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

Dimension concepts and reduced dimensions in toxicological QShAR databases as tools for data quality assessment

Paul G. Mezey [a,b,c,*], Peter Warburton [a], E. Jako [b] and Zsolt Szekeres [d]

[a] *Mathematical Chemistry Research Unit, Department of Chemistry and Department of Mathematics and Statistics, University of Saskatchewan, 110 Science Place, Saskatoon, SK, Canada, S7N 5C9*
E-mail: mezey@sask.usask.ca
[b] *Institute for Advanced Study, Collegium Budapest, Szentháromság u. 2, 1014 Budapest, Hungary*
[c] *CODATA (ICSU/UNESCO), Committee for Data in Science and Technology, CODATA Secretariat, 51 Bd de Montmorency, 75016 Paris, France*
[d] *Institute of Chemistry, University of Budapest, Pázmány Péter Sétány 2, Budapest, Hungary*

Received 24 May 2000

Dedicated to the 80th birthday of Professor Frank Harary

The dimensions of databases can be defined based on a variety of concepts, ranging from the standard tools of principal component analysis to context-biased approaches. The effective dimensions of databases, in particular the effective dimensions involving continua such as electron density data, provide a set of important tools for database comparisons and for the evaluation of some aspects of database quality. The problems associated with database comparisons and database mergers, such as those occurring in the process of database unification in the actual merger of two pharmaceutical companies, provide challenging tasks and opportunities for data science. Some of the tools for effective dimension reduction and dimension expansion are reviewed in the context of database quality control and conditions for database compatibility are presented. A common misconception affecting data sampling techniques for data quality evaluation is discussed and methods for circumventing the associated sampling errors are described.

KEY WORDS: QShAR (Quantitative Shape – Activity Relations) databases, database dimension, reduced dimensions, database quality assessment, sampling errors in high dimensions

1. Introduction

It has been recognized some time ago that with the rapid growth of the volume of numerical data in the natural sciences and other areas, the mathematical and probabilistic foundations of statistical methods and their applications have acquired new significance

* Corresponding author.

0259-9791/01/1100-0375/0 © 2002 Plenum Publishing Corporation

see, e.g., [1–9]. The exceptional advances in biochemistry and biology in general have highlighted the concerns for the reliability of data and the problems associated with the compatibility of sets of data obtained by theoretical models and by various experimental means [10–20].

The problems and questions concerning the reliability of data is not limited to the above areas. In the information age, much of the knowledge accumulated by the human race is stored in databases of various forms, numerical, textual, graphical, etc., and combinations of these see, e.g., [21,22]. The medium of storage is increasingly becoming electronic or at least accessible by electronic means. Criteria for the validity of such databases are of increasing importance [23]. In this context, the problems of data pollution, often of greater significance and impact on society than the pollution of the actual environment, have been pointed out as fundamentally important aspects of data management. The quality of scientific, technical, and other databases, as well as the compatibility of several different databases required for a given task are major topics within the emerging field of Data Science, and are among those aspects which require special consideration by the practicing scientists of the new millennium.

The tasks of data quality management involve a wide range of mathematical, statistical and computational tools if the data themselves are representations of continua, although much of these continua are also represented by discrete numerical data. Continuum models are numerous, for example, weather data are ultimately samples taken from quasi-continua, and such is the case for electron density databases for the interpretation of molecular behavior, especially, within a biochemical context [24–30]. The focus of this study will be on data quality of electron density databanks, but many of the conclusions and the proposed approaches are assumed to be valid for other databanks representing continua.

The goal of many studies involving electron density databanks is to establish correlations between computable molecular properties, ultimately dependent on the molecular electron densities, and pharmacological activity, toxicity, carcinogenicity, mutagenicity, or other biochemical effects of molecules.

Some fundamental molecular properties, such as the nature of functional groups and their typical interactions are natural targets for computational analysis [31–38]. For this purpose the Quantitative Shape – Activity Relations (QShAR) approach [24,31–58] is used, (as opposed to the more traditional QSAR, Quantitative Structure – Activity Relations approach), where the letters Sh refer to the shape of electron density clouds, alternatively, where the letter h may be regarded as a reference to Planck's constant indicating the fact that most of the QShAR methods are based on Quantum Chemistry computations.

The QShAR method focuses on the fundamental information carrier in molecules: the electronic density cloud. The shape of the molecule is the shape of its three-dimensional electron density cloud, that contains far more detailed information than the conventional structural formulas and stereochemical arrangements of the formal bonding skeleton, used in traditional QSAR. A description of the mathematical and computational background for the detailed analysis of molecular shape is given in the monograph

Shape in Chemistry: An Introduction to Molecular Shape and Topology [59], and in the original references [60–124].

In principle, the three-dimensional shape of the molecular electron density cloud contains the complete information about the molecule, as it is formally established by the celebrated Hohenberg–Kohn theorem [125]. A natural question was asked: what information is carried by local regions of molecules? An early approach to apply the Hohenberg–Kohn theorem to local regions succeeded only for artificial molecular models where molecules had boundaries [126], an assumption that does not hold for real molecules. However, according to a more recent result, a theorem stronger than the Hohenberg–Kohn theorem also holds for real molecules: any small nonzero volume piece of the (*boundaryless*, nondegenerate ground state) electron density of each molecule contains the complete information about the entire molecule [127,128]. This implies that local shape features of the electronic density are also suitable for establishing shape – activity correlations, a fact that is advantageous if the toxic or other biochemical activities of large molecules are the subject of the study. Consequently, by establishing correlations between local or global shape features of the electronic density cloud and various biochemical properties, including differences in toxic activities, the success of the QShAR approach has a strong theoretical foundation [129–134].

