Journal of Mathematical Chemistry 12(1993)365-373

Topological shape analysis of chain molecules: An application of the GSTE principle

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

One of the fundamental tools of the molecular topology program is the GSTE principle: *g*eometrical *s*imilarity is treated as *t*opological *e*quivalence. The molecular topology (MT) approach to the description of molecules and the reaction topology (RT) approach to the description of reactions are the two main aspects of the topology program (TP), a research program for the reformulation and description of some basic concepts of chemistry within a differential and algebraic topological framework. In this report, a new application of the GSTE principle to the shape analysis of chain molecules, in particular, of chain biomolecules is described, with special emphasis on the dynamic aspects of conformational changes and folding processes of proteins.

1. Introduction

The study of the relations between the chemical properties and three-dimensional shapes of chain molecules, in particular, of chain biomolecules such as polypeptides, proteins, or DNA, is of fundamental importance in modern chemistry and biochemistry (for a sample of references see, e.g. refs. [1-17]). Many chemical properties are dependent on the topological shape properties of these molecules, and topological shape characterization techniques are becoming important tools of both theoretical and applied chemical research. Among the three-dimensional shape analysis methods of tertiary structures of proteins, the method of approximate polyhedral fitting [15,16] provided an early, essentially topological classification technique that was followed by several, more general topological methods for both static and dynamic shape analysis [18-23].

Topological methods have several advantages over more conventional, geometrical approaches. A fundamental consideration is that molecules are not geometrical but topological objects [24], since most small geometrical changes do not alter the identity of molecules. In dynamic processes, such as conformational rearrangements or chemical reactions, many geometrical shape features may change but some of the essential, topological shape properties can remain invariant. These invariant shape features and the domains in configuration space where these features are preserved can be characterized by algebraic topological means, suitable for algorithmic shape analysis by computer programs.

In this study, we shall describe an application of the GSTE principle to "shape globe invariance maps" (SGIM) of protein folding patterns, and a planar shape map as well as the associated shape matrices and shape graphs derived from the SGIM. The actual (P, W)-shape types [24] will be defined as follows. The shape representation P will be chosen as the protein backbone projected on tangent planes of a shape globe enclosing the protein. The topological shape descriptor W will be chosen as the pattern of invariance domains on the globe and on the associated planar map where within each domain the "fuzzy" crossing patterns of the projected backbone images on the tangent plane (with tangent points within the domain) are topologically equivalent.

In general, the shape representation P may involve some parameters, taken as the components of a k-dimensional vector $\boldsymbol{p}$ of some vector space $\boldsymbol{P}$, $P = P(\boldsymbol{p})$. In our present case, we shall consider only one such parameter p, which will be regarded as representing an energy bound $p = \varepsilon$, the energy available to the protein above a reference energy, for example, above its absolute energy minimum or above one of its local energy minima. The extreme case of $\varepsilon = 0$ corresponds to the classical, geometrically defined minimum energy conformation of the protein. For small, positive ε values, the protein enjoys a limited conformational freedom, allowing it to change its shape slightly, but not too drastically, from the reference (minimum) energy conformation. As the energy parameter ε increases, the conformational freedom also increases, leading to the possibility of more prominent shape variations, that is, to a "fuzzier" shape of the protein backbone. This fuzziness in shape can be treated by the techniques proposed for more general "fuzzy" conformational problems [25–27] and for a fuzzy set description of approximate symmetry ("syntopy" [28]), combined with the topological shape characterization methods. The approach provides a topological fuzzy set characterization of the dynamic shape problem of protein backbones.

2. Fuzzy shape globe invariance maps (FSGIM)

Consider the backbone of a protein, a curve in three-dimensional space, as the primary shape representations P^0. Enclose P^0 within a sphere S, placing the center of mass of the molecule so that it coincides with the center of the sphere S. One may choose the smallest such sphere; however, the analysis leads to identical topological results for any sphere that encloses P^0 and is centered on the center of mass of the molecule. Project P^0 onto a tangent plane $R(s)$ at each point s of the sphere S by beams perpendicular to $R(s)$. The projection $P'(s)$ of the shape representation P in each tangent plane $R(s)$ can be characterized topologically, leading to a family of topological descriptors

$$F(s) = \{I(i), i = 1, \dots, k\}. \tag{1}$$

In earlier works [19–22], these topological descriptors have been chosen in a variety of ways, for example, as multigraphs or as knots compatible with (the slightly modified) crossing patterns of the projected image [19–21] or neighbor relation graphs and matrices of the visible domain patterns of the enclosed molecule [22].

