DARESBURY LABORATORY

INFORMATION QUARTERLY for PROTEIN CRYSTALLOGRAPHY

An Informal Newsletter associated with Collaborative Computational Project No. 4 on Protein Crystallography

Number 2 January 1980

Contents

Editor: Pella Machin — Science Research Council, Daresbury Laboratory, Daresbury, Warrington WA4 4AD, England

Deputy Editor for Imperial College: Dr. Alan Wonacott — Imperial College of Science & Technology, The Blackett Laboratory, Prince Consort Road, London SW7 2BZ

Deputy Editor for Birkbeck College: Dr. David Moss — Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX

EDITORIAL

P.A. MACHIN

This second newsletter follows after another meeting of Working Group 2 and the minutes of the meeting are included here for information. They detail progress reports given by the groups as well as further discussions on data formats.

The subject of data formats in the United States was raised at the meeting and later followed up by Mike Elder. He reports on his findings here.

Finally, I include some miscellaneous brief progress reports which may be of interest to other groups: Trevor Greenhaugh (Keele) reports that the first part of OSCAR (for off-line film processing) is now working: Phil Bourne has documentation for the programs used at Sheffield including Modelfit, DISTAN and TORSION: John Campbell has the Isaacs Agarwal refinement program running for 5 space groups: the Cambridge version of PLUTO (for plotting protein molecules and maps) is at DL and has been tested for molecules (but not yet maps) - output is on the single pen Calcomp or Versatec: CAD has been implemented and is under test - at the moment it outputs "Dodson binary format" but it is intended to add the new NA2 format as an option, when it may prove a convenient program for changing between formats.

MINUTES OF A MEETING OF WORKING GROUP 2 OF THE PROTEIN CRYSTALLOGRAPHY COLLABORATIVE COMPUTATIONAL PROJECT WHICH WAS HELD AT IMPERIAL COLLEGE, LONDON, AT 11.00 ON 4TH DECEMBER, 1979

Present:	Dr. Alan Wonacott (Chairman)	Imperial College, London
	Dr. T. Bhat	" " "
	Dr. David Moss	Birkbeck College, London
	Dr. Ian Tickle	" " "
	Mrs. Eleanor Dodson	York University
	Dr. Phil Bourne	Sheffield University
	Dr. Bob Stansfield	" "
	Dr. Trevor Greenhaugh	Keele University
	Dr. Bob Diamond	M.R.C., Cambridge
	Dr. Phil Evans	" "
	Dr. Mike Elder	S.R.C., Daresbury
	Dr. John Campbell	" "
	Miss Pella Machin (Secretary)	" "

Apologies from: Dr. Keith Wilson

1. The progress of all groups had been severely restricted because of the problems with the Daresbury IBM 370/165 computer. These problems had been particularly severe over the last month. It was agreed that a letter should be sent to Brian Davies (with copy to Prof. Burke) expressing the group's dissatisfaction and that if any constructive suggestions were forthcoming they should be passed on. The Imperial and Birkbeck groups agreed to meet later to construct this letter.

2. Implementation of the Munich system PROTEIN was considered. It was agreed that Ian Tickle should find out more about the system so that if necessary a proper case for its implementation could be written.

   In general it was felt that although this appeared to be a good system which would be particularly useful for new groups it was not what was wanted for the main CCP program system. It was agreed that a more flexible system was required so that programs could be added and removed to obtain the"best programs".

/cont.

3. Progress reports were given by the various groups.

(i) David Moss reported that hardly any work had been carried out on the data processing programs because of the machine problems with the IBM. In fact Birkbeck had resorted to using the Cambridge IBM for some runs.

DIFVAL, PDPPLUS, HRS4, KSCALE, FCALCA, FCALCB, SFLSP, CONTOUR were all running at Daresbury, however, DIFVAL had not actually been used to process data because of further problems encountered when trying to read and transmit paper tape data at the workstation.

(ii) Bhat reported that the programs ROTOVATA, AGROVATA and GENDEN were all going at Daresbury. PROTIN and PAIROT were not yet running. Again great problems had been encountered because of the poor machine response. Further work had been done on implementing the NA 2 reflection data format and this is detailed in a later section (5 (i)).

(iii) Phil Bourne reported the activities of the Sheffield group. The programs MODLFT and DISTAN as implemented by DL were in use and improved CLISTS and writeups were available. A torsion angle program (TORSION) was also working. The majority of effort was being spent on consideration of refinement methods available for the high symmetry space group F432.

(iv) The Keele group was implementing the program OSCAR (from Oxford) for off-line rotation camera photograph processing.