2. The inherent dimensions and effective dimensions of continuum electron density databanks

Electron densities are three-dimensional continua, however, in the sense of a formal continuum data set, the dimension of this data set is infinity of continuum cardinality. Nevertheless, molecular electron densities exhibit important simplifying features that allows for significant reduction of dimensionality in the study of such densities, at least in a practical sense.

As long as the molecule contains only a finite number of nuclei, the electron density shape analysis can be carried out within an approximate model that can be described by techniques involving only finite effective dimensions. The simplest indication of this fact is obtained from the traditional models of stereochemistry focusing on the nuclear arrangements, and ultimately, on the molecular graph of the bonding pattern. Whereas the molecular graph does not determine the electron density uniquely, nevertheless, it does represent a significant constraint on the electron density cloud. In particular, a formal bond is represented by an edge of the molecular graph, that one may interpret simply as having the two nuclei indicated as the incident vertices of the graph within a "short distance" from one another, whereas nuclei not linked by an edge as being "far" from one another. With such simple "near" and "far" restrictions on the pairwise nuclear arrangements, typically, there exists only a finite number of energetically stable nuclear arrangements for the given molecular stoichiometry. That is, the molecular graph in fact determines much of the essential information about electron densities: if not exclusively the actual electron density of a given conformer, then the set of finitely many electron

densities of the set of all the finite number of stable conformations for the given bonding pattern.

The molecular graph itself is representable by a discrete set of finite number of integers, and by the above arguments, this set of integers provides sufficient input for the (ideally exact) machinery of quantum chemistry to determine a finite set of electron densities of the associated set of stable conformations. By regarding the laws of quantum mechanics, and by extension, the computer programs of quantum chemistry as a mere interpretative tool for information processing, the actual molecular graph is the sole source of specific information for the generation of the entire (finite) set of electron densities. These finite number of electron densities can be ordered, for example, by energy (here we are disregarding cases of degeneracies, where some additional ordering principle, such as multiplicity, can still be applied), hence, besides the dimensionality of the information of the molecular graph, a single additional dimension is all that is required. Consequently, an inherent "effective" dimensionality of electron density clouds of molecules of a finite number of nuclei is also finite.

The finite effective dimensions of individual electron densities imply that electron density databanks can be regarded as data sets in finite dimensions. This fact has advantages both in processing and ordering the actual continua of electron density data in electron density databanks, such as the databank of methyl- and ethyl-substituted polyaromatic hydrocarbon (PAH) electron density fragments, used in toxicity analysis of various PAH molecules.

3. Effective dimensionality as data quality assessment tool

The effective dimensions of electron density databanks can also be studied by direct means: the set of electron density features identified within a databank can be subjected to the standard methods of principal component analysis. Whereas some electron density shape features are likely to be strongly interdependent, a representative set of quasi-independent shape features can be determined, and the dimension of this set can be compared to those of other electron density databases as well as to the inherent dimensions determined by the molecular graph approach.

If the shape-based effective dimensions and the molecular graph-based dimensions differ, this is likely to indicate a deficiency of the electron density databank. The degree of disagreement can be used as a quality measure.

On the other hand, if the shape-based effective dimensions of two electron density databanks are compared, such a comparison may serve as a tool to assess the compatibility of the qualities of the two databanks. The difference between the shape-based effective dimensions of the two databanks is a measure of the quality-compatibility of the two electron density databanks.

4. A counter-intuitive aspect of sampling strategies in higher dimensions

In data sets of finite effective dimensions, especially in dimensions higher than three, some important counterintuitive features may influence the interpretation of the results.

One important problem is associated with two of the typical sampling techniques used in higher dimensional databanks. In one approach, the grid-based sampling method, the data points are selected from a hypercube of suitably chosen edge length, with edges aligned with the coordinate directions representing the (quasi-) independent principal components. In another approach, a reference data point is used, and data samples are taken from the interior of a hypersphere of a suitable radius, centered on the reference data point.

We may easily recognize the associated counterintuitive feature of the two sampling techniques if we consider unit hypercubes (hypercubes of unit edge lengths) and unit hyperspheres (hyperspheres of unit radii), respectively.

Consider the volume ratios of the unit hyperspheres and the unit hypercubes. By definition, the latter object has a d-dimensional volume equal to 1 in all dimensions d, hence the ratio is in fact the d-dimensional volume of the hypersphere of dimension d, denoted by β(d).

This latter volume can be computed [135,136] as

$$\beta(d) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} \tag{1}$$

where the gamma function is given by

$$\Gamma(x) = \int_0^1 \left\{ \ln(1/u) \right\}^{x-1} \mathrm{d}u. \tag{2}$$

Alternatively, using the int(y) function as the integer part of the number y, and the double factorial d!! of odd only or even only factors, one may write the concise formula

$$\beta(d) = \frac{2^{\mathrm{int}((d+1)/2)}\pi^{\mathrm{int}(d/2)}}{d!!}. \tag{3}$$

For example, for the first seven dimensions, $d = 1, 2, 3, 4, 5, 6$, and 7, these numbers $\beta(d)$ are 2, π, $(4/3)\pi$, $\pi^2/2$, $8\pi^2/15$, $\pi^3/6$, and $16\pi^3/105$, respectively.