A family $F_j(s) = \{I(i), i = 1, \ldots, k\}$ of topological descriptors $I(i)$ remains invariant within some domain C_j of points s of the sphere S. Usually, there are only a finite number of different $F_j(s) = \{I(i), i = 1, \ldots, k\}$ sets for each fixed protein backbone P^0, and each such projected shape invariance domain C_j may be regarded as analogous to a country on a global map. The families $F_j(s)$ generate the shape globe invariance map SGIM, or simply the shape globe map on S. We emphasize an important feature of these shape globe maps: the set $F(s) = \{I(i), i = 1, \ldots, k\}$ of topological descriptors assigned to each point s of the sphere S provides information on a global property of the enclosed molecule, in our case, of the enclosed backbone P^0.

In fig. 1, two of the essential steps of the SGIM method are shown, with a special choice for shape descriptor W, taken as a graph of the crossing pattern. The first step shown is the projection of a protein backbone to tangent planes of a shape globe S and the generation of the graphs (possibly multigraphs or pseudographs) of the crossing patterns associated with each tangent point s. The second step is the generation of invariance domains C_j of these graphs on the shape globe S. This approach leads to a two-dimensional representation of the shape of the molecular backbone on a spherical surface.

The shape globe invariance maps SGIM on the shape globe S can also be characterized topologically. One such characterization is by their shape groups [29,30], as defined by a specified truncation pattern obtained by eliminating invariance domains C_j of certain $F_j(s)$ types. An alternative method is based on the neighbor relations of the invariance domains C_j on the global map SGIM, leading to treatments analogous to the shape graph [31] and shape matrix methods [32]. In addition to the actual, topological shape information, the information on the size of invariance domains on the global map SGIM may also be included in the description, for example, by the ordering of the domains according to their size. These graphs and matrices, possibly augmented with size information, are regarded as alternative shape codes based on the shape globe invariance map method.

Within an earlier approach [33], information on the dynamic shape properties of a molecule has also been included within the above framework, and here we shall present two generalizations of this approach. Both generalizations are applicable to a variety of choices for shape representation P and shape descriptor W within the SGIM framework. For example, one may take P as some molecular surface enclosed within the shape globe S and consider a topological shape descriptor W such as a pattern of domains visible when projected on the tangent planes of S. Nevertheless, in this communication we shall consider explicitly the case of protein backbones and the generated topological descriptions of projected crossing patterns.

The first generalization is a direct extension of the method from two competing molecular configurations to a family of a continuum of configurations, such as

