186
4/4/79

MASTER

LBL-7939
UC-4
CONF-7806136

NRCC NATIONAL RESOURCE FOR COMPUTATION IN CHEMISTRY

COMPUTATIONAL METHODOLOGY IN CRYSTALLOGRAPHY: EVALUATION AND EXTENSION

Report of a
Workshop at Asilomar
June 19-22, 1978

LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA, BERKELEY

Prepared for the U.S. Department of Energy under Contract W-7405-ENG-48

COMPUTATIONAL METHODOLOGY IN CRYSTALLOGRAPHY:

EVALUATION AND EXTENSION

Report of a Workshop at Asilomar
June 19-22, 1978

NATIONAL RESOURCE FOR COMPUTATION IN CHEMISTRY

Lawrence Berkeley Laboratory

University of California

Berkeley, California 94720

Edited by: D. Templeton
C. Johnson

Table of Contents

Foreword v

Workshop Participants vii

Introduction 1

Summary and Recommendations 4

Systems 8

Phasing 15

Numerical Refinement 21

Graphics 26

Appendix to Graphics Section 35

Information Retrieval 38

Abstracts 47

FOREWORD

The National Resource for Computation in Chemistry (NRCC) was established as a division of Lawrence Berkeley Laboratory (LBL) in October, 1977. The functions of the NRCC may be broadly categorized as follows: (1) to make information on existing and developing computational methodologies available to all segments of the chemistry community, (2) to make state-of-the-art computational facilities (hardware and software) accessible to the chemistry community, and (3) to foster research and development of new computational methods for application to chemical problems.

Workshops are planned as an integral part of the NRCC's program. As its first major effort in this area, the NRCC sponsored the titled workshop to examine the status and needs of crystallographic computation and to seek recommendations from active scientists in the field on how the NRCC could help meet these needs. The conclusions and recommendations of the participants are summarized briefly on page 4, followed by more extensive reports prepared by panels organized in the workshop. One recommendation from the workshop--the addition of a crystallographer to the scientific staff of the NRCC--has already been implemented and the appointee is expected shortly.

The NRCC is indebted to Dr. Carroll Johnson, Oak Ridge National Laboratory and Professor David Templeton of Lawrence Berkeley Laboratory and the Department of Chemistry, University of California, Berkeley for their considerable time and effort in organizing the workshop and in editing this volume.

The National Resource for Computation in Chemistry is funded jointly by the Department of Energy and the National Science Foundation.

William A. Lester, Jr.
Director, NRCC

WORKSHOP PARTICIPANTS

Carroll K. Johnson, Oak Ridge National Laboratory, Co-Chairman

David H. Templeton, University of California, Co-Chairman

William A. Lester, Jr., Director, NRCC, Lawrence Berkeley Laboratory

William C. Harris, National Science Foundation

Lieselotte K. Templeton, Lawrence Berkeley Laboratory

SYSTEMS PANEL

David J. Duchamp, Upjohn Company, Chairman

Joseph Calabrese, University of Wisconsin

Bertram A. Frenz, Molecular Structure Corp.

Allen C. Larson, 412 Greg Avenue, Santa Fe, NM

Richard E. Marsh, California Institute of Technology

Robert A. Sparks, California Scientific Systems

Charles M. Weeks, Medical Foundation of Buffalo

Allan Zalkin, Lawrence Berkeley Laboratory

PHASING PANEL

William R. Busing, Oak Ridge National Laboratory, Co-Chairman

David Sayre, IBM Research Center, Co-Chairman

Patrick Argos, Purdue University

Gerard Bricogne, MRC Laboratory of Molecular Biology

Richard Gilardi, Naval Research Laboratory

Edward A. Green, Medical Foundation of Buffalo

Eaton E. Lattman, Johns Hopkins University

Mary J. Richardson, Brock University

William E. Thiessen, Oak Ridge National Laboratory

Donald E. Williams, University of Louisville

NUMERICAL REFINEMENT PANEL

Verner Schomaker, University of Washington, Chairman

Steven Freer, University of California, San Diego

Eric Grosse, Stanford University

Wayne A. Hendrickson, Naval Research Laboratory

George A. Jeffrey, University of Pittsburgh

Edward Prince, National Bureau of Standards

Lynn F. Ten Eyck, University of Oregon

Keith D. Watenpaugh, University of Washington

Margaret Wright, Stanford University

GRAPHICS PANEL

C. David Barry, Washington University, Chairman

Robert Diamond, MRC Laboratory of Molecular Biology

Jonathan Hanson, Brookhaven National Laboratory

William Johnson, Lawrence Berkeley Laboratory

Robert Langridge, University of California, San Francisco

Edgar Meyer, Texas A & M University

Donald Vickers, Lawrence Livermore Laboratory

William V. Wright, IBM Corp. and University of North Carolina

INFORMATION RETRIEVAL PANEL

Gerald G. Johnson, Jr., Pennsylvania State University, Chairman

Frank H. Allen, University of Cambridge

Herbert J. Bernstein, Brookhaven National Laboratory

I. David Brown, McMaster University

Lauriston D. Calvert, National Research Council of Canada

Quintin C. Johnson, Lawrence Livermore Laboratory

Alan D. Mighell, National Bureau of Standards

Graheme J. Williams, Brookhaven National Laboratory

INTRODUCTION

Carroll K. Johnson

Advances in the field of crystallography, particularly structural crystallography are closely tied to advancements in computational technology, both hardware and software. Crystallographers were among the early scientific users of computers and they also were involved in the pioneering use of minicomputers for on-line control of experiments, in particular the development of computer controlled diffractometers. Many previous conferences and international summer schools have discussed crystallographic computing but the closest analogue to the present workshop is the National Academy of Sciences conference, "Computational Needs and Resources in Crystallography," held in Albuquerque, New Mexico, April 8, 1972.

The present workshop entitled "Computational Methodology in Crystallography: Evaluation and Extension," is the first workshop of the National Resource for Computation in Chemistry (NRCC) since NRCC was established in October 1977. The workshop held its opening session at Lawrence Berkeley Laboratory, the home of NRCC, on Monday afternoon, June 19, 1978, then reconvened at Asilomar, the California State Park Convention Center near Monterey, for three more days of discussion.

The format of the workshop was derived by an ad hoc program committee at the American Crystallography Association Meeting in Norman, Oklahoma in March 1978. The committee concluded that NRCC can provide a unique opportunity for community level cooperative advancements in crystallographic computational methodology, if the specific areas where such advances are needed can be delineated. A "community level research area" is a field which has developed to the point that a synthesis of the individual results is needed before the field can provide maximum benefits to the general scientific community. Sometimes this synthesis of results is easier to obtain on neutral ground than at one of the established centers involved in the original work. The committee tried to identify a small number of candidate community research areas in crystallography. A panel of about eight experts from each of the five areas selected was invited to the workshop.

1. Systems: Hardware and Software. The first area concerns the nuts and bolts of computing. Computational concepts are useless without the hardware and software systems which implement the computational capability. Many crystallographers have found the in-laboratory minicomputer to be a desirable alternative to the large computer centers; consequently, the program committee invited several of the strongest minicomputer advocates in the crystallographic community to serve on the systems panel. Of course, not everyone has a choice of computational facilities, so the main solution is to insure maximum exploitation of whatever computational facilities are at hand.

2. Phasing by Indirect Methods. The classical phase problem of crystallography has been the subject of many crystallographic papers and conferences. The "direct methods" technique involving relations between the phases of the reflections has proven to be an excellent technique for many small to medium sized structures; however, there are still a significant number of difficult cases where combinations of several techniques must be used, hence the term "indirect methods." This subject is a frontier research area more diffuse than a community research area, as defined above, but it was concluded that at least one topic of this sort should be included to see if the workshop format could produce something new.