The actual numerical values, listed in table 1 up to dimensions 20, also shown as a plot in figure 1, are interesting. The function $\beta(d)$ initially increases and has a maximum at dimension 5, and then it decreases monotonically, becomes less than one at dimension 13, and converges to zero in high dimensions! That is, the 13-dimensional unit hypersphere has a volume *smaller* than that of the 13-dimensional unit hypercube (dimension 13 is the smallest dimension for this to occur), and in higher dimensions the volume of the hypersphere becomes negligibly small compared to the volume element, that is, to the volume of the unit hypercube! This is a surprising result if one is used to the idea that the unit circle has an area more than three times greater than the unit square,

Table 1
Unit sphere volumes in various dimensions.

Dimension	Sphere volume
0	1.0000000000000
1	2.0000000000000
2	3.1415926535898
3	4.1887902046667
4	4.9348022002626
5	5.2637890136135
6	5.1677127796068
7	4.7247659699263
8	4.0587121259528
9	3.2985089023616
10	2.5501640395129
11	1.8841038791206
12	1.3352627686256
13	0.91062875462715
14	0.59926452920089
15	0.38144328074699
16	0.23533063030508
17	0.1409811068849
18	0.082145886589997
19	0.046621601018096
20	0.025806891382638

and the unit sphere in three dimensions has a volume more than four times greater than the unit cube.

This observation has implications on the choice of data sampling in high dimensions. If random sampling is taken from a unit hypersphere drawn about a distinguished reference data point, then this sampling, in fact, accesses only a very small part of the multidimensional space, if compared to the volume accessed when sampling is taken from a unit hypercube.

If dissimilarity measures of data points taken from a sampling accessing only a unit hypersphere are large, then this may be regarded as a significant result, since significant dissimilarities already show up if taken from a more limited range of sample. On the other hand, if dissimilarity measures are small even if the data points are taken from a sampling accessing a unit hypercube, then this is a more reliable conclusion that a similar result obtained from a sampling taken from the more restricted unit hypersphere.

5. Summary

The interpretation of statistical results based on sampling strategies in various dimensions is influenced by the perception of the relative volumes of unit cubes (often thought of as blocks making up the sample space) and unit spheres (often regarded as neighborhoods of certain distinguished points in the space). The relative volumes have

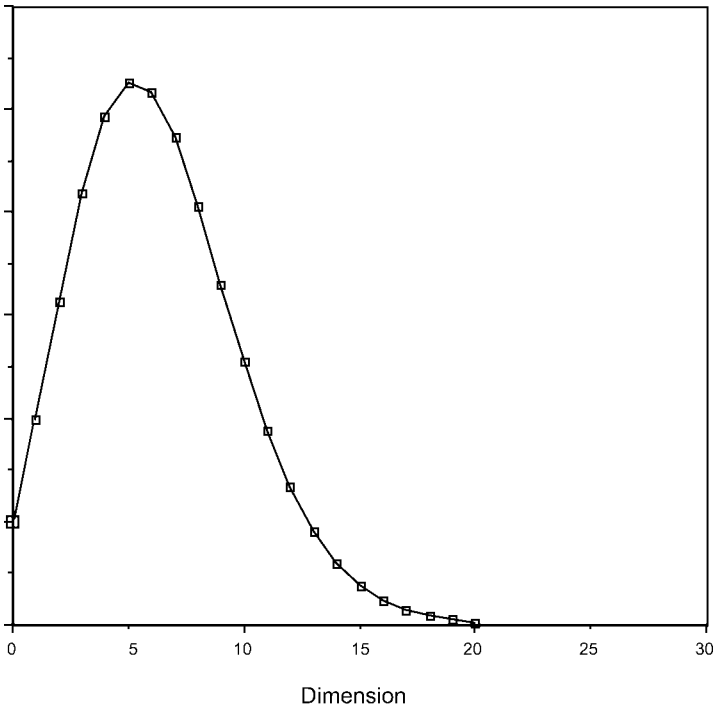

Figure 1. Plot of unit sphere volumes in various dimensions.

different significance in different dimensions. If dimension reduction is successfully accomplished without significant loss of relevant information, as it is possible in certain shape analysis studies of potential energy hypersurfaces and molecular electron density maps, then a reassessment of the sampling strategy may be required to account for the changing relative volume conditions.

Acknowledgements

It is a pleasure to acknowledge the motivating role of Professor Frank Harary in discrete mathematics applications in chemistry. The operating and strategic research grant support of the Natural Sciences and engineering Research Council of Canada, the support of CODATA Task Group on Data Quality and Database Compatibility, and the hospitality of the Institute for Advanced Study, Collegium Budapest, are gratefully acknowledged.

References

[1] J.R. Blum and J.I. Rosenblatt, *Probability and Statistics* (Saunders, Philadelphia, NJ, 1972).
[2] O. Kempthorne and L. Folks, *Probability, Statistics and Data Analysis* (Iowa State Univ. Press, Ames, IA, 1971).

