J Math Chem (2009) 45:1–6
DOI 10.1007/s10910-008-9365-8

ORIGINAL PAPER

Some dimension problems in molecular databases

Paul G. Mezey

Published online: 19 March 2008

Abstract Molecular databases obtained either by combinatorial chemistry tools or by more traditional methods are usually organized according to a set of molecular properties. A database may be regarded as a multidimensional collection of points within a space spanned by the various molecular properties of interest, the property space. Some properties are likely to be more important than others, those considered important form the essential dimensions of the molecular database. How many properties are essential, this depends on the molecular problem addressed, however, the search in property space is usually limited to a few dimensions. Two types of search strategies are related either to search by property or search by lead compound. The first case corresponds to a lattice model, where the search is based on sets of adjacent blocks, usually hypercubes in property space, whereas lead-based searches in databases can be regarded as search around a center in property space. A natural model for lead-based searches involves a hyperspherical model. In this contribution a theoretical optimum dimension is determined that enhances the effectiveness of lead-based searches in property space of molecular databases.

Keywords Molecular databases · Lead-based sampling in QSAR · Database dimension · Sampling errors in high dimensions · QshAR (Quantitative Shape-Activity Relations)

P. G. Mezey
Scientific Modeling and Simulation Laboratory, Department of Chemistry and Department of Physics and Physical Oceanography, Memorial University of Newfoundland, St. John's, NL, Canada A1B 3X7

P. G. Mezey (✉)
Institute for Advanced Study, Collegium Budapest Szentháromság u. 2, Budapest 1014, Hungary
e-mail: paul.mezey@gmail.com; pmezey@mun.ca

1 Introduction

With the rapid development and advancement of early approaches to combinatorial chemistry [1–4], molecular databases have experienced spectacular growth in recent years. Data mining techniques applied to such databases are becoming ever more important, and rational organization of search methods has acquired special significance. Organization of databanks are often accomplished in terms of various molecular parameters, molecular formulas, or even by sets of typical reactions. In particular, data entries into molecular databases obtained either by combinatorial chemistry tools or by more traditional methods are usually organized according to a set of molecular properties. In a mathematical sense a database may be regarded as a multidimensional collection of points within a space spanned by the various molecular properties of interest. This space is called the property space; some properties are likely to be more important than others, and the coordinates according to the former properties are regarded as those providing the essential dimensions of the molecular database. It is the molecular problem that determines how many properties are essential, however, the search in property space is usually limited to a few dimensions. One important application of molecular databases is in the study of quantitative structure-activity relations (QSAR).

One should point out a combinatorial chemistry aspect a database construction: due to the essence of combinatorial chemistry approaches, there exists a particular bias in the way new entries are made to the databank: the actual synthesis of a large number of molecules often follows a "controlled randomness", based on the class of reagents participating in the combinatorial chemistry process. There are historical reasons for this. The very first version of combinatorial chemistry, in a form applicable to the combinatorial synthesis of large numbers of molecules, was invented by Árpád Furka [1–4] in 1982. The use of combinatorial approaches to the production of very large numbers of molecules from specific building blocks has revolutionized synthetic chemistry, but it also reduced the variety of types as compared to the much slower traditional synthetic methods, where, nevertheless, accidental variations had a much more prominent role. Furka's original "portioning—mixing method for the synthesis of combinatorial libraries" and subsequent variations on the basic combinatorial chemistry principle were applied primarily to peptides, using individual amino acids as building blocks, hence the completeness principles were fulfilled to a high degree. However, the combinatorial chemistry principles have been generalized and extended to non-peptide molecules, and the combinatorial methodologies have also been adapted to types of molecules involving a much broader family of potential building blocks. By having a much greater variety of building blocks linked up according to various combinatorial patterns, there are some limitations required to avoid combinatorial explosion, and this also means the introduction of some biases. Hence, the databases obtained by such methods show an entry distribution that is far from uniform in property space, and the local point density of these spaces has an important effect on all data mining and search strategies.

Although the combinatorial chemistry advances have revolutionized the approaches used by the pharmaceutical industry for the production of new molecules, this has also initiated a new approach to database construction as well as data mining, for eventual

selection, tests, and optimization of the pharmacological effectiveness of potential drug molecules. Database sampling and search methods are an important stage in such processes.

2 Optimum dimension of property space for lead-based search in QSAR data-mining

Database searches may follow several strategies, however, two types of rational search strategies are related either to search by property value or search by lead compound.

The first case corresponds to a lattice model, where various intervals for the values of the relevant properties are considered, and search is based on sets of adjacent blocks with edges defined by these intervals, and the blocks usually form sets of joined hypercubes in property space, much like two-dimensional lattice animals on planar grids.

