* The method of a metric or semimetric definition can be implemented independentely on this principle making use of the Chegis–Yablonsky approach to different discrete functional models of the same class.

The proposed method of metric or semimetric definition provides an implementation of novel metric classification algorithms [12–14]. Notice that in general case, the descriptors of objects as discrete functions cannot be considered as elements of a vector metric space. Indeed, discrete characterization of feature values can relate rather to qualitative characterization of objects then its quantitative measuring. The new approach allows us to define metric avoiding numerical characterisation of objects.

3. Logical models of molecular shapes

This section is devoted to logical models of molecular shapes and their families. It relates to the previous section in that the topology of a shape is based on the partitioning of a surface on two-, one-, and zero-dimensional subsets with finite set properties. These subsets interrelate with each other through known relations. This allows implementation of the finite topology principle, based on the method of logical modeling introduced in the previous section.

3.1. Logical models of isoproperty contour surfaces

Recall that for the characterization of the shapes of molecular contour surfaces, such as MIDCOs and MEPCOs, or an interpenetration of both, the surface is subdivided into domains satisfying some local shape criteria. This is discussed in detail in chapter 5 of the referred monograph [2]. These absolute or relative shape domains satisfy some geometrical or physical properties. For example, these domains may be thought as unbounded locally convex, locally concave, or locally-saddle-type subsets of a surface or as unbounded sets deleted from the shape. More exactly, domains D can be truncated by subdivision into three sets: the unbounded set C where $C \subset D$, the unbounded set $D' \subset D$, and the common boundary J of the sets C and D, i.e., $J = \mathrm{clos}(C) \cap \mathrm{clos}(D)$, where clos is the set-theoretical operation of closure. Set C is referred to as a *deleted* set and set D' is referred to as a *truncated* set.

If all the truncated domains, nontruncated domains, deleted sets, and one-dimensional boundaries (with the possible addition of the zero-dimensional boundaries of boundaries) are taken into account, a partition of the surface can be obtained. This partition satisfies all the conditions required by the proposed in previous article [1] logical modeling method: by considering the molecular surface before deletions as a set X, the partitions act as subsets of X. Depending on the goal of modeling, an appropriate subset of these subsets can be chosen as a base for topology T. For example, all these

subsets can be chosen as a base, and in general, choosing all of the domains and deleted sets as a base proves to be most useful. Yet, regardless of which base is chosen, the elements of the topology base interrelate and possess some finite number properties. Finally, it should be noted that the same arguments apply to the adequateness of logical modeling in cases if IPCO's, VDWS's, or interpenetrating of the above types of surfaces are used.

Consider some properties of lower Boolean lattices under the assumption that all the nontruncated domains, truncated domains, as well as all the deleted sets are used as a base sets of topology T. For clarity, all of these sets are henceforth called domains and are denoted as $D_1^{i_1}, \ldots, D_1^{i_N}$.

In the general case, $T_s^{i_1,\ldots,i_s}$ (for $s \geqslant 2$) denotes the maximal finite set that contains $|T_s^{i_1,\ldots,i_s}|$ nonempty subsets of maximal connected components which are subsets of the intersection of the domains $D_1^{i_1}, \ldots, D_1^{i_s}$ closures. It is assumed that each element of the set $T_s^{i_1,\ldots,i_s}$ is obtained from one of such maximal connected component, after deleting all the points that belong to closure of at least one other domain. If the deletion of such points leads to an empty set, that should not be represented in the set $T_s^{i_1,\ldots,i_s}$.

Assuming that $T_1^{i_1} = \{D_{i_1}\}$, the sets $T_s^{i_1,\ldots,i_s}$ correspond to elements $D_1^{i_1} \cup \cdots \cup D_1^{i_s}$, $s = 1, \ldots, N$, of a lower Boolean sublattice. By taking into account the previously mentioned isomorphism φ, the sets correspond to the elements of Boolean N-cube as well. Therefore, the properties of the sets $T_s^{i_1,\ldots,i_s}$ can be used to define a logical function f, as was introduced in the previous section. In fact, the final definition depends on a preliminary geometrical topological model of a molecular shape, but it is useful to note some common agreements. For example:

$$f(\underbrace{0, \ldots, 0}_{i-1}, 1, \underbrace{0, \ldots, 0}_{N-i}) = 0, \tag{3.1}$$

if D_1 is a deleted set.