3. Numerical Refinement with Constraints. The solution of the phase problem leads to a trial structure which is in approximate agreement with the experimental observations. Numerical optimization techniques are used to improve the trial structure model, which may have several hundred to several thousand variables. In macromolecular refinement and in other special cases, the model must be constrained to maintain reasonable stereochemical properties consistent with previous experimental results. This is a community research area in that many techniques exist for carrying out such calculations. State of the art numerical analysis techniques are necessary in this work, and a rich interdisciplinary area exists for collaboration between numerical analysts and crystallographers. Both computer science and crystallographic experts were included on this panel.

4. Computer Graphics. Interactive crystallographic graphics, particularly for macromolecules, is an active research area at about a dozen locations throughout the world. Computer graphics performance has reached a level such that outside users are becoming seriously interested. Implementing and learning to use an interactive graphics device requires considerable investment in time and capital, and the experience of the established research centers should be pooled to aid the newcomers. This is an opportune community research area and the members of this panel were particularly enthusiastic about the possibilities.

5. Information Retrieval. Crystallographic data bases are being developed at a number of locations. Each data base specializes in some specific area and has evolved independently of the others. In general, the prospective user would like access to several of the data bases but each one requires learning a new set of operating procedures and satisfying different requirements for gaining access. Consequently, these extremely valuable resources are often underutilized. The members of this panel represent most of the known crystallographic data bases, and this is the first time they have assembled to discuss formally the common problems.

The charge to the workshop panels was a series of questions. (1) Where is the current state of the art? (2) What future advances are likely to occur? (3) When are the current techniques and resources inadequate? (4) Why is the field not more advanced than it is?

(i.e., what are the theoretical and practical limiting factors?) (5) Who would benefit from advances in the area? (i.e., what are the interdisciplinary factors?) and (6) How can NRCC help catalyze future advances in the area?

SUMMARY AND RECOMMENDATIONS

David H. Templeton

The workshop produced a good exchange of ideas and spirited discussion of many aspects of crystallographic computing. Agreement was reached that some of them are very worthy of attention by NRCC. Many of the participants afterward expressed satisfaction that they had taken part and that something worthwhile had been accomplished. In this report we record our conclusions concerning things which need to be done and how NRCC might help do them.

The workshop tended to concentrate on joint activities where NRCC might facilitate cooperation among groups of persons to advance computational science in ways not so likely to occur otherwise.

This report gives lesser attention to the category of individual proposals for work to be assisted by NRCC. Such a proposal must be judged on its merits in each particular case. The more original and innovative it is, the less likely it is to be anticipated by a workshop like this. The report contains suggestions of many areas for such proposals, but these are not exclusive. The NRCC program should be open-ended for proposals from creative individuals.

In the organization of the workshop, it was necessary to compromise between depth of representation of work in particular areas and broad coverage of diverse crystallographic research. As a result some important subjects receive little explicit attention in this report, and some points of view are likely to be unrepresented. As examples, the accurate measurement of electron density in simple structures and advanced techniques of electron microscopy could each have been the subject of a panel, but they are touched on only incidentally. This neglect does not imply a value judgement, but rather indicates the practical necessity of limiting the scope of the workshop.

RECOMMENDATIONS OF THE PANELS

Each panel was instructed to study what is needed to advance computation in the particular field, first without the limitation of how it might be achieved, and then to recommend what NRCC could do to help meet this need. The recommendations are set forth in the respective reports, which are the definitive versions. As a convenience for the reader, but with some risk of oversimplification and loss of context, an abridgement of the main points is given here.

Systems Panel Recommendations

1. The NRCC staff should include a person knowledgeable in crystallographic computing techniques.

2. NRCC should fund development of algorithms which will improve crystallographic calculations.

3. NRCC should aggressively acquire, maintain, test, and document a complete system of crystallographic computer programs.

4. NRCC should be a clearing house of crystallographic software.

5. NRCC should promote and support tutorials and workshops concerning programming practices, standards, and algorithms.

The first point, having a crystallographer on the staff, is the highest priority and is a prerequisite of all the others.

Phasing Panel Recommendations

1. NRCC can help strengthen crystallographic phasing techniques by support of tutorials, schools, and workshops.

2. NRCC should encourage and support proposals for research in phasing methods.

3. In selected cases, NRCC should provide computing time for solution of crystal structures.

4. NRCC should acquire, test and distribute phasing programs.

The highest priority is for peer education and communication (point 1).

Refinement Panel Recommendations

1. NRCC should bring about definitive checks of quantum chemical and crystallographic results for molecular dimensions.

2. NRCC ought to be firmly committed to refinement of macromolecular structures, including development of a high-performance graphics system.

3. NRCC should organize collaborative research sessions, after the CECAM model.*

A commitment to macromolecular refinement (point 2) is the central recommendation. Point 3 is one example of a means to this end, but is listed separately because it applies equally to many other objectives in crystallography and indeed other fields of chemistry. Point 1 is a scientific problem which must be addressed by individuals, but NRCC might have an important effect as a place where quantum chemists and crystallographers are brought together.

---

*A small group of scientists is invited to work for a few weeks at the Centre Européen de Calcul Atomique et Moleculaire (University of Paris) to carry out some major calculation.

Graphics Panel Recommendations

NRCC should undertake:

1. To develop a family of compatible, interactive graphics systems.

2. To disseminate medium performance graphics systems.

3. To develop a device-independent, general-purpose graphics package.

4. To establish access to a high quality color, computer driven microfilm facility.

Priorities are hard to set. The development of an advanced multi-level graphics system (point 1) would be very expensive, but its value would be large and lasting. Point 2 would have an immediate impact and would cater to a community not presently served. Point 3 is something the panel would like to have today. Point 4 is out of reach of most chemists except through some special channel such as NRCC.

Information Panel Recommendations

NRCC should support:

1. Education of potential users of data bases with short courses, tutorials, workshops, and other means.

2. Research using data in databases.

3. Research on data retrieval techniques.

Education is the highest priority. Much effort is going into compilation of databases, but few chemists are aware of them or know how to access them.

OVERALL CONCLUSIONS

The reports which follow and the recommendations in them are the work of the individual panels. In the final session of the workshop these recommendations were presented to and discussed by the full group. Recognizing that finite resources require some choices to be made, we searched for a consensus on relative priorities. There was broad and strong support for a major effort in graphics by NRCC as the highest priority item. As to the details of this major effort, opinions differed. One group urged the importance of development of the next generation of graphic capability; others gave priority to providing broader access to present technology. On the choice of a single highest priority item from among the specific recommendations of the various panels, no consensus was possible. A tentative straw vote, taken amid objections that a rational choice was

impossible and that such a vote was too simplistic, gave results scattered among the five alternatives almost within random statistical expectations.

Throughout the discussions, no significant objection was voiced against any of the recommended items except that NRCC could not do all of them. The ideas were expressed repeatedly that some action should be taken on several of them, and that exact priorities must reflect the practical circumstances of NRCC and its actions on other fronts. The major graphics effort, for example, probably is most realistic as a service to the whole chemical community and not as a narrow project for crystallography.

Another theme which recurred was the value of various types of workshops, tutorials, and working groups for a variety of purposes. The highly successful CECAM session last year on the structure of a virus is a model which NRCC could emulate with profit for key problems in any field of chemistry.

Another conclusion one may draw is that crystallographers, who indeed need regularly and frequently to carry out large scale computations, do not view NRCC as a major source of this computing service which already is provided to most crystallographers by other means in a more or less satisfactory manner. Rather, the potential value of NRCC is more in providing access to exceptional facilities (e.g., graphics) or in promoting human cooperation, interaction, and education.

SYSTEMS

INTRODUCTION

A crystallographic computing system is a set of programs which are mutually compatible and which are capable of passing data and parameters between one another essentially automatically. A system is the means by which we automate our crystallographic computing on our particular hardware configuration.

The reasons for automating calculations are essentially three: 1) reduce the human effort required to carry out a series of calculations, 2) obtain a quicker answer to our problem, and 3) get better results at lower cost through elimination of error. A successful system may permit the user to carry out the complicated and sometimes tedious calculations of a crystal structure analysis without being familiar with the details of the calculations or with the algorithms used to implement them. An obvious result of such systems has been the opening of the field of small-molecule crystal-structure analysis to the non-crystallographer.