[3] P.G. Hoel, S. Port and C.L. Stone, *Introduction to Probability Theory* (Houghton Mifflin, Boston, USA, 1971).
[4] T.S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach* (Academic Press, New York, 1967).
[5] G.A. Mihram, *Simulation: Statistical Foundations and Methodology* (Academic Press, New York, 1972).
[6] B.W. Lindgren, *Statistical Theory* (Macmillan Co., New York, 1968).
[7] M.H. DeGroot, *Probability and Statistics* (Addison-Wesley, Reading, MA, 1975).
[8] P.D. Lark, B.C. Craven and R.C.L. Bosworth, *The Handling of Chemical Data* (Pergamon, Oxford, 1968).
[9] A. Rényi, *Probability* (Tankönyvkiadó, Budapest, 1973).
[10] R.R. Colwell (ed.), *Biomolecular Data, A Resource in Transition* (Oxford Univ. Press, 1989, Oxford, UK).
[11] A.M. Lesk and C. Chothia, How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins, J. Mol. Biol. 136 (1980) 225–270.
[12] A.M. Lesk and C. Chothia, The response of protein structures to amino acid sequence changes, Philos. Trans. Roy. Soc. London ser. A 317 (1986) 345–356.
[13] L.T.J. Delbaere, G.D. Brayer and M.N.G. James, Comparison of the predicted model of α-lytic protease with the X-ray structure, Nature 279 (1979) 165–167.
[14] C. Chothia, A.M. Lesk, M. Levitt, A.G. Amit, R.A. Mariuzza, S.E.V. Phillips and R.J. Poljak, The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure, Science 233 (1986) 755–758.
[15] T.C. Hodgman, The elucidation of protein function from its amino acid sequence, CABIOS 2 (1986) 181–187.
[16] T.A. Jones and T. Thirup, Using known substructures in protein model building and crystallography, EMBO Journal 5 (1986) 819–822.
[17] W. Kabsch and C. Sander, On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations, Proc. Nat. Acad. Sci. USA 81 (1984) 1075–1078.
[18] W. Kabsch and C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.
[19] A. Morffew, S.J. Todd and X. Snellgrove, The use of a relational database for holding molecule data in a molecular graphics system, Computers and Chemistry 7 (1983) 9–16.
[20] S.B. Needleman and X. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. 48 (1970) 443–453.
[21] P.G. Mezey, Non-visual shape analysis by computer, in: *New Data Challanges in Our Information Age*, eds. P.S. Glaeser and M.T.L. Millward (CODATA, Paris, France, 1994) pp. 18–27.
[22] P.G. Mezey, Shape-data processing in the natural sciences and technology, in: *Data and Knowledge in a Changing World, Modeling Complex Data for Creating Information*, CODATA Series, eds. J.-E. Dubois and N. Gershon (Springer, Berlin, 1996) pp. 147–154.
[23] B.J. McConkey, P.G. Mezey, D.G. Dixon and B.M. Greenberg, Fractional simplex designs for interaction screening in complex mixtures, Biometrics 56 (2000) 824–832.
[24] P.G. Mezey, Shape-similarity measures for molecular bodies: A 3D topological approach to QShAR, J. Chem. Inf. Comput. Sci. 32 (1992) 650–656.
[25] P.G. Mezey, Z. Zimpel, P. Warburton, P.D. Walker, D.G. Irvine, D.G. Dixon and B. Greenberg, A high-resolution shape-fragment database for toxicological shape analysis of PAHs, J. Chem. Inf. Comput. Sci. 36 (1996) 602–611.
[26] P.G. Mezey, Quantitative shape – activity relations (QShAR), molecular shape analysis, charge cloud holography, and computational microscopy, in: *QSARs Environmental Toxicology – VIII. QSARs for Predicting Endocrine Disruption, Chemical Persistence and Effects*, ed. J.D. Walker (SETAC) in press (accepted May 20, 1998).