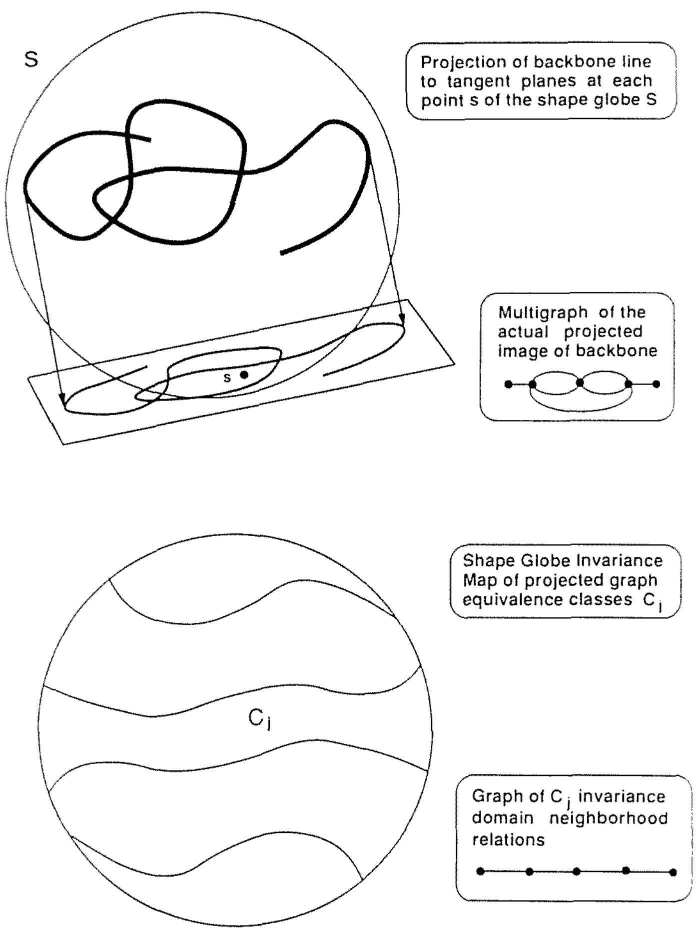

Fig. 1. Two essential steps for the generation of SGIMs. The first step is the projection of a protein backbone to tangent planes of a shape globe S and the generation of the graphs (possibly multigraphs or pseudographs) of the projected crossing patterns associated with each tangent point s, taken as a special choice for shape descriptor W. In the second step, the invariance domains C_j of these graphs are generated on the shape globe S. The resulting SGIM is a two-dimensional representation of the shape of the molecular backbone on the spherical surface S.

those occurring along reaction paths or within conformational domains. If the protein backbone undergoes a major rearrangement from conformation K to conformation K', then the SGIM may change from SGIM(K) to SGIM(K'). A given point s on the shape globe S may become re-assigned from the invariance domain $C_j(K)$ for conformation K to a different invariance region $C_{j'}(K')$ for conformation K', as the conformational change progresses. A new label can be assigned to points s of the sphere "contested" by different invariance domains C_j and $C_{j'}$. One can regard these formal "no man's land" areas, that is, the intersections

$$C_k = C_j(K) \cap C_{j'}(K') \tag{2}$$

on the shape globe S as new, separate domains.

In general, along a configurational change K to K', or within a configurational domain, such as a catchment region, a whole continuum M' of nuclear arrangements K may occur. For any given point s on the shape globe S, one may list all shape invariance domains

$$C_j,\ C_{j'},\ C_{j''},\ \ldots, \tag{3}$$

which contain s for any one of the nuclear configurations K of the family M'. Usually, there are only a finite number of different lists, and point s with the same list can be collected into equivalence classes denoted by

$$C_{j,j',j''},\ \ldots \tag{4}$$

for each list of type (3). Then these new subsets $C_{j,j',j''}, \ldots$ of shape globe S generate a new map on S that can be characterized by the same methods described above. The shape graphs and shape matrices of the resulting dynamic shape globe maps (possibly augmented with size information) are new shape codes, characterizing the dynamic shape properties of the molecule, if the molecule is confined within the conformational domain M'.

The second generalization involves an energy threshold ε and a fuzzy set approach. Consider a reference configuration K, for example, that of a local (or global) energy minimum for the molecule, and the associated map SGIM(K) on the globe S. We may consider SGIM(K) as a secondary shape representation P and one may select an appropriate topological shape descriptor W, for example, neighbor relation graph g_k(SGIM(K)) or shape matrices $\mathbb{M}_k$(SGIM(K)) of domains C_j of SGIM(K), where k is a serial index. Consider now various changes of the configuration K. The kth graph g_k(SGIM(K)) or the kth matrix $\mathbb{M}_k$(SGIM(K)) remain invariant for some conformational changes within M', but they may change with extensive configurational changes. It is possible to generate the corresponding g_k(SGIM(K))-preserving (or the $\mathbb{M}_k$(SGIM(K))-preserving) invariance domains M'_k of the family M' of configurations. This provides a partitioning of family M', somewhat similar to the symmetry domain partitioning of the configurational space M, an analogy we shall exploit below. We may also consider the energy constraint represented by the threshold ε in a manner similar to that in the syntopy model [28]. If ε is small, then only small conformational flexibility is allowed, hence the assignment of an SGIM to the corresponding energy-constrained, dynamic molecular species involves only a small degree of fuzziness. By contrast, if ε is large, then the assignment of any specific SGIM to the molecular species is becoming more fuzzy. The above considerations can be precisely formulated by formally replacing the point symmetry invariance domains G_i of the syntopy model [28] with the $g_k(\mathrm{SGIM}(K))$-preserving (or the $\mathbb{M}_k(\mathrm{SGIM}(K))$-preserving) invariance domains M'_k of the family M' of configurations. With this replacement, the exact derivation of the entire section 3 of ref. [28] can be repeated, leading to a fuzzy membership function $\mu(K, k)$ for $g_k(\mathrm{SGIM}(K))$ graphs (or $\mathbb{M}_k(\mathrm{SGIM}(K))$ matrices) and to an energy-dependent, fuzzy set characterization of molecular shape variations in dynamic processes.