Some of the goals which systems developers attempt to achieve are:

a. Single entry of data -- For example, once the unit cell dimensions are entered into the calculations, they should be retained for all subsequent calculations on that structure.

b. Minimum input -- This can be achieved by having defaults for all options and by never requiring the user to enter anything that the computer can calculate.

c. Minimization of input errors -- Good input design is critical here. All program input should be "user friendly".

d. Continual operation until a "human" decision is needed.

e. Minimum effort for decision. Decision making must be facilitated by well designed organization and presentation of relevant data.

f. Minimum restart effort once a human decision is made.*

There are many problems in achieving all these goals, not the least of which are those imposed by hardware configuration and monitor.

Systems may be classified into two types: the loosely-coupled system, in which programs within the system run essentially independently as separate jobs or job steps, but have mutually compatible input/output,

---

*A reviewer suggested that good documentation should be added to this list.--Ed.

and tightly-integrated systems in which the system looks to the operating system of the computer as one super program, so that one execution can run several programs in sequence.

Programs are added to crystallographic systems in two ways: 1) a program can be written specifically for inclusion in a particular system, and 2) a program written independently can be adapted for a particular system. Sometimes programs are difficult to add to systems. Certain features may make a major reworking of the program necessary. The integration of a program into a system has in some cases been so large a task that the system developer has elected instead to write a "splice program" which changes one program's I/O format into an I/O format acceptable to another program.

PRESENT STATE OF COMPUTER SYSTEMS

Crystallographic programs exist for computers of virtually every manufacture. However, there is a wide variance in the sophistication of these programs and, in many cases, these programs do not constitute crystallographic program systems as defined above. Many state-of-the-art program packages accomplish most, if not all, of the goals outlined above. In addition to systems developed for large computers, there has been introduced, in the past few years, a number of program packages for small computers. These are designed for exclusive use by one research group, or in some cases, for an entire department. Several of these packages are available commercially, and the hardware and software can be purchased for as little as $30K.

TRENDS IN THE COMPUTER INDUSTRY

Historically, the development of crystallographic computer systems has paralleled the development of the computer industry, and it is useful to take a brief look at where the computer industry is going. There is a lot of activity in this area. Computer memories are becoming much cheaper. Bulk storage is becoming increasingly available with the introduction of bigger, faster, and less expensive disk drives, and with the emergence of new technology devices, such as bubble memory. Distributed processing, sharing of tasks between computers, has already been implemented in chemical computing and shows promise of being very useful in crystallography. The increasing availability of array processors at reasonable cost is especially worth noting since these devices can enormously increase the computational speed of small, as well as of large, computers.

Microcomputers constitute another dynamic area in computer development. When the 8-bit micro "computer-on-a-chip" came out a few years ago, most crystallographers correctly concluded that they were too limited and too slow to do much in crystallographic calculations. However, very recently several manufacturers have introduced a new generation of microcomputer chips -- 16-bit devices, with large addressing capability and speeds reported to be as fast as the PDP 11/45.

If history repeats itself, within five years another generation will be introduced which will have 5 to 10 times the capabilities of these new impressive devices.

In the area of user input/output devices, the punched card, which many people still use, is being replaced with interactive job submission. The computer graphics area is undergoing rapid development. High capability, low-cost graphics terminals are becoming available as the raster scan technology progresses. These terminals will probably be available to all crystallographers in the near future, instead of being available only to a few laboratories.

One of the driving forces in the computer industry today is the development of suitable computers for the "home computer market." This part of the industry seeks to put a computer in millions of homes. At the recent National Computer Conference (June 1978), a panel concluded that, to achieve this, such a machine must have the following features: color graphics, voice input, voice output, the ability to control things around the home, and a price under $1000. The large development effort in this area will generate hardware which could be useful to crystallographers.

To summarize, the computing power of the small computer (mini and micro) will continue to grow, making "grind" type calculations increasingly less expensive. In many situations, the minicomputer already has become the cost effective way to do crystallographic calculations. (Minicomputers and Large Scale Computation, P. Lykos (Ed.), ACS Symposium Series No. 57, Washington, D.C., 1977). The availability of better and less expensive graphics terminals and mass storage devices will benefit the crystallographic community. New devices, the array processor in particular, show promise for helping with large computing jobs.

TRANSPORTABILITY OF PROGRAMS

To facilitate a flexible and useful exchange of crystallographic programs by providing painless, efficient, and low cost implementation of such programs, it is essential that certain reasonable guidelines and standards be established. FORTRAN is the most universal computer language in use today and its continued use is fundamental. ANSI FORTRAN is an accepted standard and should be adopted as the standard for the programs in the exchange list. Machine independence should be maintained at as high a level as possible by use of good programming methods. Workshops on good programming techniques would be very useful to all chemists in developing reasonable and acceptable programming standards. With the establishment of guidelines for software development, the ability to integrate programs to a variety of computers, both mini and maxi types would considerably improve the chemist's options in dealing with computer problems.

CLEARINGHOUSE FOR CRYSTALLOGRAPHIC PROGRAMS

This panel believes that there is a clear need for a central clearinghouse for computer programs dealing with crystallographic calculations.

Background: Crystallographic computing has developed in a random fashion. In the early days of digital computing, programs were written in machine language and, by necessity, incorporated algorithms that were tuned to the particular computer. Since the programmer was usually the user as well, documentation of these programs was sketchy, and they were not widely disseminated. High-speed, large-memory processors have removed much of the need for machine language programming and individualized algorithms. This has not led to standardization of programming styles or, with some notable exceptions, to wide dissemination. And, in those cases where documentation has been extensive and dissemination wide, such as ORTEP, MULTAN and the XRAY system, the burden of distribution has become increasingly difficult for the originating laboratory to bear. Even more important, there exists an imposingly large collection of programs and algorithms, tested and at least partially documented, that are effectively unavailable to others in the community because of the difficulties and the inertias involved in their distribution.

Previous attempts to establish a central clearinghouse for crystallographic programs, most notably a continuing project by the American Crystallographic Association, have met with less than overwhelming success. The lack of success of these projects has been due, we are quite sure, to the fact that they have been passive projects: it has been left to the original programmers to submit their programs as they see fit, with or without adequate documentation, and the recipient has had no guarantee that the program is workable and no help in its implementation. Not surprisingly, the few successful attempts at wide dissemination have resulted only when the original programmer has been willing to invest large amounts of time and effort.

We believe that an active clearinghouse of computer programs would be of immense help to the crystallographic community. We see the following features as necessary to the successful operation of such a clearinghouse:

1. A supervisor, familiar with crystallography and with the crystallographic community, to act as liaison between program contributor and user.

2. Continuing, active solicitation of programs and distribution of lists of available programs.

3. A facility for checking and evaluating programs. Although we realize that it is not feasible to routinely check all submitted programs, such a facility would greatly ease the communication between programmer and user.

4. Aid and advice in the preparation of documentation.

5. A facility for transcribing accepted programs, either for distribution via tape, cards, or disc, or for use by remote access.

NEW ALGORITHMS AND TECHNIQUES

There is a need for development of new algorithms for use by crystallographers. It is important that these algorithms be adequately documented and that provisions be made for their dissemination to all those crystallographers who could use them.

The following are examples of algorithms that would be valuable to a large number of crystallographers and chemists.

1. The number-crunching power of many maxi, midi and minicomputers, has been greatly enhanced with the addition of an array processor. For example, R. E. Kaplan of the University of Southern California added a $35,000 Floating Point Systems AP 120B Array Processor to his Data General Eclipse minicomputer. He then compared the execution time of his fluid mechanics programs on this system with the same programs run on the Illiac IV and CDC Star computers. He found that the minicomputer with the array processor executed these programs as fast as the Illiac IV and half as fast as the CDC Star computer.