[27] P.G. Mezey, Z. Zimpel, P. Warburton, P.D. Walker, D.G. Irvine, X.-D. Huang, D.G. Dixon and B.M. Greenberg, Use of QShAR to model the photoinduced toxicity of PAHs: Electron density shape features accurately predict toxicity, Environ. Toxicol. Chem. 17 (1998) 1207–1215.
[28] P.G. Mezey, Molecular structure – reactivity – toxicity relationships, in: *Soil Chemistry and Ecosystem Health*, ed. P.M. Huang (SSSA, Pittsburgh, PA, 1998) pp. 21–43.
[29] P.G. Mezey, Relations between computational and experimental engineering of molecules from molecular fragments, Molec. Engrg. 8 (1999) 235–250.
[30] J.-E. Dubois and P.G. Mezey, A functional group database: A charge density – DARC approach, Molec. Engrg. 8 (1999) 251–265.
[31] P.G. Mezey, Functional groups in quantum chemistry, Adv. Quant. Chem. 27 (1996) 163–222.
[32] P.G. Mezey, Computational microscopy: Pictures of proteins, Pharmaceutical News 4 (1997) 29–34.
[33] P.G. Mezey and P.D. Walker, Fuzzy molecular fragments in drug research, Drug Discovery Today (Elsevier Trend Journal) 2 (1997) 6–11.
[34] Q. Du, G.A. Arteca and P.G. Mezey, Heuristic lipophilicity potential for computer-aided rational drug design, J. Comput. Aided Mol. Design 11 (1997) 503–515.
[35] Q. Du and P.G. Mezey, Heuristic lipophilicity potential for computer-aided rational drug design: Optimizations of screening functions and parameters, J. Comput. Aided Mol. Design 12 (1998) 451–470.
[36] P.G. Mezey, K. Fukui, S. Arimoto and K. Taylor, Polyhedral shapes of functional group distributions in biomolecules and related similarity measures, Internat. J. Quantum Chem. 66 (1998) 99–105.
[37] P.G. Mezey, Molecular similarity and host–guest interactions, Theoret. Comput. Chem. 6 (1999) 593–612; chapter 23 in: *Pauling's Legacy: Modern Modelling of the Chemical Bond*, eds. Z. Maksic and W.J. Orville-Thomas (Elsevier, Amsterdam, 1999) pp. 593–612.
[38] P.G. Mezey, *Topological Methods of Molecular Shape Analysis: Continuum Models and Discretization*, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 51 (2000) 267–278.
[39] P.D. Walker and P.G. Mezey, Molecular electron density lego approach to molecule building, J. Amer. Chem. Soc. 115 (1993) 12423–12430.
[40] P.G. Mezey, Quantum chemical shape: New density domain relations for the topology of molecular bodies, functional groups, and chemical bonding, Canad. J. Chem. 72 (1994) 928–935 (special issue dedicated to Prof. J.C. Polanyi).
[41] P.G. Mezey, Iterated similarity sequences and shape ID numbers for molecules, J. Chem. Inf. Comput. Sci. 34 (1994) 244–247.
[42] P.D. Walker and P.G. Mezey, *Ab initio* quality electron densities for proteins: A MEDLA approach, J. Amer. Chem. Soc. 116 (1994) 12022–12032.
[43] P.D. Walker and P.G. Mezey, Realistic, Detailed images of proteins and tertiary structure elements: *Ab initio* quality electron density calculations for bovine insulin, Canad. J. Chem. 72 (1994) 2531–2536.
[44] P.D. Walker and P.G. Mezey, A new computational microscope for molecules: High resolution MEDLA images of taxol and HIV-1 protease, using additive electron density fragmentation principles and fuzzy set methods, J. Math. Chem. 17 (1995) 203–234.
[45] P.D. Walker and P.G. Mezey, Towards similarity measures for macromolecular bodies: MEDLA test calculations for substituted benzene systems, J. Comput. Chem. 16 (1995) 1238–1249.
[46] P.G. Mezey, Shape analysis of macromolecular electron densities, Struct. Chem. 6 (1995) 261–270.
[47] P.G. Mezey, Molecular similarity measures for assessing reactivity, in: *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*, ed. R. Carbó (Kluwer Academic, Dordrecht, 1995) pp. 57–76.
[48] P.G. Mezey, Methods of molecular shape-similarity analysis and topological shape design, in: *Molecular Similarity in Drug Design*, ed. P.M. Dean (Chapman & Hall/Blackie, Glasgow, 1995) pp. 241–268.
[49] P.G. Mezey, Density domain bonding topology and molecular similarity measures, in: *Topics in Current Chemistry*, Vol. 173, *Molecular Similarity*, ed. K. Sen (Springer, Heidelberg, 1995) pp. 63–83.