Moreover, $f(a_1, \ldots, a_i, \ldots, a_N) = 0$, if

$$(T_s^{i_1,\ldots,i_s}) = \emptyset,$$

where i_j are the numbers of the binary elements $a_{i_j} = 1$.

Furthermore, for MIDCO- or MEPCO-based molecular shapes, the function is defined such that

$$f(\underbrace{0, \ldots, 0}_{i-1}, 1, \underbrace{0, \ldots, 0}_{N-i}) = 1 + \mu(D_i), \tag{3.2}$$

and

$$f(\underbrace{0, \ldots, 0}_{i-1}, 1, \underbrace{0, \ldots, 0}_{j-i-1}, 1, \underbrace{0, \ldots, 0}_{n-i-j}) = \begin{cases} 1, & \text{if } \operatorname{clos}(D_i) \cap \operatorname{clos}(D_j) = T_2^{i,j} \neq \emptyset, \\ 0, & \text{otherwise,} \end{cases} \tag{3.3}$$

where $\mu(D_i)$ is the curvature index of the domain D_i.

The function f is referred as a *molecular contour function* (MCOF).

3.2. Logical models of van der Waals surfaces

For the logical modeling of VDWSs, additionally, two items need to be accounted for:

(a) the sets

$$T_2^{i_1,i_2}, \quad s = 2,$$

can contain more than one set; and

(b) the sets

$$T_s^{i_1,\ldots,i_s}, \quad s > 2,$$

can be nonempty. Note that these sets can consist of more than one, but less than N, maximal connected components, where N is the number of nuclei. All elements of the sets

$$T_s^{i_1,\ldots,i_s}, \quad s > 2,$$

are either single-membered sets or empty sets. The single-membered sets consist of a unique element that belongs to a sphere or spheres represented on the VDWS by some domains

$$D_1^{i_1}, \ldots, D_1^{i_s}.$$

If the number of spheres to which an element belongs is denoted as k, then, in the general case, s does not equal k.

Although the theory of logical modelling can be generalized, in the present discussion for clarity it will be restricted to the case where shapes satisfy the criterion that a *VDWS contains at least one domain from each sphere*. In this case elementary geometry shows that, in the case of $s > 2$, the sets $T_s^{i_1,\ldots,i_s}$ satisfy the following properties:

(a) $|T_s^{i_1,\ldots,i_s}| \leqslant 2$;

(b) $|T_s^{i_1,\ldots,i_s}| = 2$ implies the equality $k = s$ for elements of single-membered sets from $T_s^{i_1,\ldots,i_s}$;

(c) if $|T_s^{i_1,\ldots,i_s}| = 1$ when $s > 2$ then $k \in \{s, s+1\}$ for an element of a single-membered set from $T_s^{i_1,\ldots,i_s}$.

According to the proposed logical modeling method, the construction of logical models of molecular shapes represented by VDWSs can be achieved by utilizing the following rules.

The topological space (VDWS, T), where VDWS is the modeled van der Waals surface, and T is a finite topology based on the set B of the VDWS domains

$$D_1^j, \quad j = 1, \ldots, N,$$

should be chosen. It is assumed that some of these domains have been deleted, while others have been left in the VDWS.