   Array processors have not been used by crystallographers and it is not clear how useful they will be for crystallographic applications. It may be that the short word length (38 bits or less) will be a serious disadvantage. To evaluate the array processors, it would be desirable to rewrite certain large, number-crunching crystallographic algorithms so that they could take advantage of the power of these processors. These algorithms should be written so that they could be transferred easily to other host computers and, if possible, to other array processors. Such an evaluation will give information about the cost effectiveness of the array processor and the importance of the length of the floating point word for crystallographic applications.

2. It would be desirable to write new or improved algorithms for calculation of quantities from crystallographic data that can be related to other experimental chemical data or that can be related to certain theoretical calculations. For example, algorithms could be developed for a more precise determination of electron densities and for more precise and meaningful molecular thermal parameters.

3. There is a need to develop new algorithms and techniques for use with distributive computer systems. It is especially important to disseminate these algorithms and techniques to potential users.

These are but a few examples and as new technology develops, there will be a need for many additional algorithms.

RECOMMENDATIONS

The Systems Panel makes the following recommendations for NRCC support of crystallographic computing.

1. The NRCC staff should include a person knowledgeable in crystallographic computing techniques. It is essential that at least one NRCC staff member be familiar both with crystallography and with the local computing facility in order that there be an easy interface between NRCC and the crystallographer.

2. NRCC should fund the development of algorithms which will improve crystallographic calculations. In addition to supporting entirely new computational procedures, NRCC should also support proposals to investigate new methods of performing old computations, as well as proposals to find methods appropriate for new hardware as it becomes available. One example of the latter is the development of efficient algorithms for machines equipped with array processors.

3. NRCC should aggressively acquire and maintain a complete system of crystallographic computer programs for both large and small molecule structure determinations. The accuracy of these programs should be verified by running a standard set of test problems, such as that suggested by the IUCr. The documentation should meet certain minimum standards to be determined by the NRCC staff. These programs would be available for client or customer use as needed.

4. NRCC should act as a clearinghouse of crystallographic software for both maxi and minicomputers. The storage and dissemination of this software could be delegated to an existing depository. The NRCC should aggressively induce crystallographers to submit programs, and if necessary, should test these programs and improve their documentation before they are passed on. In addition, NRCC could function as a clearinghouse for the distribution of algorithms for crystallographic computations.

5. NRCC should actively promote and support a series of tutorials and workshops. The topics of these meetings should include (a) good programming practices, (b) programming standards for crystallographic programs, and (c) crystallographic computing algorithms. These workshops and tutorials might be appropriately held in conjunction with ACA meetings.

PHASING

TECHNICAL BACKGROUND

Methods

The methods available for phasing and structure solution today are indicated in the figure. The four ovals running down the middle of the figure (magnitudes, phases, structure image, and parameters of the structure model) represent the quantities involved in structure determination. Any sequence of paths which leads to the second oval is a phasing method. The processes ranged round the periphery of the figure are:

The diffraction experiment (DIFF);

Fourier and inverse Fourier transformation (F and F^{-1});

Direct methods (DIR);

Phase extension and refinement (PREF);

Model building (MOD). This may be based on magnitudes only or on ρ, and frequently involves a wide range of chemical and other information;

Refinement and Diamond-type refinement (REF and DREF);

Microscopy, usually electron (MIC);

Structure-factor calculation (SF);

Calculation of structure image from structure model (unnamed);

Physical phasing methods (PHYS). The principal examples today are the heavy-atom phasing methods, but other coherent reference-beam methods, such as Post's multiple-reflection technique and x-ray holography, are future possibilities.

Solid lines in the figure indicate that the process involved is certain to succeed, or nearly so; dashed lines that success of the process is uncertain.

Two of the processes (REF and DREF) are the subject of another panel, and will not be dealt with further here.

Assessment of the Methods

Uncertainty. It will be noticed that there is no certain phasing method except for microscopy followed by inverse Fourier transformation. The resolution of microscopy today is such that this path is available for low-to-medium resolution phasing only. Thus phasing today, while usually

successful, is not certain, and this leads to one of the main driving forces on the field, namely to reduce this uncertainty. The situations which are most serious in inducing uncertainty today are: non-rigid molecules with more than 100 non-hydrogen atoms and no heavy atoms, e.g., peptides with 10 or more amino acid residues; disorder, particularly if it involves heavy-atom locations; pseudosymmetry, especially the types involving pseudotranslation along an axis or molecular locations such that the apparent space group is of higher symmetry than the true one; twinning; and absence for any cause of high-resolution magnitudes.

Cost. There are computation costs associated with the diffraction experiment (diffractometer control, data reduction), Fourier and inverse Fourier transformation, direct methods, phase extension and refinement, model building (computer graphics, energy minimization, use of known structural information), refinement and Diamond-type refinement, structure-factor calculation, model to image conversion, and some cases of physical phasing. These costs can be large and represent another major driving force on the subject. Non-computational costs can also be highly important, especially in the macromolecular case (problems of crystal growth to permit the diffraction experiment, problems of producing isomorphous crystals for physical phasing).

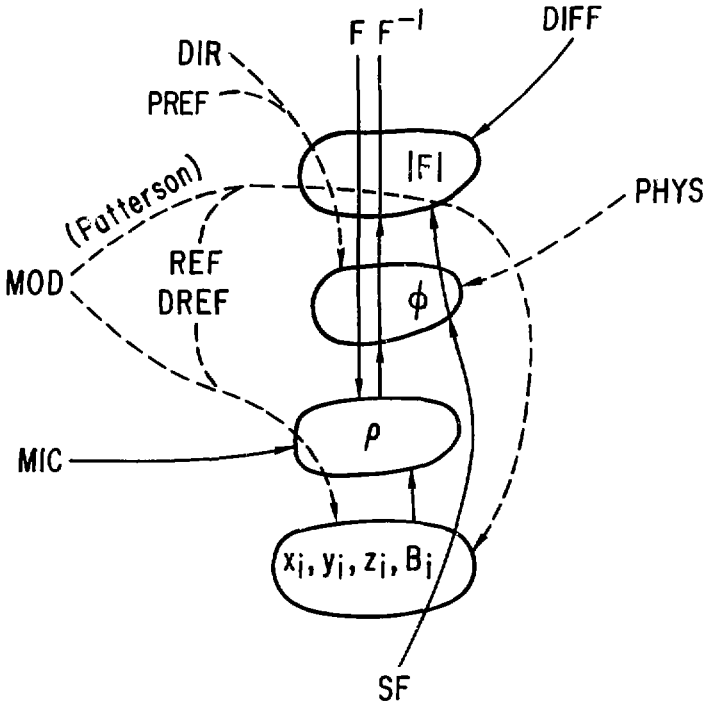
Pressure from Structure Size. There exists a further strain on the subject, arising from the existence of highly important biological structures lying at or beyond the current limits of crystallographic technique. This reservoir of larger structural problems will provide a third driving force for development of improved phasing methods throughout the foreseeable future.

Current Technical Responses

Fortunately, there exists today a highly vigorous level of technical innovation in phasing methods, which is succeeding (though sometimes only marginally) in keeping the techniques at an adequate level for the needs of crystallographers. The following list mentions some of those of apparently greatest promise.

PHYS	Use of synchrotron radiation in anomalous dispersion. Multiple-reflection techniques.
DIFF	Small-sample crystal growth techniques. Dynamical theory needed for microcrystalline diffraction using electrons.
F, F^{-1}	Winograd superfast algorithms.
DIR	Higher-order structure invariants and seminvariants. New complete direct-method systems (e.g., MAGLIN).

PHASING OF DIFFRACTION DATA (SCHEMATIC)



PREF Use of non-crystallographic symmetry in large proteins and viruses.
Least-squares and matrix methods in smaller proteins.

MOD Computer graphics (see report of Panel #4).
Molecular replacement, i.e., determination of unknown structures from similar known structures.
Use of energy minimization, packing calculations, etc.

MIC Near-atomic resolution of inorganic materials.
Electron microscopy crystallography, i.e., use of electron microscopy with electron diffraction as protein structure-solving technique.