[50] P.D. Walker, P.G. Mezey, G.M. Maggiora, M.A. Johnson and J.D. Petke, Application of the shape group method to conformational processes: Shape and conjugation changes in the conformers of 2-phenyl pyrimidine, J. Comput. Chem. 16 (1995) 1474–1482.
[51] P.D. Walker, G.M. Maggiora, M.A. Johnson, J.D. Petke and P.G. Mezey, Shape group analysis of molecular similarity: Shape similarity of six-membered aromatic ring systems, J. Chem. Inf. Comput. Sci. 35 (1995) 568–578.
[52] P.G. Mezey, Local shape analysis of macromolecular electron densities, in: *Computational Chemistry: Reviews and Current Trends*, Vol. 1, ed. J. Leszczynski (World Scientific, Singapore, 1996) pp. 109–137.
[53] P.G. Mezey, Descriptors of molecular shape in 3D, in: *From Chemical Topology to Three-Dimensional Geometry*, ed. A.T. Balaban (Plenum, New York, 1997) pp. 25–42.
[54] P.G. Mezey, Fuzzy measures of molecular shape and size, in: *Fuzzy Logic in Chemistry*, ed. D.H. Rouvray (Academic Press, San Diego, CA, 1997) pp. 139–223.
[55] P.G. Mezey, Quantum chemistry of macromolecular shape, Internat. Rev. Phys. Chem. 16 (1997) 361–388.
[56] P.G. Mezey, Shape in quantum chemistry, in: *Conceptual Trends in Quantum Chemistry*, Vol. 3, eds. J.-L. Calais and E.S. Kryachko (Kluwer Academic, Dordrecht, 1997) pp. 519–550.
[57] P.G. Mezey, Shape analysis, in: *Encyclopedia of Computational Chemistry*, eds. P.V.R. Schleyer, N.L. Allinger, T. Clark, J. Gasteiger, P.A. Kollman, H.F. Schaefer III and P.R. Schreiner, Vol. 4 (Wiley, Chichester, 1998) pp. 2582–2589.
[58] P.G. Mezey, Combinatorial aspects of biomolecular shape analysis, Bolyai Soc. Math. Stud. 7 (1999) 323–332.
[59] P.G. Mezey, *Shape in Chemistry: an Introduction to Molecular Shape and Topology* (VCH, New York, 1993).
[60] P.G. Mezey, Group theory of electrostatic potentials: A tool for quantum chemical drug design, Internat. J. Quantum Chem. Quant. Biol. Sympos. 12 (1986) 113–122.
[61] P.G. Mezey, Tying knots around chiral centres: Chirality polynomials and conformational invariants for molecules, J. Amer. Chem. Soc. 108 (1986) 3976–3984.
[62] P.G. Mezey, The shape of molecular charge distributions: Group theory without symmetry, J. Comput. Chem. 8 (1987) 462–469.
[63] P.G. Mezey, Group theory of shapes of asymmetric biomolecules, Internat. J. Quantum Chem. Quant. Biol. Sympos. 14 (1987) 127–132.
[64] P.G. Mezey, From geometrical molecules to topological molecules: A quantum mechanical view, in: *Molecules in Physics, Chemistry and Biology*, ed. J. Maruani, Vol. II (Reidel, Dordrecht, 1988) Chapter 2, pp. 61–81.
[65] J. Maruani and P.G. Mezey, The concept of "syntopy": A continuous extension of the symmetry concept for quasi-symmetric structures using fuzzy set theory, Compt. Rend. Ser. II 305 (1987) 1051–1054; 306 (1988) 1141.
[66] G.A. Arteca and P.G. Mezey, A topological characterization for simple molecular surfaces, J. Mol. Struct. Theochem 166 (1988) 11–16.
[67] G.A. Arteca, V.B. Jammal, P.G. Mezey, J.S. Yadav, M.A. Hermsmeier and T.M. Gund, Shape group studies of molecular similarity: Relative shapes of Van der Waals and electrostatic potential surfaces of nicotinic agonists, J. Molec. Graphics 6 (1988) 45–53.
[68] P.G. Mezey, Shape group studies of molecular similarity: Shape groups and shape graphs of molecular contour surfaces, J. Math. Chem. 2 (1988) 299–323.
[69] G.A. Arteca, V.B. Jammal and P.G. Mezey, Shape group studies of molecular similarity and regioselectivity in chemical reactions, J. Comput. Chem. 9 (1988) 608–619.
[70] G.A. Arteca and P.G. Mezey, Shape characterization of some molecular model surfaces, J. Comput. Chem. 9 (1988) 554–563.
[71] P.G. Mezey, Global and local relative convexity and oriented relative convexity; application to molecular shapes in external fields, J. Math. Chem. 2 (1988) 325–346.

[72] G.A. Arteca and P.G. Mezey, Methods of topological characterization of molecular surfaces, Folia Chimica Theoretica Latina 15 (1988) 115–154.
[73] G.A. Arteca and P.G. Mezey, Shape description of conformationally flexible molecules: Application to two-dimensional conformational problems, Internat. J. Quantum Chem. Quant. Biol. Sympos. 15 (1988) 33–54.
[74] F. Harary and P.G. Mezey, Graphical shapes: Seeing graphs of chemical curves and molecular surfaces, J. Math. Chem. 2 (1988) 377–389.
[75] J. Pipek and P.G. Mezey, Dependence of MO shapes on a continuous measure of delocalization, Internat. J. Quantum Chem. Sympos. 22 (1988) 1–13.
[76] G.A. Arteca and P.G. Mezey, Molecular conformation and molecular shape: A discrete characterization of continua of van der Waals surfaces, Internat. J. Quantum Chem. 34 (1988) 517–526.
[77] P.G. Mezey, Topology of molecular shape and chirality, in: *New Theoretical Concepts for Understanding Organic Reactions*, eds. J. Bertran and I.G. Csizmadia, Nato ASI Series (Kluwer Academic, Dordrecht, 1989) pp. 77–99.
[78] G.A. Arteca and P.G. Mezey, Shape group theory of van der Waals surfaces, J. Math. Chem. 3 (1989) 43–71.
[79] G.A. Arteca and P.G. Mezey, Discrete characterization of crossections of molecular surfaces, Theoret. Chim. Acta 75 (1989) 333–352.
[80] G.A. Arteca and P.G. Mezey, Molecular similarity and molecular shape changes along reaction paths: A topological analysis and consequences on the Hammond postulate, J. Phys. Chem. 93 (1989) 4746–4751.
[81] P.G. Mezey, The topology of molecular surfaces and shape graphs, in: *Computational Chemical Graph Theory*, ed. D.H. Rouvray (Nova Publications, New York, 1990) pp. 175–197.
[82] P.G. Mezey, Three-dimensional topological aspects of molecular similarity, in: *Concepts and Applications of Molecular Similarity*, eds. M.A. Johnson and G.M. Maggiora (Wiley, New York, 1990) pp. 321–368.
[83] J. Pipek and P.G. Mezey, A fast intrinsic localization procedure applicable for ab initio and semiempirical LCAO wavefunctions, J. Chem. Phys. 90 (1989) 4916–4926.
[84] G.A. Arteca and P.G. Mezey, Two approaches to the concept of chemical species: Relations between potential energy and molecular shape, Internat. J. Quant. Chem., Sympos. 23 (1989) 305–320.
[85] P.G. Mezey, Molecular surfaces, in: *Reviews in Computational Chemistry*, eds. K.B. Lipkowitz and D.B. Boyd (VCH, New York, 1990) chapter 7, pp. 265–294.
[86] P.G. Mezey and J. Maruani, The concept of "syntopy": A continuous extension of the symmetry concept for quasi-symmetric structures using fuzzy-set theory, Mol. Phys. 69 (1990) 97–113.
[87] G.A. Arteca, G.A. Heal and P.G. Mezey, Comparison of potential energy maps and molecular shape invariance maps for two-dimensional conformational problems, Theor. Chim. Acta 76 (1990) 377–390.
[88] P.G. Mezey, Point symmetry groups of all distorted configurations of a molecule form a lattice, J. Math. Chem. 4 (1990) 377–381.
[89] G.A. Arteca and P.G. Mezey, Analysis of molecular shape changes along reaction paths, Internat. J. Quantum Chem. 38 (1990) 713–726.
[90] 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*), eds. C. Ogretir and I.G. Csizmadia, Nato ASI Series (Kluwer Academic, Dordrecht, 1991) pp. 261–288.
[91] P.G. Mezey, Molecular point symmetry and the phase of the electronic wavefunction; Tools for the prediction of critical points of potential energy surfaces, Internat. J. Quantum Chem. 38 (1990) 699–711.
[92] G.M. Maggiora, P.G. Mezey, B. Mao and K.C. Chou, A new chiral feature in α-helical domains of proteins, Biopolymers 30 (1990) 211–215.