On the other hand, lead-based searches in databases can be regarded as search around a center in property space, where this center is defined by the properties of the lead compound. The search is focused on the vicinity of the lead compound in property space, and in this case not a set of property intervals, but a property space distance from this center is guiding the search. A natural model for lead-based searches involves a hyperspherical model, with a radius defined in terms of a property space distance condition.

Whereas both of the above search approaches have advantages, the intuitive evaluation of their efficiency may become very highly misleading in higher dimensions. As it has been pointed out earlier in the context of molecular databases [5], also relevant for electron density databases for combinatorial quantum chemistry [6] and more general computational database problems [7–15], the "capture rate" of hyperspherical searches in high dimensions is affected by a counter-intuitive feature of hyperspheres: in high dimensions the volume of hyperspheres of unit radius become very small as compared to the multidimensional volume element.

The local point density of entries in QSAR property spaces of databanks is practically never uniform; based on the traditional chemical information, as well as based on earlier lead compounds, there are typically local ranges which show much higher point density than the average density in the relevant part of property space. Hence, there are necessarily many factors influencing the effectiveness of search strategies, and the actual variations of local data density may counteract the effectiveness of any general data search strategy. However, general strategies are still valuable, and for their overall evaluation and ranking, it is often best to assume a fairly uniform distribution of data entries.

First we shall briefly review a property of hypersphere sequences in various dimensions, discussed earlier in the context of toxicological databases [5]. One may compare the dimension dependence of hypersphere volumes, and compare this volume to the unit volume of hypercubes in each dimension, Essentially, hyperspheres are the tools of lead-based search methods, whereas search by hypercubes in lattices is a tool of property-based search methods.

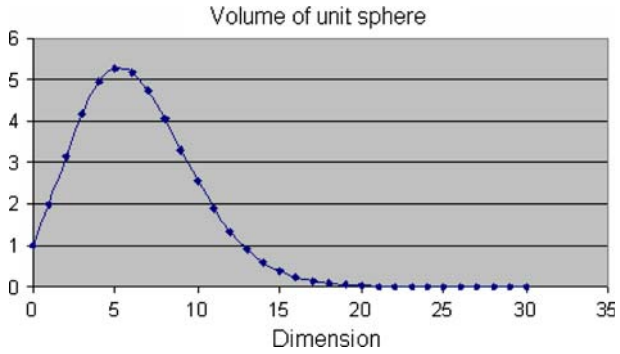

Fig. 1 The change of the volume $\beta(d)$ of hyperspheres as a function of dimension d

Two formulas [16,17] are commonly used to calculate the volume β(d) of a sphere of dimension d. The first such formula is

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

where the gamma function is given by the integral

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

The alternative formula is

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

using the int(y) function defined as the integer part of the number y, and the double factorial d !! of odd only or even only factors.

In Fig. 1 the volume of a hypersphere is plotted as a function of dimension for dimensions 0–30, noting that this volume converges to zero az dimension d grows to infinity.

The results for the first few dimensions are well known, however, the most important for our purposes are the actual quantities for dimensions between 4 and 7. The results for the first seven dimensions are given below in concise formula as well as numerically (for four decimals only):

$$\begin{array}{rccccccc} d = & 1, & 2, & 3, & 4, & 5, & 6, & 7, \\ & 2, & \pi, & (4/3)\pi, & \pi^2/2, & 8\pi^2/15, & \pi^3/6, & 16\pi^2/105, \\ & 2.0000, & 3.1416, & 4.1888, & 4.9348, & 5.2638, & 5.1677, & 4.7248, \end{array}$$

respectively.

We may note that the volume is maximum in dimension 5, this value of 5.2637890136135... is actually only slightly higher than the volume 5.1677127796068 ... for dimension 6.

The shape of the distribution of volumes and the presence of a maximum has implications on lead-based searches in property spaces of molecular databases. If no additional information is available, it is natural to assume some average point density of entries in the database within the relevant range of the property space. Since in high dimensions the hyperspherical volume of unit radius becomes negligible as compared to the volume of the unit cube (that is by definition equal to 1 in every dimension, since the hypercube is the volume element), the "capture efficiency" of the hyperspherical approach of lead-based searches becomes reduced, when compared to grid-based (property-based) searches. In high dimensions the volume of the hypercube has a far better chance of actually capturing data entries, than the diminishing volume of the hypersphere.

However, at relatively low dimensions, the unit radius hypersphere captures a larger volume than the hypercube, and in dimension 5, this volume is maximum. Consequently, if the actual property space where the search is carried out has dimension 5, than the lead-based hyperspherical search is the most effective, assuming uniform distribution of data entries. Dimension 6 is almost as advantageous than dimension 5, however, hyperspherical search in much higher dimensions become rapidly less efficient for capturing data points.