RECOMMENDATIONS

Education and Communication

One way in which the NRCC can contribute to the solution of the phase problem is by the support and sponsorships of tutorials, schools, and workshops at various levels of knowledge. In some cases the methods are well understood by the experts and programs are available, but researchers need specific instructions and hands-on experience in their use. Schools should provide such instruction in the areas of direct methods, multiple isomorphous replacement, molecular replacement, and molecular averaging methods. Participants would bring unsolved problems to these schools and would learn to apply the available techniques. Tutorials might also be sponsored to educate chemists on mathematical subjects such as fast Fourier transforms or special functions used in diffraction theory.

Advanced workshops would concentrate on bringing together researchers with a common interest in solving some frontier problem. For example, the sessions organized by the Centre Europeen de Calcul Atomique et Moleculaire (CECAM) at Orsay, France, have proven very successful. These involve about a dozen people for a period of two months and provide them with ample computational facilities. In the area of crystallography these workshops aim at initiating collaboration between experimentalists who are struggling with difficult structures and theoreticians engaged in developing methods to solve them. A recent Workshop on Virus Crystallography (July - August, 1977) has had a marked impact on the field, making possible the large computations involved in the structure determination of Tomato Bushy Stunt Virus.

Support of Research in Phasing Methods

The NRCC should encourage proposals from individuals or groups for the development and testing of new methods for solving the phase problem. Support for this research could be in the form of computer time needed for the work. It could also provide for an extended or sabbatical visit by the researchers to the NRCC where they would have access not only to

computing facilities but also to consultation with mathematicians, programming experts, and other scientists.

Problems which such research programs might attack include:

Development of methods and programs for combining information from physical experiments with statistical methods,

Solution of structures with pseudo-symmetry or disorder,

Solution of problems involving 100 or 200 atoms in the asymmetric unit,

Determination of structures when the amount of diffraction data is limited,

Solution of fiber and polymer structures,

Optimizing large Fourier transformations by improving software and hardware,

Solving super-large structures such as viruses, membranes, and chromatin,

Determining structures which exhibit polysynthetic twinning,

Developing software packages for electron microscopic crystallography or for utilizing diffraction data from microcrystals.

Providing Computing Facilities

In certain cases the NRCC should accept applications for computer time for the solution of a particular crystal structure. These should be cases which would provide a genuine advance in chemistry but which could not be undertaken by the researcher without support. They might require computing facilities not available to the applicant or programs which are not readily transportable.

Brookhaven National Laboratory has had extensive experience in serving crystallographic users with Control Data computers. Their experience should be examined to identify problems involved in remote computing.*

*One might include consultation with remote users of computers, including those of the other National Laboratories.--Ed.

Maintaining a Program Library

The NRCC should acquire a library of programs useful for solving the phase problem. These programs should be checked out and documentation should be available. Many of these programs would be acquired as a result of schools and workshops held at NRCC. Others should be solicited from their authors.

Distribution of these programs could be undertaken at some appropriate level. Participants in schools should be able to take home with them the program packages which they have learned to use.

Comparison and testing of phasing programs might be done by resident staff members or as a research project in phasing methods. As an aid to such testing, the NRCC could also establish a library of data for structures which were difficult to solve.

NUMERICAL REFINEMENT

The computer and diffractometer, in the course of 25 or 30 years, have brought the art of refining small crystal structures to a remarkably satisfying but still perceptibly immature state. For large structures such as proteins, nucleic acids, and viruses these years have seen stupendous achievements of discovery and verification but only the beginnings of much needed and eagerly awaited structure refinement. The NRCC can make important contributions to the art of refining both classes of structures.

For small structures, standard full-matrix least squares on such computers as are available to almost all workers gives good results quickly at rather moderate cost. This isn't to say that improvements in speed, cost, and convenience of both computers and program systems aren't to be expected -- only that the situation in these respects isn't felt to be critical. The only commonplace serious difficulty is that the usual type of model (spherical form factors, individual anisotropic U's, some attention to absorption and extinction) usually gives statistically unsatisfactory agreement. Goodness of Fit equal to 3 (if honestly estimated) is common for reasonably large crystals, small unit cells, and long counting times for reasons that vary from case to case but in the computation domain always include contributions (whether dominant or not) from curvilinear or otherwise anharmonic motions and from deviations from the spherical-form-factor approximation. Sometimes there is difficulty with convergence and always there is uncertainty whether appreciable structural errors are induced. About the gross unreliability of x-ray C-H, O-H, and N-H distances there is, of course, no doubt, but the associated errors in C-O, C-C, and C-N are not known. There are indications that a simple modification of the least-squares procedure to make it more robust and resistant will often be helpful. One such, (see the abstract by E. Prince) has indeed produced good results (convergence and apparently correct parameter values) in some cases where the usual method had failed, but it is hardly clear how effectively it will reduce the above-mentioned induced parameter errors. Nor is it clear whether refinement with constraints, for example, of non-crystallographic symmetry or of coupling the motions of rigid or quasi-rigid groups, will prove to be consistently effective.

What is necessary for the maturity of this part of structural crystallography is to bring to full test the reality and accuracy of its numerical results -- the bond lengths, angles, and torsions; the Uij's; the overlap densities; etc. -- against results from quantum mechanics, crystal spectroscopy, and where possible, experimental studies of the gas molecule. The confrontation is inevitable. It will almost certainly be as revealing on all sides as was the similar serious meeting a few years back of gas-molecule electron diffraction and microwave spectroscopy.

Our greatest present need and opportunity is probably to afford the best possible comparison with the results of current quantum chemistry. Current state-of-the-art x-ray and neutron single-crystal structure analysis ($R \approx 2$ percent) can provide the most complete experimental

descriptions of the structure of small* molecules that are possible. Electron densities can be mapped to standard deviations of 0.01-0.05/Å^3 and molecular dimensions can be measured to σ's of 0.001 Å in distances and 0.1° in angles. These electron densities and dimensions are averages that refer to a particular molecule undergoing thermal motions within the potential fields of the other molecules in the crystal. The same chemical molecule in another crystal structure (i.e., a polymorph) will have a different electron density and different dimensions because it is in a different crystal environment undergoing different thermal motion.

The same molecules observed in the liquid or gas phases or calculated theoretically (i.e., isolated molecules at rest) will also have different electron densities and dimensions.

We need to be able to make the connection between crystallography, nmr, electron diffraction, and optical molecular spectroscopy by correlating the observable numerical quantities to a common state for the same molecule; either the isolated molecule at rest or the observable average molecule in the crystal.

To be able to do this we need to develop a well-tested and well-documented series of programs that, with appropriate computer hardware, will perform the following:

1. Carry out a theoretical thermal motion analysis and make the thermal motion adjustments to electron densities and molecular dimensions, without the necessity to assume rigid-body motion or harmonic motion.

2. Determine the distortions in a molecule that are due to intermolecular interactions with the other molecules in the crystal.

Such programs should be capable of application in both directions:

- To reduce an observed structure to that of an isolated molecule at rest.

- To extrapolate from an isolated molecule at rest to the molecule in the known crystal structure.

The latter would be of primary interest to the theoretical chemist, who was seeking experimental support for his quantum mechanical approximations to the wave equation. The former would be of primary interest to the crystallographer who is seeking a common base for correlating his results with those from other types of experiments.

*A small molecule will fit into a small crystal structure, which in the sense used in this report can be defined pragmatically as one that can be refined by full-matrix least squares using present day computers.

If both disciplines approach this problem from these opposite directions, major progress may be made in the understanding of the electronic structure of the molecules.

For large structures the picture is quite different. Even the biggest computers fall far short of accommodating refinements by full-matrix least squares, despite the usual lack of enough observed reflections for a decent level of overdetermination for even a crude model without solvent molecules and with isotropic temperature factors. Although a single large-structure refinement (of rubredoxin) has been pushed into the range characteristic of small structures refined with the same type of model, it stands alone. With but few exceptions, all others have hung up at levels ($R \approx 20\%$) that for small structures would signal probable gross error. Rubredoxin gives great hope, but in fact we don't yet know whether the large structures typically are capable of being refined to 10% or so in R and to realistic σ's of less than 0.1Å in atomic positions.

