Chimia 48 (1994) 112
© Neue Schweizerische Chemische Gesellschaft
ISSN 0009–4293

3D-Databases, Database Mining and Electronic Screening

Hans Peter Weber*

In the context of this short article, the term '3D-database' is used for an ordered and formatted collection of molecular structures in a database containing 3D-coordinates fully describing the conformation of each entry, and possibly other molecular properties such as *e.g.* partial charges. In many cases the 3D-database is associated to a '2D-database' in which the same molecules are described by '2D-data' such as connectivity, bond orders, definition of chiral centers, and other characteristic items (*e.g.* chemical name, internal registry number, cross-references to other databases).

Large and public 3D-databases of organic and biological compounds exist since more than 25 years; the most famous among them are the Cambridge Crystal Structure Database, comprising today more than 100 000 experimentally determined crystal structures of organic compounds, and the Brookhaven Protein Data Bank with some 1 000 experimental biopolymer structures. Both have been maintained since many years and are continuously updated, and both represent a source of structural information of perennial validity since they are not based on theoretical models (of varying validity) but on established *experimental* results. The Cambridge Crystal Structure Database has traditionally been a full database system, *i.e.* the Cambridge Crystal Data Centre not only distributes a comprehensive collection of published organic crystal structures, but in addition a complete software system allowing intelligent retrieval of structural information. The Brookhaven Protein Data Bank, however, is a simple depository of carefully checked and documented biopolymer structures, based on published X-ray and NMR structural analyses.

Since a few years a new type of 3D-databases is emerging and gaining wide interest, particularly in pharmaceutical research laboratories: large, proprietary 3D-databases consisting of molecular structures are created from a parent 2D-database by *theoretical* methods. The build-up and extensive use of 2D-chemical databases – cross-referenced to biological data – has a long tradition in pharmaceutical research. These in-house 2D-databases, usually consisting of hundreds of thousands of entries, represent an invaluable treasure for a company: the accumulated results of many years of pharmaceutical research. Continuously updated these databases are a source of great potential for *lead finding* and *rediscovery*. Although a lot of information can be retrieved from 2D-chemical databases using intelligent software and clever query formulations, one hopes to further increase the potential use of this treasure by adding the 3D-database to it. With this in view, quite some effort has recently been devoted to create 3D-databases in an automated procedure from 2D-databases, *i.e.* to build 3D-molecular structures starting from the 2D-molecular description. Many programmes which do that have been written, among the best known are, *e.g.*, CONCORD (R.S. Pearlman, TRIPOS Ass., St. Louis), WIZARD/COBRA (D.P. Dolata, *J. Comp. Aided. Mol. Des.* **1988**, *2*, 107, CORINA (J. Gasteiger, *Anal. Chim. Acta.* **1992**, *265*, 233), to name just a few.

As clear as the goal is, as difficult it is to achieve it in practice: There are many problems involved in these 2D → 3D conversion programmes: one of the most prominent one is speed. Since these conversion programmes are designed to be applied to large databases, an upper limit of *ca.* 1 s in average per structure generation on a reasonably modern computer is acceptable, if the procedure consumes 10 s or more per structure, the processing times to produce (and update) a 3D-database is becoming unrealistic. This restriction allows model bulding only by fast algorithms like *e.g.* a distance geometry, or build-up by connecting fragments from a library, but it would exclude full structure optimization by molecular mechanics. These fast procedures produce generally acceptable, low quality 3D-models, which usually have correct local stereochemistry but often have bad contacts between remote parts. Another problem is that these methods will produce just one, and always the same, molecular conformation, which may be stereochemically reasonable but not necessarily representative for a low (or minimum) energy conformation. Connected to this problem is the treatment of molecular *flexibility*, or how to generate a representative ensemble of low energy conformations: again here the speed factor constrains to severe compromises. Although these problems in 2D → 3D conversion appear to be purely technical, which will be solved as more powerful computers become available, there are in fact still a number of unsolved basic problems, *e.g.* the conformation of rings larger than six, or the representative sampling of the conformational space.

In spite of all these problems, mostly recognized by pharmaceutical research management, there is a lot of effort put into the buildup of such 3D-databases. The driving force behind it and the main purpose are two novel, and potentially powerful approaches to drug design and drug discovery: '*database mining*' and '*electronic screening*'. With database mining it is hoped to locate molecules in a 3D-database which have a similar 3D-shape, or a similar 3D-disposition of several functional groups (*i.e.* a common pharmacophore), to a reference compound, thus finding (existing) compounds which may become a new lead. With the other method, electronic screening, one tries to dock the molecular models of a 3D-database one by one into a protein binding site, calculating a figure of fit for each and retaining the top-scoring compounds. These molecules, a selection of say 100 compounds 'electronically screened' out of some 100 000 candidates, will then be tested in a biological assay, hoping that by this procedure the probability to find the active compounds is (almost) as good as testing the complete set of 100 000 compounds.

Thus, the rational behind these efforts is obvious, and even if some of the methodology applied at present is still a bit dubious, database mining and electronic screening are believed to make a selection of compounds, if not optimal, so still better than random. And the hit rate will improve as the quality of 3D-databases is improving, and as the problems of molecular flexibility are being solved, and as shape recognition and docking procedures are getting better. It is a difficult and costly enterprise to do all this, however, the prospect of better exploiting the accumulated wealth of data in a company's most valued treasure, seems worthwhile.

**Correspondence*: Dr. H.P. Weber
Sandoz Pharma AG
Preclinical Research Discovery and Technology
CH–4002 Basel