3. The rate of shape change along conformational paths using shape globe invariance maps

The approach described below is generally applicable to any shape globe representation; however, here we shall consider explicitly the case of protein backbones. Consider a configurational change along a formal reaction path $p(t)$ from

$$p(0) = K \tag{5}$$

to

$$p(1) = K', \tag{6}$$

where the usual parametrization of the path is considered [26], taking parameter values t from the unit interval $I = [0, 1]$. Parameter t along the path can be chosen as proportional to the arc length in configurational space M, where the conventional metric of space M is applied [26]. Alternatively, one may consider a parametrization t proportional to time, based on the time scale of typical conformational changes in protein backbone folding processes. In either case, the *rate of shape change* along the conformational path $p(t)$ can be defined in terms of the rate of change of areas $A(C_j)$ of the C_j invariance domains on the SGIM. Note that in general the areas of different invariance domains may change at different rates, and some invariance domains may entirely disappear at intermediate stages of the conformational change. For this reason, not each invariance domain is suitable to give a proper indication of the rate of shape change. One can choose a favored invariance domain $C_{j'}$ that persists along the entire conformational change and follow the shape variation by formally taking

$$\mathrm{d(shape')}/\mathrm{d}t = \mathrm{d}A(C_{j'})/\mathrm{d}t \tag{7}$$

as the rate of shape change. It is clear, however, that this choice is ultimately arbitrary. According to a somewhat less arbitrary definition, one may take the sum of absolute values of the rates of all area changes:

$$\mathrm{d(shape)}/\mathrm{d}t = \sum_j |\mathrm{d}A(C_j)/\mathrm{d}t|. \tag{8}$$

The above quantity is well defined along the entire conformational path $p(t)$. It allows one to provide a quantitative measure for interrelating the extent and speed of conformational changes, as measured by displacements in the metric configurational space M (or by other, less general methods), with the extent and speed of changes of the essential, topological shape properties.

4. Planar representations of shape globe invariance maps and fuzzy shape globe invariance maps

In order to simplify the analysis and to obtain an easily visualizable and recognizable description, it is advantageous to generate a planar representation of SGIMs. One such technique is illustrated in fig. 2. The shape globe S with the SGIM on its surface is placed within a hemisphere H of radius twice that of S. (Note