The Boolean function $f : B_{m_1} \to \{0, 1, \ldots, N\}$ is defined as follows:

$$f(\underbrace{0, \ldots, 0}_{i-1}, 1, \underbrace{0, \ldots, 0}_{N-i}) = \begin{cases} 0 & \text{if } D_i \text{ is a deleted domain,} \\ 3 & \text{otherwise,} \end{cases}$$

according to the rules (3.1) and (3.2);

$$f(\underbrace{0, \ldots, 0}_{i-1}, 1, \underbrace{0, \ldots, 0}_{j-i}, 1, \underbrace{0, \ldots, 0}_{N-i-j}) = \left|\left(T_2^{i,j}\right)\right|, \tag{3.6}$$

in the case, where the weight (the number of nonzero elements) of the N-tuple $r(a_1, \ldots, a_n)$, is greater than two, the indices $i_1, \ldots, i_s$, $s > 2$, correspond to the position of the nonzero elements of the binary N-tuple

$$f(a_1, \ldots, a_n) = \begin{cases} 0 & \text{if } |\,(T_s^{i_1 \ldots i_s})'| = 0, \\ 1 & \text{if } |(T_s^{i_1 \ldots i_s})'| = 1, \text{ and } s = k, \\ 3 & \text{if } |(T_s^{i_1 \ldots i_s})'| = 1, \text{ and } s \neq k, \\ 2 & \text{if } |(T_s^{i_1 \ldots i_s})'| = 2. \end{cases} \tag{3.7}$$

As previously stated, k is the number of spheres to which a unique element of the set $(T_s^{i_1 \ldots i_s})'$ belongs.

The domain of f coincides with the N-cube isomorphic to lower Boolean sublattice of the finite topology T. Assuming that the VDWS contains at least one domain from each sphere, it can be shown that the function $f(x_1, \ldots, x_n)$ is well defined. This logical function can be called the *VDWS function*.

Corresponding examples are given in [1].

4. Definition of a metric on molecular shape family

4.1. Logical models of molecular shape families

In this section, we describe one example of molecular shape family and logical models for the shapes from this family.

Example 4.1. Consider a four-atom molecule of type AB_3. The shapes shown in figures 1–5 consequently represent all the different geometrical topologies of molecular shapes in bending oscillation corresponding to the "umbrella inversion" of the system AB_3 from [15]. All of the VDWSs contain four or five subdomains from the set

$$\left\{D_1^1, D_1^2, D_1^3, D_1^4, D_1^5\right\}$$

and maximal connected components D_k^j of VDWS that points belong to exactly k, $k > 1$, van der Waals spheres.

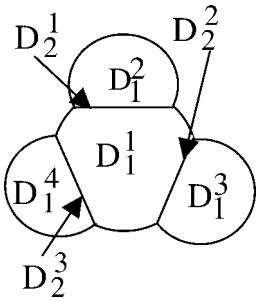

Figure 1. The first topologically different VDWS in the bending oscillation corresponding to the "umbrella inversion" of an AB_3 type molecular system [15] with the bond angle $\gamma = \gamma_1$.

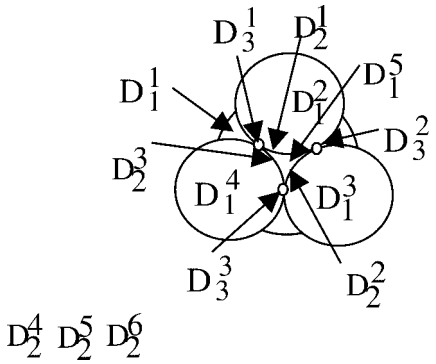

Figure 2. The second topologically different VDWS in the bending oscillation corresponding to the "umbrella inversion" of an AB_3 type molecular system [15] with the bond angle $\gamma = \gamma_2$.

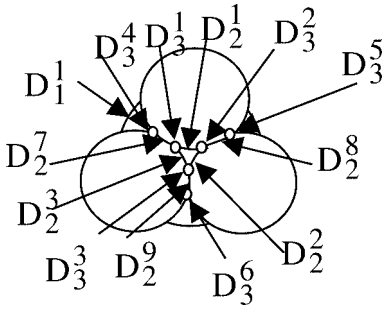

Figure 3. The third topologically different VDWS in the bending oscillation corresponding to the "umbrella inversion" of an AB_3 type molecular system [15] with the bond angle $\gamma = \gamma_3$.