Eleanor Dodson said that she had spent some effort rationalising the FFT routines and modifying the Isaacs-Agarwal refinement (particularly concerning its treatment of form factors).

(v) The activity at Daresbury had also been severely limited by machine problems. The main effort had been put into implementing the latest York versions of the Isaacs-Agarwal refinement, and the versions for space groups 19 and 92 were almost working.
John Campbell had also given some thought as to the form of a full

/cont.

program suite and the data formats which would be necessary in such a system. The programs CAD (from York) and PLUTO (from Phil Evans at Cambridge) were being worked on.

It was agreed that the Phil Evans version of PLUTO which plots map and molecule (but without map extension or removal) should be implemented as soon as possible at Daresbury (using 3 colour pens) for the 3 devices FR80, Versatec and Calcomp. It was agreed that Hugh Savage (Birkbeck), who wanted this program urgently, be enlisted to help with its implementation, and that Eleanor Dodson and Phil Evans should be available to act as consultants.

It was noted that as agreed by working group 1 CCP funds should if possible be available to help with this sort of work; for example, to contribute to travel costs.

(vi) It was further noted that the CRAY computer was available at Daresbury and that the Oxford group (Bill Pulford) had run the Konnert refinement program on the CRAY with gains in speed of a factor of 15 over the ICL 2980. Oxford are planning to use the CRAY for production runs as soon as file transfer from the CRAY to the IBM is available.

4. Consideration was given to the "standard Crystallographic File Structure" proposed by the international working party set up by the computing and data commissions of the International Union.

(i) Bob Diamond, a member of the IUCr working party had circulated relevant documentation to group members before the meeting and requested their consideration of this topic.

(ii) Alan Wonacott introduced the topic pointing out that the International format was mainly concerned with formats for the interchange of data, whereas previous discussions by the CCP group related to binary file formats for internal use. Both formats have a use.

(iii) Bob Diamond gave some background to the decisions which had been made by the panel and stressed that the proposed file structure was designed for and by small molecule crystallographers. As a representative of the protein crystallographers Bob wished to gather the opinions of the assembled group so that he could report their views to the International Panel.

/cont.

(iv) Atom card format.

Everyone agreed that the proposed format did not allow sufficient space in the atom identifier field for protein work, where residue number and name were necessary in addition to the atom name. A total of 12 characters are needed for protein work. It was suggested that the standard deviation field in the proposed format could be used for this extended atom identifier in the protein case.

(v) It was agreed that the definition of the HKL card was insufficient for protein work and it was thought that in this area the protein work requirements differed from the small molecule work significantly. It was suggested that an alternative 'card' HKLPROT could be introduced to cope with the necessary additions and that the following classes of data should be allowed.

h k L F σ Δ anom σΔ anom

2 phases (most probable and centroid) and figure of merit

A B C D coefficients.

Various sets of Fcalc.

5. A discussion followed on the suggested CCP data formats.

Two items of information were circulated to the group, the first being a letter from Professor Tony North advocating an alternative co-ordinate data format and the second being a letter from George Reeke relating to data formats in the U.S.

(i) h k L format.

Unanimous support was given to the NA2 format which was being developed and implemented by the Imperial group. Some further details of the implementation of this format were distributed and some concern was expressed that it might be too complicated to use

/cont.

(for example in "jiffy programs") in its current form. It was agreed that Alan Wonacott, Bhat, David Moss, Ian Tickle should form a sub-committee to discuss the details of the implementation of this format.

(ii) Co-ordinate data formats.

Due consideration was given to the arguments put forward by Professor North and proponents of other xyz formats. However, the group finally agreed that the Brookhaven co-ordinate format should be adopted on a trial basis. This format is used internationally and contains all necessary information without any obvious problems. The atom card format to be adopted is therefore

6A1, I5, 1X, A4, A1, A3, 1X, A1, I4, A1, 3X, 3F8.3, 2F6.2, 1X, I3

for

1-4	ATOM
7-11	Atom serial number (starting from N terminal)
13-16	Atom name
17	Alternate location indicator
18-20	Residue name
21-27	Sequence identifier
31-38	X)
39-46	Y } Orthogonal Å co-ordinates
47-54	Z)
55-60	Occupancy
61-66	Temperature factor
68-70	Footnote number

It was noted that the naming conventions for water molecules often cause problems.

6. (i) To enable groups to communicate on a daily basis the Daresbury group had set up a TSO file PC.NEWS. TEXT (i.e. under the identifier PC) on the Daresbury computer.

/cont.