[93] P.G. Mezey, A global approach to molecular symmetry: Theorems on symmetry relations between ground and excited state configurations, J. Amer. Chem. Soc. 112 (l990) 3791–3802.
[94] P.G. Mezey, Fivefold symmetry in the context of potential surfaces, molecular conformations and chemical reactions, in: *Quasicrystals, Networks, and Molecules with Fivefold Symmetry*, ed. I. Hargittai (VCH, New York, 1990) pp. 223–238.
[95] P.G. Mezey, Topological Quantum Chemistry, in: *Reports in Molecular Theory*, eds. H. Weinstein and G. Náray-Szabó, Vol. 1 (CRC Press, Boca Raton, 1990) pp. 165–183.
[96] G.A. Arteca and P.G. Mezey, A method for the characterization of foldings in protein ribbon models, J. Mol. Graphics 8 (1990) 66–80.
[97] A.A. Arteca and P.G. Mezey, A quantitative approach to structural similarity from molecular topology of reaction paths, Internat. J. Quantum Chem. Symp. 24 (1990) 1–13.
[98] P.G. Mezey, The role of shape analysis in drug design, in: *IEEE Engrg. in Med. & Bio. Soc. 11th Annual Int. Conf.* (1989) pp. 1905–1906.
[99] G.A. Arteca and P.G. Mezey, Quantitative measures of molecular similarity, in: *IEEE Engrg. in Med. & Bio. Soc. 11th Annual Int. Conf.* (1989) pp. 1907–1908.
[100] P.G. Mezey, The degree of similarity of three-dimensional bodies; Applications to molecular shapes, J. Math. Chem. 7 (1991) 39–49.
[101] P.D. Walker, G.A. Arteca and P.G. Mezey, A complete shape characterization for molecular charge densities represented by Gaussian-type functions, J. Comput. Chem. 12 (1991) 220–230.
[102] F. Harary and P.G. Mezey, Chiral and achiral square-cell configurations and the degree of chirality, in: *New Developments in Molecular Chirality*, ed. P.G. Mezey (Kluwer Academic, Dordrecht, 1991) pp. 241–256.
[103] P.G. Mezey, A global approach to molecular chirality, in: *New Developments in Molecular Chirality*, ed. P.G. Mezey (Kluwer Academic, Dordrecht, 1991) pp. 257–289.
[104] J. Maruani and P.G. Mezey, From symmetry to syntopy: An extension of the symmetry concept to quasi-symmetric structures using fuzzy set theory, J. Chim. Phys. 87 (1990) 1025–1047.
[105] G.A. Arteca and P.G. Mezey, Energy and shape analysis along reaction paths of chemical reactions. The case of hydrogen–deuterium exchange, J. Mol. Structure Theochem 230 (1991) 323–338.
[106] G.A. Arteca and P.G. Mezey, Configurational dependence of molecular shape, J. Math. Chem 10 (1992) 329–371.
[107] G.A. Arteca and P.G. Mezey, A topological analysis of macromolecular folding patterns, in: *Theoretical and Computational Models for Organic Chemistry*, eds. S.J. Formosinho, I.G. Csizmadia and L.G. Arnaut (Kluwer Academic, Dordrecht, 1991) pp. 111–124.
[108] P.G. Mezey, New symmetry theorems and similarity rules for transition structures, in: *Theoretical and Computational Models for Organic Chemistry*, eds. S.J. Formosinho, I.G. Csizmadia and L.G. Arnaut (Kluwer Academic, Dordrecht, 1991) pp. 93–110.
[109] G.A. Arteca and P.G. Mezey, Algebraic approaches to the shape analysis of biological macromolecules, in: *Computational Chemistry, Structure, Interactions and Reactivity*, Part A, ed. S. Fraga (Elsevier, Amsterdam, 1992) pp. 463–487.
[110] G.A. Arteca, O. Tapia and P.G. Mezey, Implementing knot-theoretical characterization methods to analyze the backbone structure of proteins: Application to CTF-L7/L12 and carboxypeptidase A inhibitor proteins, J. Mol. Graphics 9 (1991) 148–156.
[111] G.A. Arteca and P.G. Mezey, A measure of roughness of cross-sections of molecular surfaces, Theor. Chim. Acta 81 (1992) 79–93.
[112] F. Harary and P.G. Mezey, Similarity and complexity of the shapes of square-cell configurations, Theor. Chim. Acta 79 (1991) 379–387.
[113] P.G. Mezey, The alpha-Hull and the T-Hull of a point set: Tools for the analysis of shapes and relative orientations of objects in 3D-space, J. Math. Chem. 8 (1991) 91–102.
[114] G.A. Arteca, A. Hernández-Laguna, J.J. Rández, Y.G. Smeyers and P.G. Mezey, A topological analysis of molecular electrostatic potential on van der Waals surfaces for histamine and 4-substituted derivatives as H_2-receptor agonists, J. Comput. Chem. 12 (1991) 705–716.