Overall, these shapes correspond to the various nuclear configurations that represent the elements of the dynamic sequence of equivalence classes of nuclear configurations. The representative nuclear configurations differ from each other in bond angle γ. In this particular case, all the domains are convex.

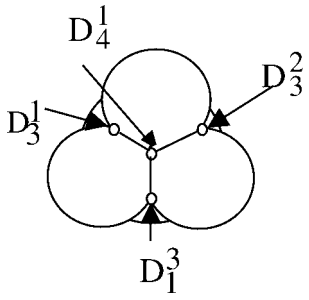

Figure 4. The fourth topologically different VDWS in the bending oscillation corresponding to the "umbrella inversion" of an AB_3 type molecular system [15] with the bond angle $\gamma = \gamma_4$.

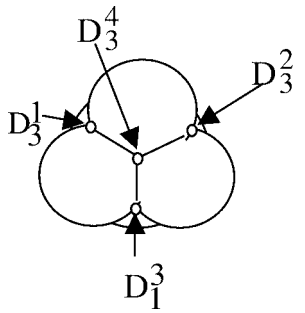

Figure 5. The fifth topologically different VDWS in the bending oscillation corresponding to the "umbrella inversion" of an AB_3 type molecular system [15] the bond angle $\gamma = \gamma_5$.

The partitions of molecular surfaces in figures 1–5 resulting from bond angles $\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5$ such that $\gamma_1 > \gamma_2 > \gamma_3 > \gamma_4 > \gamma_5$ generate the following sets:

(a) $\gamma = \gamma_1$: $D_1^1, D_1^2, D_1^3, D_1^4, T_2^{1,2}, T_2^{1,3}$ and $T_2^{1,4}$;

(b) $\gamma = \gamma_2$: $D_1^1, D_1^2, D_1^3, D_1^4, D_1^5, T_2^{1,2}, T_2^{1,3}, T_2^{1,4}, T_2^{2,5}, T_2^{3,5}, T_2^{4,5}, T_2^{2,3}, T_2^{2,4}, T_2^{3,4}$, $T_3^{1,2,3}, T_3^{1,2,4}, T_3^{1,3,4}, T_3^{2,3,5}, T_3^{2,4,5}$ and $T_3^{3,4,5}$;

(c) $\gamma = \gamma_3$: $D_1^1, D_1^2, D_1^3, D_1^4, D_1^5, T_2^{1,2}, T_2^{1,3}, T_2^{1,4}, T_2^{2,5}, T_2^{3,5}, T_2^{4,5}, T_2^{2,3}, T_2^{2,4}, T_2^{3,4}$, $T_3^{1,2,3}, T_3^{1,2,4}, T_3^{1,3,4}, T_3^{2,3,5}, T_3^{2,4,5}$ and $T_3^{3,4,5}$;

(d) $\gamma = \gamma_4$: $D_1^1, D_1^2, D_1^3, D_1^4, T_2^{1,2}, T_2^{1,3}, T_2^{1,4}, T_2^{2,3}, T_2^{2,4}, T_2^{3,4}, T_3^{1,2,3}, T_3^{1,2,4}, T_3^{1,3,4}$ and $T_3^{2,3,4}$;

(e) $\gamma = \gamma_5$: $D_1^1, D_1^2, D_1^3, D_1^4, T_2^{1,2}, T_2^{1,3}, T_2^{1,4}, T_2^{2,3}, T_2^{2,4}, T_2^{3,4}, T_3^{1,2,3}, T_3^{1,2,4}, T_3^{1,3,4}$ and $T_3^{2,3,4}$.

Here the sets $T_2^{i,j}$, $i, j \in \{1, 2, 3, 4, 5\}$, $i \neq j$, contain the one-dimensional set that are the subsets of closures of both the domains D_1^i and D_1^j; the sets $T_3^{i,j,k}$, $i, j, k \in \{1, 2, 3, 4, 5\}$ are maximal finite sets that contain the single-membered components D_3^s of closure of three domains D_1^i, D_1^j and D_1^k.