This file is available for access by all groups and it can be listed or modified. The file itself contains details of the 'rules' imposed and of methods of adding messages to the front of the file.

John Campbell agreed to assume responsibility for the housekeeping associated with the file.

(ii) A discussion followed on available documentation methods.

Mike Elder suggested that of the 3 main resources available - word processor (at DL), IBM computer (at DL), typist - that the word processor was the best solution at the present time. Machine methods were favoured above a typist for ease of editing and updating. The current IBM computer problems meant that TSO response and machine time were sufficiently valuable that manual documentation assumed a low priority compared with important job submission and file editing, and many hours at a terminal typing in a manual could not be justified.

However, Phil Evans pointed out that for the purpose of distribution manuals should be in machine readable form and that existing manuals on magnetic tape should be used for input for documentation.

Ian Tickle suggested that an advantage of "computer manuals" was their availability to computer users at the terminal.

Taking these points into consideration it was agreed that existing documentation in machine readable form be kept. However, accepting the current problems with the IBM it was thought that the potential of the word processor should be explored further, in particular regarding the possibility of transfer of data from its store to other machines.

/cont.

NRCC - NEWS OF AN AMERICAN INITIATIVE

DR. MIKE ELDER (DARESBURY LABORATORY)

A National Resource for Computation in Chemistry (NRCC) has been set up in the United States. In many respects the functions of the NRCC parallel those of the various Collaborative Computational Projects at Daresbury. In particular, the NRCC has a crystallographic section which is concerned with the standardization of data formats and program systems, so it seems important that readers of this newsletter should be aware of developments at NRCC and that there should be interaction between the two groups.

The functions of the NRCC are broadly categorized as:

1: to make information on both new and existing computational methods available throughout the chemistry community;

2: to make state-of-the-art computational facilities (hardware and software) available to the community;

3: to foster research and development of new computational methods in chemical problems.

The first phase of the project was the establishment of a group at the Lawrence Berkeley Laboratory of the University of California under Dr. W.A. Lester, Jr. Staff members co-ordinate activities in chemical kinetics, crystallography, quantum chemistry etc. Phase two, which may involve the purchase of computing facilities, awaits the evaluation of the first phase.

The crystallographer on the scientific staff of NRCC is Dr. Arthur J. Olson. He edits a NRCC newsletter entitled "Computers in Crystallography"[1], which originated from the crystallography group at a conference on software

/cont.

standards in Utah, July, 1979. The first edition of the newsletter contains the crystallographic report from the conference, which is worth summarizing since it contains so much of relevance to CCP4. In answer to the question: why bother about software and data format standards now, after 25 years of crystallographic programs? the group state their purpose:

1: to facilitate program exchange among laboratories and various processors;

2: to enable programs from different people, different places to access a common, standard data file;

3: to avoid the manpower waste associated with computer changes;

4: to minimize the dislocations which will occur if FORTRAN compilers are revised;

5: to ensure that as new computational methods are developed they are programmed in a portable and machine-independent fashion.

I am sure that no one would disagree with these aims: they are a somewhat idealistic statement of the aims of CCP4. The recommendations may contain some surprises.

1: FORTRAN standardization: concern was expressed that the proposed FORTRAN 77 standard will not be compatible with existing FORTRANS. NRCC should encourage programmers to use PFORT[2], and code supported and distributed by NRCC should be accompanied by a PFORT verification report.

2: RATMAC: the use of RATMAC[3], a macro-enhanced version of RATFOR[4], is recommended for crystallographic software development. RATMAC has the advantages that it is freely available, maintained by a group concerned with crystallographic software portability. It is a preprocessor language with structured programming capabilities, producing standard FORTRAN code with macro calls to handle machine dependencies.

3: Binary Data File: the BDF designed by Jim Stewart et al for the XTAL80 system[5] be used by as many programs as possible, whether or not they are part of XTAL80. NRCC to distribute software and documentation.

/cont.

4: IUCr Formatted File: NRCC to remain aware of progress toward the definition of an international standard for archival storage and exchange of crystallographic data and encourage programmers to provide code for reading and writing this file.

5: XTAL80 Conventions: NRCC should support financially the development of a pilot system of crystallographic programs using the standardized coding conventions of the XTAL80 system.

6: Computer Graphics: encourage users to follow the graphics standard proposed by SIGGRAPH[6] in order to promote device independence.

