Preface

The Collaborative Computing Projects [CCPs (http://www.dl.ac.uk/CCP/main.html)] were set up originally under a UK Science Research Council initiative to encourage collaborative computational work between UK groups. CCP4 is the Collaborative Computational Project No. 4 in protein crystallography (http://www.dl.ac.uk/CCP/CCP4/main.html), now funded by the UK Biotechnology and Biological Sciences Research Council and administered from Daresbury Laboratory (a UK national laboratory which was founded in 1962 and is now part of the Central Laboratory of the Research Councils). CCP4 is perhaps best known for its suite of programs which are created and developed by working crystallographers. The programs are distributed in source form, and are intended to be a resource for crystallographers worldwide. Both the use of and contribution to the suite are encouraged.

CCP4 also runs various courses and workshops, each dealing in depth with a specialized area of protein crystallography such as data collection, molecular replacement or refinement. In particular, since 1980 CCP4 has organized study weekends, now held on a regular annual basis, which are aimed primarily at teaching postgraduate and postdoctoral crystallographers about both the fundamentals and the latest advances in some aspect of the subject.

In the past the proceedings of these study weekends have been available as technical reports from the Daresbury Laboratory, and have proved a valuable reference source of specialized information not always obtainable elsewhere. This method of distribution was simple when the numbers of participants and the number of protein crystallographers were both relatively small, but recent weekends have attracted attendances of 300–400, with several times that number potentially wishing to obtain the proceedings. Therefore, the decision was made to reach a wider audience by publishing the proceedings in *Acta Crystallographica Section D*.

Therefore, the articles in this issue are from the Study Weekend on Databases for Macromolecular Crystallographers held in January 1998 at Reading University.

The explosion of macromolecular structural determinations is well documented, and has implications for the validation, deposition, curation, exploration and exploitation of data, which are discussed in several papers in this volume. The crystallographic community showed great foresight in the establishment of central databases (CCDC, PDB); but these are not static entities, and have a continuing need to respond to increases in the volume of data, developments in good practice in data storage and handling, and to an increasingly diverse user community. Databases derived from this primary data, either focusing on a specialized area or classifying some aspects of the structures, are valuable resources; they require both the optimum quality of deposition to the primary databases and an appreciation by the user of the principles on which they are created, and any inherent limitations.

Database use now underpins many stages of a structure determination, from sequence searching through solution, modelling and refinement to structure comparison. In parallel with the data explosion, increased accessibility of numerous search and comparison tools, many available *via* the Internet, has provided a large, and at times bewildering, choice for a structural biologist. Within the confines of a two-day workshop comprehensive coverage of the field of databases is impossible, but speakers were chosen to illustrate current areas of activity, in the hope of fostering an appreciation of some of the considerations involved both in creating and using databases intelligently. A complete list of speakers is available on the WWW at http://www.cryst.bbk.ac.uk/~ubcg09j/ccp4urls_1.html.

The introductory article by P. Murray-Rust explores the revolution in technology and outlook in which we are participating, willingly or not. The path data $\rightarrow$ information $\rightarrow$ knowledge $\rightarrow$ wisdom is not straightforward, and the whole community needs to be aware of the issues, rather than devolving them to information specialists.

The creation and evolution of the *Iditis* database (S. Gardner) and its underlying programs explains some of the design considerations in creating a robust database and flexible query system.

The next three papers are concerned with comprehensive databases of protein and nucleic acid structure – the primary databases upon which other derived and more specialized databases rely. J. Sussman *et al.* describe the Protein Data Bank, in terms of its history, the transition to automated deposition of data and future plans for conversion to the new 3DB-Base, a relational database system which should facilitate both archiving and querying. The Nucleic Acid Database (NDB) (Berman *et al.*) was created as a database of DNA and RNA structures, which would also provide specialist information and tools for understanding nucleic acid structures. It is now a direct deposition site for these structures; the underlying technology has also been applied to other macromolecular databases. P. Bourne discusses some of the deficiencies, anomalies and discrepancies in present practice, and suggests some ways in which these might be remedied in the future.

The next two papers are concerned chiefly with the processes of data deposition and validation, and the potential extension to recording material for deposition during the course of the structure determination. Keller *et al.* explain the philosophy behind the current auto-

matic deposition procedure with *AutoDep*, and discuss the prospect of data harvesting, a major change in the way of working which will require the collaboration of both software writers and instrument manufacturers to achieve its full potential. E. Dodson then discusses the importance of validation techniques, and in particular the work of the EU-funded network project on validation, and their relationship with refinement. This has wider implications for the refereeing process since until now journals, *via* their referees, have had full responsibility for judging a structure, but requirements for deposition and validation imply a refereeing role for the PDB also.

Model validation and analysis form a part of G. Kleywegt's use of databases as implemented in the suite of programs from the Uppsala Software Factory. Tools to assist structure determination also include those for the creation of hetero-compound dictionaries and for fold recognition. J. Thornton presents several new databases for analysing structures including TOPS (protein topology) and of protein–nucleic acid interactions, and a WWW server for analysis of interactions in multimeric structures. We were also reminded of the useful material provided by CCP11, a sister CCP concerned with sequences and genomics. Sequence alignment is offered by several WWW sites, and G. Barton suggests ways of producing optimal results.

The three following papers (Hubbard *et al.*; Orengo *et al.*; Sowdhamini *et al.*) are all concerned with classifications of protein structures. The last few years has seen the production of several domain databases and these three examples show something of the similar, but non-identical, approaches to classification and searching.

The remaining papers are primarily concerned with the ligands and metal atoms in proteins.

ReLiBase (M. Hendlich) is a new database which can be used to identify and analyse ligands and protein–ligand complexes. The inclusion of chemical and biochemical data together with good query tools provides a good tool for rational ligand (drug) design. The Cambridge Structural Database (CSD) is the repository for small-molecule structures (other than purely inorganic), and has been used by protein crystallographers, for example in the derivation of target bond lengths and angles for refinement. The Cambridge Crystallographic Centre also provides several programs for analysis of structural data, and the example chosen here (R. Taylor *et al.*) is that of hydrophobic interactions. G. Orpen describes detailed geometry around metal ions determined by analysis of the CSD. Tabulations of bond lengths for *d*- and *f*-block metals are organized by ligand contact atom and by the type and bonding mode of the ligand. Use of the CSD to detect intermolecular interactions is illustrated by the case of metal chloride bonds acting as hydrogen-bond acceptors.

Islam *et al.* have assembled data on the preparation and characterization of heavy-atom derivatives into a heavy-atom database (HAD). Much of this data was gleaned from the literature, and by contact with individual crystallographers and is not readily available elsewhere.

One speaker was unable to provide a written version of their talk; B. Hardy spoke on the carbohydrate resources available on the WWW, which include structure, nomenclature and chemistry.

Judith Murray-Rust
Liz Potterton
Ben Luisi
Eleanor Dodson
Sue Bailey

CLRC Daresbury Laboratory,
Warrington WA4 4AD, England