The five shapes in figures 1–5 are represented as VDWS functions in table 2.

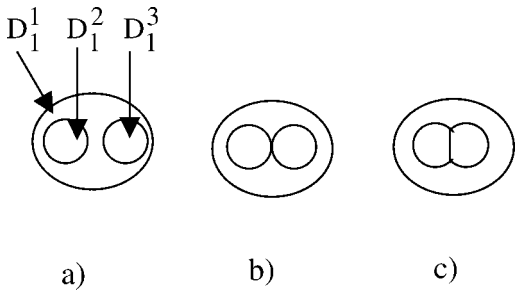

Figure 6. The topologically different VDWSs in the bending oscillation corresponding to the "umbrella inversion" of an AB_2 type molecular system.

4.2. A metric and semimetric on molecular shape family

In this section, the logical method of definition of a metric on the family of molecular shapes will be described. This method explores the Chegis–Yablonsky approach (section 2) to data analysis. To implement it, we should possess unified functional descriptions of shapes from an embedding shape family. As such descriptors, matrix codes [2], shape group codes [2], or functional models can be implemented. The main advantage of unified functional descriptors is the possibility of their comparison and representation of comparison results in a form of discriminate functions. From discriminate functions one may construct the metric or semimetric distances between shapes following the method described in section 2. Let us consider two examples. In this section we implement this method in a frame of finite topology principle (section 2).

Example 4.2. Three logical models, obtained according to the rules (3.2), (3.3), (3.6), (3.7) for the VDWSs in figure 6, where the topologically different VDWSs in the bending oscillation corresponding to the "umbrella inversion" of an AB_2 type molecular system. are represented as logical functions in table 2.

To define the structural peculiarities, discriminating the three molecular shapes under consideration, we should construct the discriminate functions $f_{ab}(x_1, x_2, x_2)$, $f_{ac}(x_1, x_2, x_2)$, $f_{bc}(x_1, x_2, x_2)$ according to the rule (2.1). The corresponding binary table is shown as a part (see the three right columns) of table 2.

The weights w define one of many possible variants of metric on the set of three molecular shapes in figure 6.

The minimal covering $\{(1, 1, 1)\}$ contains only one row and we can see that the value $f_x(1, 1, 1)$, $x \in (a, b, c)$, allows to identify the function from the set $\{f_a, f_b, f_c\}$. This result can be interpreted in terms of preliminary WDVS models as follows: Three WDVS differ from each other because the set $T_3^{1,2,3}$ contains distinct number of single-membered sets; for WDVS in figures 6(a)–(c) these numbers are equal to 0, 1, and 2, correspondingly.

Topological spaces corresponding to the shapes in figure 6 contain isomorphic topologies. Let us construct the metric on the set of shapes in figures 1–5 that does not

Table 3
VDWS functions for the VDWSs in figures 1–5.

$a_1, \ldots, a_5$	f_a	f_b	f_c	f_d	f_e
00000	0	0	0	0	0
00001	0	3	3	0	0
00010	3	3	3	3	3
00011	3	1	1	3	3
00100	3	3	3	3	3
00101	3	1	1	3	3
00110	0	0	1	1	1
00111	0	0	1	1	1
01000	3	3	3	3	3
01001	3	1	1	3	3
01010	0	0	1	1	1
01011	0	0	1	1	1
01100	0	0	1	1	1
01101	0	0	1	1	1
01110	0	0	0	3	1
01111	0	0	0	3	1
10000	3	3	3	3	3
10001	3	0	0	3	3
10010	1	1	1	1	1
10011	1	0	0	1	1
10100	1	1	1	1	1
10101	1	0	0	1	1
10110	0	0	1	1	1
10111	0	1	0	1	1
11000	1	1	1	1	1
11001	1	0	0	1	1
11010	0	0	1	1	1
11011	0	1	0	1	1
11100	0	0	1	1	1
11101	0	1	0	1	1
11110	0	0	0	0	0
11111	0	0	0	0	0

satisfy this property. Nevertheless, introducing fictitious variables we can make them comparable and construct a metric on this set.