Since the CCP4 Working Group 2 briefly considered and conclusively rejected the possibility of using an XRAY-system-like BDF, for protein crystallographic programs it may be as well to record NRCC experience. As a follow up to recommendation 5 NRCC set up a workshop of 10 scientists in November, 1979. The participants (Richard Alder, Steven Freer, Robert Munn, George Reeke, Steven Sheriff, James Stewart, Jurgen Sygush, Lyum Tentyck, Keith Watenpaugh and Sid Hall) had 8 days to design, write and test a program that would represent the state of the art in multiple isomorphous replacement phasing, in the functional environment of Stewart's XTAL80 system including the BDF and nucleus routines. In the event, 8 days sufficed to write and compile all the major pieces of code, and the participants were pleased with the progress of their modular co-operative programming. A future workshop will be held to tune, optimise and test the code.

I have asked Arthur Olson for more information and will report NRCC news in future issues. Meanwhile, readers may like to think about the following questions before our next meeting:

1: should we reconsider our attitude to standardized binary files of crystallographic data, and perhaps experiment with the XTAL80 BDF?

2: should we attempt to gain experinece with RATMAC and for PFORT with a view to perhaps standardizing upon them or equivalent alternatives?

3: should the CCP send a representative to NRCC in order to promote co-operation and program exchange?

/cont.

References:

(1) NRCC - "Computers in Chemistry". Vol. 1, No. 1, Winter 1979.
Editor: A.J. Olson, LBL, Bldg B50B, University of California,
Berkeley, California 94720.

(2) B.G. Ryder - "Software Practice and Experience". Vol. 4, 359-377 (1974).
PFORT (portable FORTRAN) is a verifier which determines the degree of
agreement with the ANS-FORTRAN standard.

(3) R.J. Munn and J.M. Stewart, University of Maryland Computer Sciences Centre,
Technical Report TR-675 (1978).

(4) B.W. Kernigham and P.J. Plauger - "Software Tools", Addison-Welsey,
Reading, Mass. 1976.

(5) S.R. Hall and J.M. Stewart, University of Maryland Computer Sciences Centre,
Technical Report TR-700 (1978).

(6) Computer Graphics 13 ~~#~~3 (1979).

BINARY DATA FORMAT FOR REFLECTION DATA

DR. T.N. BHAT (IMPERIAL COLLEGE)

Interchangeability of reflection datasets between different programs is one of the major problems faced by programmers and users. One way to solve the problem is to follow a rigid format with a specific location reserved for each item of information. This type of data format, though simple to use, is not very convenient when users with different requirements are involved. Therefore, we propose here a more flexible binary data format for reflection data.

Identification of Items by Header Label

The main feature of the data format is a header label for each column of the dataset. With the aid of the header label, a program dynamically relates column number associated with a given header label to an internal variable in that program. Suppose a user program wishes to obtain values for the internal variables IH, IK, IL, FHEAVY and PHASE from a dataset allocated as unit 'IN'. A user assigns a dataset to unit 'IN' in his JCL. The dataset has a header label H K L M FO FC PH. Then, in the GO step of the program, the user will type in IH=H IK=K IL=L FHEAVY=FO PHASE=PH (in any order). When the program is executed, the contents of column headed H is assigned to IH, the contents of column headed K is assigned to IK, and so on.

There is no limit on the number of characters allowed for each header label, and blanks are used as delimiters. The advantage of the procedure is that one need no longer remember what data are stored in each of the columns and any meaningful symbol may be used to identify these columns. A user can dynamically allocate column numbers for the internal variables of the program in the run mode. This method also allows interchangeability of datasets between programs, as the order in which a set of variables is stored in the dataset is no longer relevant.

/cont.

The other feature of the dataset is that it allows practically unlimited length for title information and storage of cell dimension information, etc. Also, the dataset can be sorted by the IBM sort/merge program. This allows fast rearrangement of the records, based on the contents of certain columns. Record length and block size are controlled at run time, and hence a user need not necessarily specify the DCB of the dataset prior to execution. Reading and writing of the dataset is done by routines, written in assembler language. This leads to a substantial gain in efficiency. The routine allows for data of types real, integer (4-bytes or 2-bytes) and literal. However, for reflection records, there should be no need to work with other than 2-byte integers. Cell-dimensions will be stored as 4-byte real variables and titles and headers will be stored as literal data.

The group at Imperial College, London, is developing and implementing the necessary program packages. FORTRAN will be used as the main language for programming. In certain cases, PL1 and ASSEMBLER languages have been used to allow greater flexibility and efficiency. We plan to provide subroutines to handle the dataset at two levels. At the higher level, the routines are usable without detailed knowledge of the data format; at the lower level, the full flexibility of the data format is available to a programmer.