This is not to imply that refinement of macromolecular structures is not a very active field of research. It is. Or that progress has not been made. It has. But the need for model accuracy at the very highest possible level is extremely great. Many important biochemical questions cannot be answered without such accuracy, and the value of the potential results can hardly be overestimated. Protein structures of very high accuracy are needed for work in the fields of enzyme catalysis, drug design, molecular interactions, and molecular dynamics, to name only a few. Unfortunately, we don't quite know how to proceed. We have to deal with a very large number of parameters, relatively low resolution,* and noisy data.

All of these factors affect refinement in ways that aren't well characterized. In any case, the size of the structures precludes the use of full-matrix least squares, for which the storage required increases as the square of the number of parameters and the processor time faster still. And even if this were not bad enough, the basic difficulty of non-linearity may well prove to be so much more important

---

*The resolution limit is due to expense of measuring enough reflections for higher resolution, instability of the crystals in the X-ray beam, or sheer absence of measurable high-angle reflections. Biological macromolecules are intrinsically rather flexible and the lattice contacts between such molecules in crystals are much less restrictive than are those in the tightly knit crystals of small molecules. This often results in large thermal vibrations, dynamic disordering among possible local conformations and static variations between the structures within different unit cells. In turn, the diffraction intensities diminish rapidly with scattering angle. Consequently, the atomic structure may be poorly specified by the diffraction data. Indeed, the number of atomic parameters may actually exceed the number of independent observations.

than it is for simple structures that it defeats any automatic procedure based on the small-structure methods.

The present procedure for refining macromolecular structures is to adjust automatically by some numerical procedure a molecular model to improve the fit between observed and calculated diffraction patterns while paying due attention to known stereochemistry. In all known cases and by all known methods this adjustment step converges again and again to a local false minimum. At each such halt, visual inspection of difference maps and manual adjustment of the model are then required. When the automatic adjustment is then re-started, it converges to another, usually better, but still only local minimum, and so on. The inspection-and-adjustment procedure is the limiting step even for small proteins; most protein refinements spend 99% of their time in this phase.

The difficulty of this limiting step is seriously hampering research. It takes so long that alternative methods of refinement and alternative models are not being adequately compared and evaluated. Important questions about the radii of convergence and rates of convergence of different algorithms are not being answered--to say nothing of the delay in producing refined coordinate sets.

There is a general consensus among workers in macromolecular crystallography that a high-capacity interactive computer-graphics facility would be a great aid in the manual adjustments. It would provide a direct method for making adjustments to a model and in addition would be an invaluable aid in developing automatic or semi-automatic adjustment procedures.

Although, in the way just described, the laborious interactive aspect of large-structure refinement seems to us the more critical now, the computational limitations are also of course, extreme. Ways to improve or even revolutionize the numerical searches doubtless exist, but no one, it seems, knows how to find them for sure. Here is a compact summary of the situation from this point of view, not without some hope, as seen by our non-crystallographic members: In refining large molecules it is not possible to store the matrix of observation and restraint data, or even the corresponding normal matrix. Only matrix-vector products are available, often in implicit form (e.g., FFT evaluation of the gradient), so that some variant of conjugate gradients, either applied directly to the non-linear refinement or to a linearized subproblem, appears to be most practical. Only a few steps are typically used before the refinement slows down; local rebuilding of the model is then carried out, and the cycle is repeated. Although conjugate gradient techniqes could be speeded up, in various ways, this alone would not provide a major advance in refinement. In this context, the promising areas for numerical work are a deeper analysis of the non-linear least squares problem and aids to map interpretation. Improved diagnostic information, such as standard error estimates for parameters, would also be helpful. Good numerical techniques are available for linear constraints, where the associated matrix is either small and dense or large and sparse. Effective

techniques are also well developed for least-square problems when some parameters enter linearly and some non-linearly, as in fitting occupancy factors. Other numerical methods under active development, for example, in non-linearly constrained optimization, should be useful in crystallography, since non-linear structural relationships could be treated directly rather than linearized.

The NRCC could provide valuable support for research in macromolecular structure refinement in the following ways:

1. By developing an interactive computer graphics facility for macromolecular crystallography. Such systems will be too expensive for most laboratories for some years to come. Several laboratories have or are getting graphics systems; the NRCC facility could act as a standardization center and clearing house for programs in this field.

2. By making an adequate commitment of facilities and personnel to continuing work in macromolecular refinement and in supporting such refinements. This could involve development of a loosely coupled system of programs and data formats in which different types of refinement algorithms could be tested--perhaps by workers interested in optimization, but without protein data of their own to work on.

3. By maintaining a program exchange in this area.

RECOMMENDATIONS

1. We feel that the NRCC should do what is necessary to bring about definitive checks of quantum chemical and crystallographic results for molecular dimensions (bond lengths, bond angles, etc.) This will require considerable computational and theoretical developments for which, in view of its prospective users and facilities, the NRCC should be uniquely well suited.

2. We feel that the NRCC ought to be firmly committed to the crystallographic refinement of macromolecules. An essential facet of this commitment would be the development of a high-performance interactive graphic system for molecular modeling.

3. The NRCC should organize extended collaborative research sessions, after the CECAM model, to program and execute computational work on crystallographic problems. The working scientists would bring unsolved or partially solved problems relevant to the theme of the session with the aim of solving the problems and testing, improving, extending, or devising the methods to be used. There should be good access to a major computer for both the planned work and immediate trial of new ideas.

GRAPHICS

The panel recommends that the NRCC should become actively involved in the development and application of interactive computer graphics to molecular structure studies. Molecular graphics has already been demonstrated to be an important tool in the area of X-Ray crystallography. We recommend that the NRCC should take steps to improve the availability of graphics systems and graphics software for this community. It should also respond to the needs of the broader community in the physical sciences by making graphics tools more readily available for use in the presentation and examination of the data.

Specifically we advocate that the NRCC undertake:

1. The development of a family of compatible, interactive computer graphics systems to be based on a range of hardware configurations. The spectrum of configurations should include a very high performance, dedicated system suitable for operation by NRCC as a national facility, as well as low cost terminals loosely coupled to a time shared host or more tightly coupled to a laboratory mini computer which may be dedicated to graphic applications for periods of time.

2. The dissemination of medium performance graphics systems for the chemical community.

3. The development and export of a device independent, general purpose scientific graphics package for use by the NRCC community.

4. Establish access to a high quality, color, computer driven microfilm facility.

The panel has examined each of these broad recommendations in more detail considering especially the impact they may be expected to have on the general community served by NRCC, their cost and their time to fruition. Feasibility was also discussed, the results of these discussions being presented in terms of a plan for implementation of the recommendation.

A summary of our deliberations on each of these recommendations follows. An appendix to the graphics panel report is at the end of this section. It provides an outline of how the various facilities available through the NRCC could be used in support of a significant application of molecular structure research, namely DRUG DESIGN.

DEVELOPMENT OF A FAMILY OF COMPATIBLE INTERACTIVE COMPUTERS GRAPHICS SYSTEM

World-wide, 12 or so groups are currently involved in the development of high performance molecular graphics systems. Upward of 100

man-years of effort have gone into these developments and this effort should be preserved through future hardware generations. In a few instances communities have evolved by either adopting common hardware and software or by sharing a common facility. However, in general, there has been very little transfer of expertise from one system to another than that produced by the continuity of personnel. The reasons for this perhaps are diverse but a major contributing factor has been the machine dependence of programs resulting from the optimization necessary to produce adequate performance levels for the chosen spectrum of applications.

With the establishment of the NRCC it is appropriate to consider steps that can be taken to break away from machine dependent programming and to develop a compatible family of systems for a range of hardware. Some performance may have to be sacrificed to achieve this goal but with current performance characteristics for state-of-the-art hardware, the trade off of the next potential gain in performance for a degree of machine independence and a compatible family of systems seems an attractive option for NRCC to exploit on behalf of the national community.