Of course, the nature and number of properties which are important for a given chemical problem are determined by the problem itself. However, by a suitable ranking of the importance of various molecular properties, a hierarchy can be established, and one should be able to select a low number of "active dimensions" of the search space with respect to these most important properties. If the dimension of this "active search space" is 5, then one is able to carry out the hyperspherical, lead-based search under optimum dimension conditions. Note, however, that dimensions only slightly different from 5 are still effective for hyperspherical capture of data entries, and the intuitive idea that a unit sphere has more volume than a unit cube is valid for all low dimensions, in spite of the fact that this intuitive idea fails in a dramatic way in high dimensions.

3 Summary

In this contribution a theoretical optimum dimension of the active search space has been determined that enhances the effectiveness of lead-based searches in property spaces of molecular databases. In the application of molecular databases in the study of quantitative structure-activity relations (QSAR), the effective search dimension usually can be restricted in order to remain near the optimum dimension five. Database approaches have been extended to the ultimate molecular information carriers, to molecular electron density clouds and their shapes [7–11], and databases of computed electron density information is used in quantitative shape-activity relations (QShAR), where effective search methods are also important.

Acknowledgements The research support received from Natural Sciences and Engineering Research Council of Canada, the Canada Foundation for Innovation, the Canada Research Chairs Program, and from the Memorial University of Newfoundland, Canada, as well as the hospitality of Collegium Budapest, Hungary are gratefully acknowledged.

References

1. A. Furka, Notarized notes (1982), http://www.win.net/kunagota
2. A. Furka, F. Sebestyen, M. Asgedom, G. Dibo, in *Highlights of Modern Biochemistry. Proceedings of the 14th Internatational Congress of Biochemistry*, vol. 5 (VS Publishers, Utrecht, The Nederlands, 1988), p. 47
3. A. Furka, F. Sebestyen, M. Asgedom, G. Dibo, Abstract P. 168, in *Abstract, 10th International Symposium on Medicinal Chemistry*, (Budapest, 1988), p. 288
4. A. Furka, F. Sebestyen, M. Asgedom, G. Dibo, Int. J. Pept. Protein Res. **37**, 487–492 (1991)
5. P.G. Mezey, P. Warburton, E. Jako, Z. Szekeres, Dimension concepts and reduced dimensions in toxicological QShAR databases as tools for data quality assessment. J. Math. Chem. **30**, 375–387 (2001)
6. P.G. Mezey, Computational aspects of combinatorial quantum chemistry. J. Comput. Meth. Sci. Eng. (JCMSE) **1**, 99–106 (2001)
7. P.G. Mezey, The degree of similarity of three-dimensional bodies; applications to molecular shapes. J. Math. Chem. **7**, 39–49 (1991)
8. P.G. Mezey, *Shape in Chemistry : An introduction to Molecular Shape and Topology* (VCH Publishers, New York, 1993)
9. P.G. Mezey, Quantum chemistry of macromolecular shape, Int. Rev. Phys. Chem. **16**, 361–388 (1997)
10. P.G. Mezey, Quantum chemical shape: new density domain relations for the topology of molecular bodies, functional groups, and chemical bonding. Can. J. Chem. **72**, 928–935 (1994) (Special issue dedicated to Prof. J.C. Polanyi)
11. P.G. Mezey, Functional groups in quantum chemistry. Adv. Quantum Chem. **27**, 163–222 (1996)
12. P.G. Mezey, Local electron densities and functional groups in quantum chemistry. in *Topics in Current Chemistry, Correlation and Localization*, vol. 203, ed. by P.R. Surjan (Springer-Verlag, Berlin, Heidelberg, New York, 1999), pp. 167–186
13. P.G. Mezey, Molecular similarity, quantum topology, and shape. in *Computational Medicinal Chemistry and Drug Discovery*, ed. by P. Bultinck, J.P. Tollenaere, H. De Winter, W. Langenaeker (Marcel Decker Inc., New York, 2004), pp. 345–364
14. J.-E. Dubois, P.G. Mezey, Relations among functional groups within a stoichiometry: a nuclear configuration space approach. Int. J. Quantum Chem. **43**, 647–658 (1992)
15. J.-E. Dubois, P.G. Mezey, A functional group database: a charge density—DARC approach. Mol. Eng. **8**, 251–265 (1999)
16. M. Berger, *Geometry* (Springer-Verlag, Heidelberg, 1987)
17. J. Bourgain, V.D. Milman, New volume ratio properties of convex symmetric bodies in R^n. Invent. Math. **88**, 319–341 (1987)