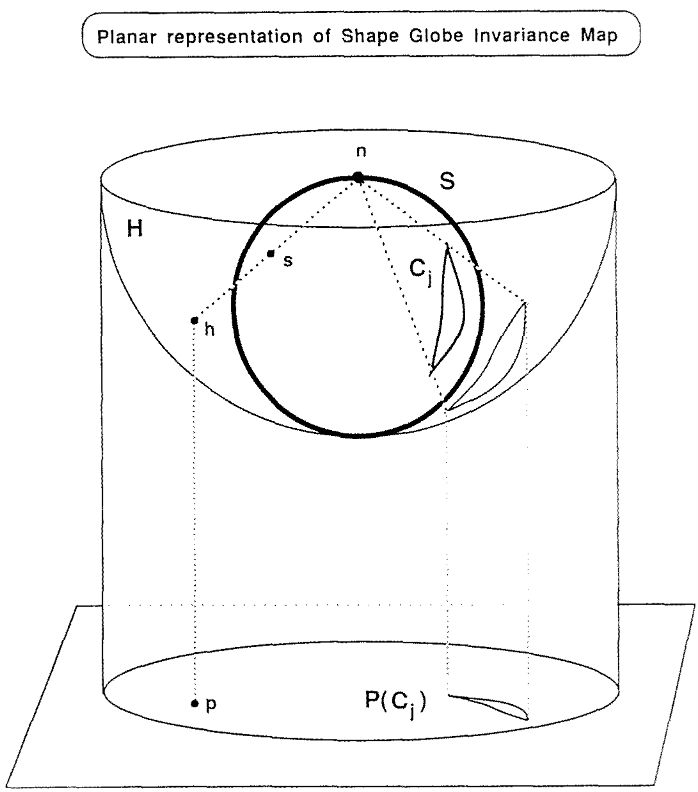

Fig. 2. A planar representation P(SGIM) of the shape globe invariance map. The shape globe S with an SGIM is placed within a hemisphere H of radius twice that of S, over a plane parallel with the perimeter of H, as shown. (For simplicity, only one invariance domain C_j is indicated.) From the "north pole" n of S, a line is issued to each point $s \neq n$ of S, piercing H at a unique point h. A second line issued from point h perpendicular to the plane defines a unique point p of the plane, defining a bijection $P: S\backslash n \rightarrow D$ between the punctured shape globe $S\backslash n$ and an open disc D of the plane. The perimeter of the hemisphere H, as well as the perimeter of the open disc D, are assigned to the north pole n of the shape globe S, completing the generation of a planar representation of the entire SGIM of the shape globe S.

that only one invariance domain C_j of the actual SGIM is shown in the figure.) For simplicity, we assume that H is placed so that it is concave from above and the perimeter of H is horizontal. Then, S is placed to the bottom of H, and a horizontal plane is placed below H. The maximum point of S is identified as the "north pole" n. For each point $s \neq n$ of S, the line issued from n and passing through s pierces H at a unique point h. The line issued from point h perpendicular to the plane defines a unique point p of the plane. This defines a bijection

$$P : S \backslash n \to D \tag{9}$$

between the punctured shape globe $S\backslash n$ and an open disc D of the plane. Furthermore, we can assign the perimeter of the hemisphere H, as well as the perimeter of the open disc D to the north pole n of the shape globe S, completing the generation of a planar representation of the entire shape globe S.

In the planar representation P(SGIM), all neighbor relations between the projected invariance domains $P(C_j)$ are going to be the same as they are on the original SGIM on the shape globe S, with the exception of the special case where a boundary point of a C_j invariance domain falls on the "north pole" n of S. In this case, all projected $P(C_j)$ domains which have boundary points falling on the perimeter of the disc D are also regarded as neighbors, in addition to the usual neighbor relations of projected domains within the disc D. Using this approach, the topological pattern of the interrelations among the invariance domains (although not the ordering of invariance domains by their sizes on the shape globe S) can be obtained directly on the planar disc D.

5. Summary

Two generalizations of the earlier shape globe invariance map technique are presented, designed for applications to the study and nonvisual shape analysis of folding patterns of protein backbones. The first generalization provides a characterization of all shapes occurring within a conformational domain, that is, for an infinite but constrained family of arrangements of the molecular backbone. The second generalization provides a description of some of the dynamic aspects of the protein folding problem based on an energy-dependent fuzzy set representation analogous with the syntopy model of approximate point symmetry, developed earlier. A measure is proposed for the rate of shape change in conformational processes, based on the rate of area changes of invariance domains on shape globes. In addition, a planar description of shape globe invariance maps is proposed, using a simple technique that facilitates the study and visualization of results of all shape globe based analyses.

Acknowledgement

Research work leading to this report has been supported by both operating and strategic research grants from the Natural Sciences and Engineering Research Council of Canada.