[115] X. Luo, G.A. Arteca and P.G. Mezey, Shape analysis along reaction paths of ring opening reactions, Internat. J. Quantum Chem. Sympos. **25** (1991) 335–345.
[116] I. Rozas, G.A. Arteca and P.G. Mezey, On the inhibition of alcohol dehydrogenase: Shape group analysis of molecular electrostatic potential on van der Waals surfaces of some pyrazole derivatives, Internat. J. Quantum Chem. Quant. Biol. Sympos. 18 (1991) 269–288.
[117] G.A. Arteca and P.G. Mezey, Similarities between the effects of configurational changes and applied electric fields on the shape of electron densities, J. Mol. Struct. Theochem 256 (1992) 125–134 (special volume on Electrostatics in Molecules, ed. G. Náray-Szabó and W.J. Orwille Thomas).
[118] G.A. Arteca, N.D. Grant and P.G. Mezey, Variable atomic radii based on some approximate configurational invariance and transferability properties of the electron density, J. Comput. Chem. 12 (1991) 1198–1210.
[119] P.G. Mezey, Similarity analysis in two and three dimensions using lattice animals and polycubes, J. Math. Chem. 11 (1992) 27–45.
[120] P.G. Mezey, On the allowed symmetries of all distorted forms of conformers, molecules, and transition structures, Canad. J. Chem. 70 (1992) 343–347 (special issue dedicated to Prof. S. Huzinaga).
[121] X. Luo, G.A. Arteca and P.G. Mezey, Shape similarity and shape stability along reaction paths. The case of the PPO–OPP isomerization, Internat. J. Quantum Chem. 42 (1992) 459–474.
[122] P.G. Mezey, Topological shape analysis of chain molecules: An application of the GSTE principle, J. Math. Chem. 12 (1993) 365–373.
[123] P.G. Mezey, Dynamic shape analysis of molecules in restricted domains of a configuration space, J. Math. Chem. 13 (1993) 59–70.
[124] 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 Academic, Dordrecht, 1992) pp. 83–104.
[125] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, Phys. Rev. 136 (1964) B864–B865.
[126] J. Riess and W. Münch, The theorem of Hohenberg and Kohn for subdomains of a quantum system, Theor. Chim. Acta 58 (1981) 295–300.
[127] P.G. Mezey, Generalized chirality and symmetry deficiency, J. Math. Chem. 23 (1998) 65–84.
[128] P.G. Mezey, The holographic electron density theorem and quantum similarity measures, Mol. Phys. 96 (1999) 169–178.
[129] P.G. Mezey, Holographic electron density shape theorem and its role in drug design and toxicological risk assessment, J. Chem. Inf. Comput. Sci. 39 (1999) 224–230.
[130] P.G. Mezey, The holographic principle for latent molecular properties, J. Math. Chem., in press.
[131] P.G. Mezey, A uniqueness theorem on molecular recognition, J. Math. Chem., in press.
[132] P.G. Mezey, The holographic electron density theorem and some of its consequences, in: *Computational Chemistry Approaches to Molecular Similarity*, ed. R. Carbó-Dorca (Kluwer Academic/Plenum, New York) in press (accepted May 29, 2000).
[133] P.G. Mezey, Macromolecular density matrices and electron densities with adjustable nuclear geometries, J. Math. Chem. 18 (1995) 141–168.
[134] P.G. Mezey, Quantum similarity measures and Löwdin's transform for approximate density matrices and macromolecular forces, Internat. J. Quantum Chem. 63 (1997) 39–48.
[135] M. Berger, *Geometry* (Springer, Heidelberg, 1987).
[136] J. Bourgain and V.D. Milman, New volume ratio properties of convex symmetric bodies in $\mathbb{R}^n$, Inventiones Math. 88 (1987) 319–341.