To calculate the metric we represent the VDWS functions (figures 1–5) in the unified form (introducing fictitious variables) in table 3. The discriminate functions and their weights (distances between functions) are represented in table 4.

One minimal test set consists of three binary 5-tuples (test vectors), for example, (0, 1, 1, 1, 0), (0, 0, 0, 0, 1) and (0, 0, 1, 1, 0).

It has to contain 5-tuples (0, 1, 1, 1, 0) or (0, 1, 1, 1, 1) because only these 5-tuples distinguish the shapes in figures 4 and 5, and it must contain at least two more 5-tuples because no single 5-tuple can distinguish shapes in pairs shown in figures 1 and 2, in

Table 4
VDWS discriminate functions for the VDWSs in figures 1–5 and their weights w'.

$a_1, \ldots, a_5$	f_{ab}	f_{ac}	f_{ad}	f_{ae}	f_{bc}	f_{bd}	f_{be}	f_{cd}	f_{ce}	f_{de}
00000	0	0	0	0	0	0	0	0	0	0
00001	1	1	0	0	0	1	1	1	1	0
00010	0	0	0	0	0	0	0	0	0	0
00011	1	1	0	0	0	1	1	1	1	0
00100	0	0	0	0	0	0	0	0	0	0
00101	1	1	0	0	0	1	1	1	1	0
00110	0	1	1	1	1	1	1	0	0	0
00111	0	1	1	1	1	1	1	0	0	0
01000	0	0	0	0	0	0	0	0	0	0
01001	1	1	0	0	0	1	1	1	1	0
01010	0	1	1	1	1	1	1	0	0	0
01011	0	1	1	1	1	1	1	0	0	0
01100	0	1	1	1	1	1	1	0	0	0
01101	0	1	1	1	1	1	1	0	0	0
01110	0	0	1	1	0	1	1	1	1	1
01111	0	0	1	1	0	1	1	1	1	1
10000	0	0	0	0	0	0	0	0	0	0
10001	1	1	0	0	0	1	1	1	1	0
10010	0	0	0	0	0	0	0	0	0	0
10011	1	1	0	0	0	1	1	1	1	0
10100	0	0	0	0	0	0	0	0	0	0
10101	1	1	0	0	0	1	1	1	1	0
10110	0	1	1	1	1	1	1	0	0	0
10111	1	0	1	1	1	0	0	1	1	0
11000	0	0	0	0	0	0	0	0	0	0
11001	1	1	0	0	0	1	1	1	1	0
11010	0	1	1	1	1	1	1	0	0	0
10011	1	0	1	1	1	0	0	1	1	0
11100	0	1	1	1	1	1	1	0	0	0
11101	1	0	1	1	1	1	0	0	1	0
11110	0	0	0	0	0	0	0	0	0	0
11111	0	0	0	0	0	0	0	0	0	0
w'	11	17	14	14	12	20	19	12	13	2

figures 1 and 3, and in figures 2 and 3, that are not distinguished considering 5-tuple (0, 1, 1, 1, 0) as well.

The metric on the set of shapes in figures 1–5 can be given in tabular form as is shown in table 5. If distances are defined as weights of discriminate function, as it was accepted above, then they reflect the number of structural differences of compared molecular shapes.

Consequently, the introduced logical models can be used to represent molecular shape families. Moreover, the molecular shape families can be compared and distinguished by means of metric distances obtained from comparison of functional models. The same can be concluded with respect to functional models of cross-sections of mole-

Table 5
Metric on the set of shapes in figures 1–5.