References

[1] G.E. Schulz and R.H. Schirmer, *Principles of Protein Structure* (Springer, New York, 1979).
[2] M. Le Bret, Biopolymers 18(1979)1709.
[3] C.R. Cantor and P.R. Schimmel, *The Conformation of Biological Macromolecules, Biophysical Chemistry*, Part I (Freeman, San Francisco, 1980).
[4] J.S. Richardson, Adv. Protein Chem. 34(1981)167.
[5] M. Karplus and J.A. McCammon, Ann. Rev. Biochem. 53(1983)263.
[6] P. De Santis, S. Morosetti and A. Palleschi, Biopolymers 22(1983)37.
[7] R. Franke, *Theoretical Drug Design Methods* (Elsevier, Amsterdam, 1984).
[8] J.S. Richardson, Meth. Enzymol. 115(1985)359.
[9] M.N. Liebman, C.A. Venanzi and H. Weinstein, Biopolymers 24(1985)1721.
[10] A.M. Lesk and K.D. Hardman, Meth. Enzymol. 115(1985)381.
[11] T. Kikuchi, G. Némethy and H.A. Scheraga, J. Comput. Chem. 7(1986)67.
[12] P.M. Dean, *Molecular Foundations of Drug-Receptor Interaction* (Cambridge University Press, New York, 1987).
[13] F.M. Richards and C.E. Kundot, Protein Struct. Funct. Genet. 3(1988)71.
[14] R.A. Abagyan and V.N. Maiorov, J. Biomol. Struct. Dynam. 5(1988)1267.
[15] A.G. Murzin and A.V. Finkelstein, J. Mol. Biol. 204(1988)749.
[16] C. Chothia, Nature 337(1989)204.
[17] M.-H. Hao and W.K. Olson, Biopolymers 28(1989)873.
[18] G.M. Maggiora, P.G. Mezey, B. Mao and K.C. Chou, Biopolymers 30(1990)211.
[19] G.A. Arteca and P.G. Mezey, J. Mol. Graphics 8(1990)66.
[20] G.A. Arteca, O. Tapia and P.G. Mezey, J. Mol. Graphics 9(1991)148.
[21] G.A. Arteca and P.G. Mezey, Algebraic approaches to the shape analysis of biological macromolecules, in: *Theoretical Chemistry, Structure, Interactions and Reactivity*, Part A, ed. S. Fraga (Elsevier, Amsterdam, 1992).
[22] P.G. Mezey, Dynamic shape analysis of biomolecules using topological shape codes, in: *The Role of Computational Models and Theories in Biotechnology*, ed. J. Bertran (Kluwer, Dordrecht, 1992).
[23] P.G. Mezey, *Shape in Chemistry: Introduction to Molecular Shape and Topology* (VCH Publishers, New York, 1993).
[24] P.G. Mezey, Three-dimensional topological aspects of molecular similarity, in: *Concepts and Applications of Molecular Similarity*, ed. M.A. Johnson and G.M. Maggiora (Wiley, New York, 1990).
[25] P.G. Mezey, Differential and algebraic topology of chemical potential surfaces, in: *Mathematics and Computational Concepts in Chemistry*, ed. N. Trinajstić (Ellis Horwood, Chichester, 1986).
[26] P.G. Mezey, *Potential Energy Hypersurfaces* (Elsevier, Amsterdam, 1987).
[27] P.G. Mezey, From geometrical molecules to topological molecules: A quantum mechanical view, in: *Molecules in Physics, Chemistry and Biology*, Vol. II, ed. J. Maruani (Reidel, Dordrecht, 1988).
[28] P.G. Mezey and J. Maruani, Mol. Phys. 69(1990)97.
[29] P.G. Mezey, Int. J. Quant. Chem. Quant. Biol. Symp. 12(1986)113.
[30] P.G. Mezey, J. Comput. Chem. 8(1987)462.
[31] P.G. Mezey, J. Math. Chem. 2(1988)299.
[32] P.G. Mezey, Non-visual molecular shape analysis: Shape changes in electronic excitations and chemical reactions, in: *Computational Advances in Organic Chemistry (Molecular Structure and Reactivity)*, ed. C. Ogretir and I.G. Csizmadia (Kluwer, Dordrecht, 1991).
[33] P.G. Mezey, J. Math. Chem., in press.