22.4. The relevance of the Cambridge Structural Database in protein crystallography

By F. H. Allen, J. C. Cole and M. L. Verdonk

22.4.1. Introduction

At its inception in the late 1960s, the Cambridge Structural Database (CSD: Allen, Davies *et al.*, 1991; Kennard & Allen, 1993) was one of the first scientific databases for which numerical data were the primary objective of the compilation. Thus, the CSD provides not only a fully retrospective bibliography of the structure determination of organic and metallo-organic compounds, but also gives immediate access to the primary results of each diffraction experiment: the space group, cell dimensions and fractional coordinates that define each structure at atomic resolution. In the late 1960s, the world output of small-molecule structures was just a few hundred per year and it was possible to use existing printed compilations to ensure that the developing CSD was fully retrospective. Despite this comprehensive nature, it has taken time for the CSD to have significant scientific impact as a research tool in its own right, and to be recognized as a source of structural knowledge that is applicable across a broad spectrum of structural chemistry.

There are two reasons for this rather gradual uptake. First, it took time to devise and implement software for the validation and organization of the data. Secondly, and most importantly, it was necessary to develop software for database searching, particularly for locating chemical substructures, and for data analysis and visualization. It was not until the late 1970s that the first comprehensive software systems became available and began to be widely distributed to scientists in academia and industry. Nevertheless, a number of highly influential database analyses were performed prior to 1980, and the proper numerical analysis and statistical treatment of bulk geometrical data began to receive attention (see *e.g.* Murray-Rust & Bland, 1978; Murray-Rust & Motherwell, 1978; Taylor, 1986). This software and its successors at last allowed the types of geometrical surveys, analyses and tabulations carried out manually by early practitioners such as Pauling (1939), Sutton (1956, 1959) and Pimental & McClellan (1960) to be executed automatically in a few minutes of increasingly powerful CPU time.

The early development of applications software simultaneously with methods for the acquisition and validation of new structural data was crucial for the CSD. Developments in structure-determination theory, allied to technological improvements in data collection and the ever increasing speed and capacity of modern computers, led to such a rapid expansion that the archive of May 1999 now contains more than 200 000 crystal structures, a total that doubles approximately every seven years. The literature is now so vast, so chemically diverse and so widely spread that it is virtually impossible for individual scientists to maintain current awareness without recourse to database facilities. It is now impossible to carry out viable systematic analyses without recourse to database technology. This chapter focuses primarily on the structural knowledge that is provided by such analyses, and that is relevant to the determination, refinement, validation and systematic study of macromolecular structures. However, the validity of these results depends crucially on two factors: the *completeness* of the archive and the *accuracy* with which the data are recorded. Hence, it is appropriate to preface the chapter with some comparative comment on these fundamental issues as they apply to the small-molecule and macromolecular structure archives.

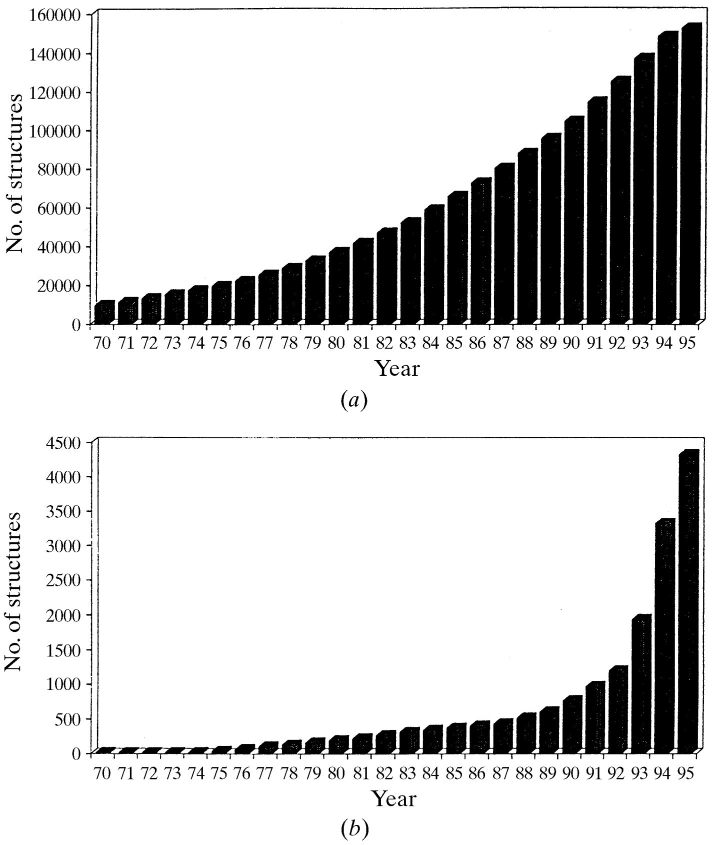

Fig. 22.4.2.1. (*a*) Growth rate of the CSD and (*b*) growth rate of the PDB, in terms of the numbers of structures published per annum for the period 1970–1995.

22.4.2. The CSD and the PDB: data acquisition and data quality

22.4.2.1. *Statistical inferences*

With a current total of 200 000 structures and a doubling period of seven years (Fig. 22.4.2.1*a*), we may expect at least half a million small-molecule crystal structures to be in the CSD by the year 2010. The Protein Data Bank (PDB) (Abola *et al.*, 1997; Berman *et al.*, 2000), which began operations in the mid-1970s, and the Nucleic Acid Database (NDB) (Berman *et al.*, 1992) are the international repositories for macromolecular structure information. Input to the PDB was initially slow but is now showing a rapid growth rate reminiscent of the CSD of the 1970s (Fig. 22.4.2.1*b*). The PDB archive has a current total of *ca* 8500 structures (mid-1999) and a doubling period of close to two years. As with the CSD, this early *high rate* of growth will almost certainly decrease, thus increasing the doubling period. Nevertheless, by the year 2010, we might expect the PDB to contain more than 100 000 structures.

22.4.2.2. *Data acquisition and completeness*

Given the size and diversity of the CSD, it is amazing that searches for some common chemical substructures often yield far fewer hits than might have been expected. Sometimes, the absence of just a few key CSD entries would have negated a successful systematic analysis: some points in a graph would have been missing and a correlation would not have been detected. Similarly, completeness of the PDB is vital for the future of 'data mining' or 'knowledge engineering' in the macromolecular arena.

Data acquisition by the PDB has always had one valuable advantage in comparison with the CSD. The volume of numerical data generated by a protein structure determination is far too large

references