Figure	11	12	13	14	15
11	0	11	17	14	14
12	11	0	12	20	19
13	17	12	0	12	13
14	14	20	12	0	2
15	14	19	13	2	0

cular surfaces, energy hypersurfaces, or spherical projections of folding macromolecules etc.

5. Concluding remarks

A new discrete matematical method for modelling and comparison of molecular shapes and their families was proposed in the previous article and further developed for the chemical applications in molecular shape analysis evaluated in the present contribution. As a result, the following Finite Topology Principle of logical modeling of molecular shapes was founded:

(i) Choose appropriate quantum-chemical or classical preliminary model.

(ii) Choose a domain of logical model basing on General Finite Topology Property.

(iii) Assign logical function values to domain elements and to subsets of these elements reflecting their properties and interrelations.

(iv) Implement the method of logical diagnostics to describe differences of objects and introduce a metric (or semimetric).

(v) Implement metric classification algorithms to classify the objects.

The most important features of the proposed method of molecular shape analysis can be summarized as follows:

- The variables as well as the values of the logical models possess physical meaning.
- The method reflects symmetrical relations of arbitrary degree on the modeled object structure.
- The method allows one to define a metric or semimetric on the set of objects, providing implementation of recently developed metric classification algorithms as well as to making use of earlier for application of methods of logical diagnostics for analysis of molecular shape dissimilarity.
- The diagnosis and the metric on the factor space does not depend on coding of functional models.

In conclusion, we outline the possible further implementations of the new method:

- Logical modelling and metric description of families of molecular shape cross-sections.
- Logical modelling of spherical projection of folding macromolecules and other spherical projections.
- Logical modelling of potential energy hypersurfaces.
- Evaluation of shape complementarity
- Decision-making systems organization.

Acknowledgements

Authors are grateful to Prof. A.A. Bolotov (LSI Logic, USA) for useful consultation and discussion and to Peter Warburton (Department of Chemistry, University Saskatchewan, Canada) for linguistic help. The article was written in Collegium Budapest Institute for Advanced Study.

References

[1] A. Frolov, E. Jako and P.G. Mezey, J. Mathematical Chemistry (30) (2001) 389–409.
[2] P.G. Mezey, *Shape in Chemistry: An Introduction to Molecular Shape Topology* (VCH, UK, 1993).
[3] R. Lidl and G. Pilz, *Applied Abstract Algebra* (Springer, 1984).
[4] A. Mishchenko and A. Fomenko, *A Course of Differential Geometry and Topology* (Mir, Moscow, 1988).
[5] I.A. Chegis and S.V. Yablonsky, Trudy Mat. Inst. Steklov. 51 (1958) 5 (in Russian).
[6] E.J. McCluskey, *Logic Design Principles* (Prentice-Hall, Englewood Cliff, NJ, 1986).
[7] A.B. Frolov, *Models and Methods of Technical Diagnostics* (Znanie, Moscow, 1990) (in Russian).
[8] E. Jako and P. Ittzés, Abstracta Botanica 22 (1998) 121.
[9] S.V. Yablonsky, *Introduction to Discrete Mathematics* (Mir, Moscow, 1988).
[10] A.E. Andreev, A.A. Bolotov and A.B. Frolov, *Discrete Mathematics: Discrete Optimisation Problems and Complexity of Algorithms* (MPEI Publisher, Moscow, 2000).
[11] A.V. Aho, J.E. Hopcroft and J.D. Ulman, *The Design and Analysis of Computer Algorithms* (Addison-Wesley, Reading, MA, 1976).
[12] A.A. Bolotov and A. Shajeb, Dokl. Akad. Nauk (1987) 297; 3 (1987) 527 (in Russian).
[13] A.A. Bolotov, Diskret. Mat. 8(4) (1996) 62 (in Russian).
[14] A.A. Bolotov and A.B. Frolov, *Classification and Recognition in Discrete Systems* (MPEI, Moscow, 1997) (in Russian).
[15] G.A. Arteca and P.G. Mezey, Internat. J. Quantum Chem. 34 (1988) 517.