The design of software for a computer system involves 3 phases, conception, representation and implementation. The panel has recommendations covering three levels of hardware (recommendations 1B, C, D.). Development of systems for these levels can be made compatible by the simple expedient of using a common conceptual and representational design, divergence only being necessary at the lowest (implementation) level. Systems developed in this manner will be compatible at the level of man-machine protocols--these being defined in the conceptual design phase and at the level of data structures--these being defined during the representational design.

Recommendation 1A: A conceptual and representational design study for molecular graphics systems. An appropriate design goal would be one aimed at achieving a synthesis of existing systems. This should facilitate incorporation of new ideas emerging from research done on any or more of the existing systems. The conceptual design phase would be carried out by a single person versed in the use of computers for molecular modeling and display. We anticipate that this stage would be approximately one man-year (mythical?) and anticipate that it would involve:

Evaluation of existing systems and identification of underlying concepts;

Surveying of user needs in all appropriate application areas;

Selection of man-machine protocols.

It would lead to a functional description of the system and publication of a user manual. This manual would stand as an authoritative document--a blue print for a well designed molecular

graphics system. It should be accepted as establishing a standard for these systems and in this way have a major impact on the field.

The representational design phase will take place in parallel with the later stages of the conceptual design. It involves:

Establishing data structures including those involved in links to other programs;

Identifying operational primitives;

Selecting graphical representation of data;

Surveying languages available and standards proposed for graphics. (See also systems panel report and the ACM SIGGRAPH computer graphics standards committee report.)

This stage should take approximately one year and involve in addition to the conceptual designer, a computer scientist versed in data structure theory and a senior programmer. The programmer would also direct the various implementational efforts--ensuring adherance to the representational design. Documentation produced during this phase would represent an authoritative contribution to the sparse literature on graphics system design and would play a significant role in the establishment or refinement of standards in the area of highly interactive three-dimensional graphics.

While this design study would require a budget sufficient to support three senior personnel, with extensive travel for at least the conceptual designer (we feel it imperative that he/she have extensive personal contact with both system designers and user communities), it represents, in effect, an effort to salvage and rationalize 100 man years of effort in this field. If it is not done, evidence is that the next generation of system designers will again pursue machine dependent approaches. Only an organization such as NRCC with a responsibility to a large community has either the man power potential or motivation to carry out this study. Selection of the conceptual designer must be made carefully but individual panel members felt that appropriate candidates do exist and that the task of synthesis of existing systems can be accomplished with cooperation from the existing community of system designers.

Recommendation 1B: A High Performance Resource. In the long term the NRCC should aim to advance graphics technology and its application to molecular problems by providing a pioneering installation of exceptional power. Such a system should encompass the capabilities of all existing systems and have the ability to tackle problems not feasible on any machine now extant. We envisage that this would combine the power of frame buffer and calligraphic systems and be backed with sufficient computer power to permit studies not only of electron density model fitting, but of molecular dynamics and energetics of macromolecular systems. Such a system may ultimately lead to a solution to the

polypeptide chain folding problem, and perhaps enable the design of peptide drugs and analogs by sequence. The specifications for the hardware and software will develop from studies of the present status and future possibilities in high performance interactive graphics which should begin at once.

Recommendation 1C: Moderate cost ($10-15K) dedicated graphics system. Although the common language of chemistry is basically graphical (topological) in nature, the full potential of current computer graphics technology has yet to be felt in the various branches of chemistry. Interactive graphics is widely found in engineering design environments. In those laboratories where graphics systems have been established, novel and productive research possibilities have resulted. A medium price, medium performance graphics system (optimal software design for targeted hardware, intelligent terminals) needs to be designed to facilitate the availability of current graphics systems to the broader chemical (and scientific) community. An intelligent graphics terminal has "stand alone" ability to process high-level computer languages (Basic, Fortran, Pascal). It has refresh graphics capability (e.g., to permit immediate rotation, interactive query of molecular geometries) and communication interface to most computers. Current displays offer a choice of black and white, color, line drawing, or raster graphics. The extra purchase price would help minimize communication costs of terminals described in 1D.

Recommendation 1D: Exportable software for display and manipulation of molecules on inexpensive terminals. There are two large groups of chemists who do not have access to molecular graphics but would benefit from easy access to it. The first group of chemists are interested in the relation of structure to other physical properties. The second group are protein crystallographers who need ways for manually refitting their electron density maps during refinement.

We feel that the needs of these chemists could be fulfilled by making available a software package that allows for display and manipulation of molecular structures on inexpensive terminals connected to "in-house" computers.

This software could make use of conceptual and representational design proposed above (1A). However, it would be developed for terminals using storage CRT (e.g., Tektronix) or raster display. If interrogation and/or manipulation of the model is desired then capability to read x, y position from the screen would be required. This equipment could be obtained for around $5000. The software must be easy to use, easy to implement, have careful documentation, and be available within the next two years.

Medium Performance Graphics

Although low to medium cost graphics display hardware has been available for years, virtually no use is made of this equipment in synthetic, computational, or structural chemistry and biochemistry

The NRCC could provide the key, missing elements: a user-friendly package and initial hand-holding to help the bench and desk chemist take advantage of available technology. By casting program development in the broadest possible sense, this effort would benefit members of the various disciplines of chemistry.

The medium cost/performance intelligent terminal envisioned in recommendation 1C above would be used to develop a panoply of programs for the chemical community. A survey of available program packages and desirable functions would be made to effect an exportable system with short term implementation possessing long term potential impact. We recognize that 1A is an ambitious project and that it may prove impossible to design a unified scheme for such a wide spectrum of hardware capabilities. In this event, we consider that project 2, which is conceived as an interim implementation, may be promoted to a permanent solution for its class of problems.

Without providing real time (hardware) rotation, this terminal would still permit software generated manipulation and interactive modelling functions. In addition, graphics operations useful to the quantum mechanist and spectroscopist (e.g., model building, spectral display) and synthetic chemist (input to synthesis programs, structural data base querying) would be implemented. As key programs are distributed and a user group develops, specialized programs developed at individual sites could be made available to other sites.

A Suggestion for the Role of NRCC in Providing a Device Independent, Transportable Graphics Capability to the Chemistry Community

NRCC could provide a general purpose, transportable, device-independent, scientific graphics system to the chemistry community, in particular those scientists who don't work with large computation centers which provide this type of capability. Such a system would provide general scientific graphics capability in the area of two and three dimensional line graphs, bar charts, polar graphs, general coordinate transformation capability, display of two and three dimensional scalar fields as contour plots, grey level plots and surface plots, two and three dimensional vector fields and complex surface display. This would run at LBL and could provide the basis of a high quality color film and movie making system, as discussed in recommendation four. Such a system would interface to the graphic device drivers through a SIGGRAF standard interface, and would therefore be capable of operating a wide variety of non-interactive and interactive devices.

In the large, this system already exists at LBL and is called the Integrated Data Display System (IDDS). NRCC could support future development in areas of specific interest to chemists and assist in completing the task of transforming the existing code to a machine-independent Fortran with a SIGGRAPH interface. The design philosophy of this system, which is a subroutine library, has included the concepts that the user should have to provide a minimum of information to get "first try" results, but that the potential for a great deal of control

exists for more sophisticated work. It is believed that the user should be provided with many different types of graphic display techniques with which to examine his/her data in order to emphasize the data characteristics of interest. A number of graphic related utilities are provided for data transformation, smoothing, interpolation and extrapolation. Finally the design philosophy includes the concept that the code should be readily accessible to the user and that it be well documented internally so that the user may adapt, modify and specialize the code.

The impact at any given site would depend upon several factors. One easily assessable factor is that this type of software speaks to a fairly large user community and serves a variety of needs, from diagnostic graphics all the way through publication quality output. The variety of easily used tools available in IDDS would allow the user to use graphics routinely in situations when he/she might have been discouraged previously by the difficulties involved. The time to fruition of such an undertaking is again site dependent. Those sites with hardware sufficiently similar to LBL (i.e., CDC-60 bit machine) could probably make almost immediate use of the code. It is estimated that to produce a machine independent version of the code would take from six months to a year depending upon the level of effort. The problems in producing a version to run on 16 bit mini computers would require additional effort to work out overlay structures and examine the algorithms involved for minimum word length considerations.

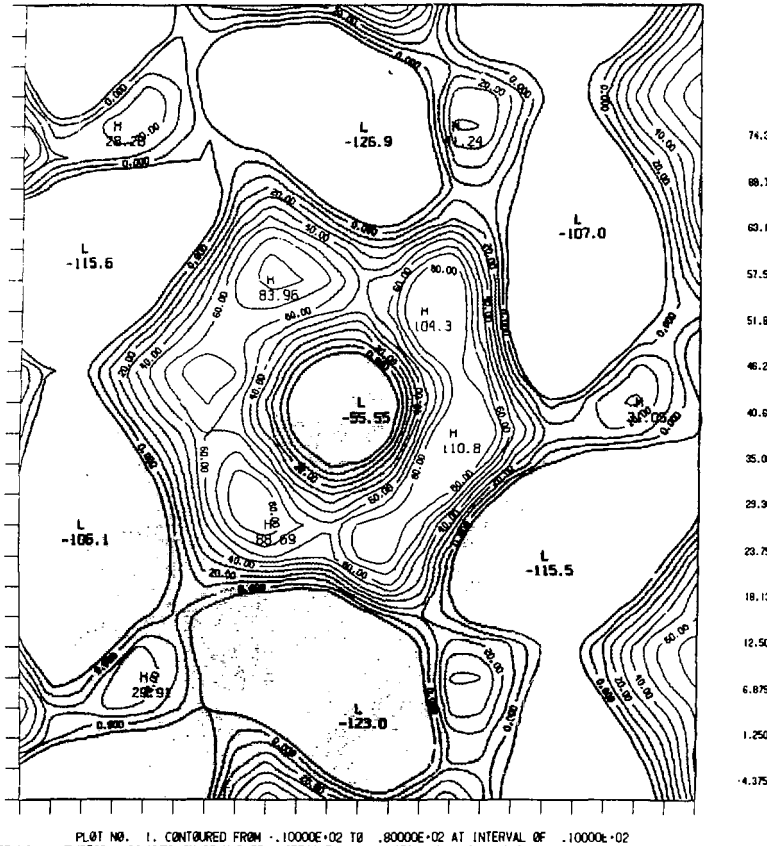
The cost of this project is estimated at 7 man months for the completion of the machine (main frame) independent version of the code, and probably a comparable effort for a minicomputer version. There is an unknown cost involved at individual sites in interfacing IDDS to local drivers if they are not SIGGRAF standard, but this is probably a fairly small effort. For sites without any type of device independent graphics system the problem would be somewhat greater. The effort over and above the implementation phase, namely the development of new algorithms and features of special interest to chemists, would be problem dependent. There would be some on going cost involved in consulting, maintenance and distribution; the magnitude of this would depend upon the size of the user community generated by NRCC, but would probably never exceed 0.5 FTE* for these specific functions, if they were desired.

Such a project is feasible. The code already exists, and is receiving active support in development, documentation and maintenance at LBL. The cost per user and even per site, of the items mentioned above, should be quite small if the code is distributed to several sites.

Figure 1 illustrates the use of multiple graphic techniques brought to bear on one set of data (electron micrograph information), the grey level display being used to provide qualitative impact to the more quantitative contour display. This graph was produced by David Grano of UC Donner Lab using IDDS at UCLBL.

---

*"Full time equivalent" -- Ed.



PLØT NØ. 1. CØNTØURED FRØM -.10000E+02 TØ .80000E+02 AT INTERVAL ØF .10000E+02
PLATE NØ. JEM5763 . SCANNED ØN 27 MAR 77 . SPECIMEN NAME S.SERPENS . ARRAY SIZE· 74 RØWS BY 64 CØLUMNS.
CØMMENTS - NEG STAIN-NØ OO-NØ CTF CØRRECTIØN-4X2 UC-CASE 1
PØRTIØN ØF ARRAY PRØCESSED · RØWS 25 TØ 51, AND CØLUMNS 23 TØ 45.

Fig. 1. XBL 7810-11921.

High Quality Color Film Recording System

It was the opinion of the computer graphics panel that in the area of graphics NRCC give high priority to establishing access to high quality color film recording capability to be available to the national community of crystallographers and physical scientists. Such a facility would give all such scientists a valuable tool which few if any of them now have and at the same time would centralize the large expense and pool of technical expertise required to operate a film recorder and processor. It is something NRCC can do which will have almost immediate impact, with an outlay of money and not of manpower.

A facility to produce gray-level, color film is needed by the scientist because it will:

- Make it possible simultaneously to see and comprehend more variables expressed as colors and gray levels, and to display time dependent processes.
- Make it easier to visualize complex 3-D structures and complicated motions and interaction.
- Quickly represent complex models which are much more accurate than sketches and also much less expensive than ball and stick models.
- Provide a more effective means of communication with fellow scientists. (As scientists become more and more specialized and deal with more and more complicated concepts, such presen tations become increasingly important.)

In addition, such a film-producing facility will aid in education and in public awareness.

It appears there are three ways in which film recording capability could now be acquired:

- It could be borrowed or rented from LLL or others. Such a solution is probably only temporary.
- A facility could be purchased and maintained by NRCC. This is a costly and long-term commitment.
- A facility could be purchased jointly with LBL. This is particularly attractive if LBL agrees to operate and maintain the equipment.

Because the film recording facility will serve a national community it should be able to handle large loads and should have as many of the currently available graphics as can be justified. Output capabilities should include:

- 16 mm black and white and color movies
- 35 mm black and white and color slides
- 4 in. x 5 in. color positives or negatives for making publishable prints

If user demand warrants, black and white fiche capability may be added. To be compatible with many graphics systems currently in use, the film recorder should accept input in a variety of commonly used forms. The film recorder should be set up to take advantage of current film recording software available. and should operate from a mini computer equipped with a keyboard terminal and disc pack drive. To ensure high quality, the recorder should have the following features:

- Additive color
- Glass filters
- 256 levels of gray
- Vector and raster output (in order to produce high quality line drawings and digital images and to allow interface with highly interactive calligraphic terminals and raster terminals.)

In order to be of most use to the scientist, turnaround must be short. Passing of data and film can be done by mail; film should be processed within one day after receipt of data.

The cost of a high-quality film recording system as described above is about $300,000. Cost of the film and processing if done in house is about 1¢ to 5¢ per frame. Yearly maintenance is estimated to be about $10,000. Mailing costs must also be considered.

Film is an excellent graphics medium because of the wide range of graphical forms it can represent. These include wire frame drawings, hidden-line drawings, hidden-surface drawings, and shaded tone drawings--all in black and white or color. Film also serves as a superior archival storage medium for interactive graphics systems and can outlive such a system by many years.

APPENDIX TO GRAPHICS SECTION

DRUG DESIGN

SUMMARY

Drug design illustrates an area in which the integrated contributions of crystallography, computation and graphics can make a considerable impact. By judicious and informed support, the NRCC can accelerate and influence developments in this field.

I. INTRODUCTION

Drug design has two main phases:

a. Discovery of a lead compound (e.g., penicillin). This was historically serendipitous, though this situation is changing, and as more three dimensional structures of large biological molecules are determined, the potential for the rational design of a specific lead compound increases.

b. Modification of the lead compound to improve its properties (e.g., potency, reduced side effects, transport, solubility). These developments are usually guided by structure-activity relationships, experimental and/or theoretical.

II. CRYSTALLOGRAPHIC DATA

Ideal:

Know the accurate three-dimensional structures of both drug and receptor.

Needed:

a. Accurate structures of drugs (e.g. plant opiates (such as morphine; endogenous opiate peptides (e.g. enkephalin, Medical Foundation of Buffalo).

b. Accurate (to the limit of the data) structures of biological molecules with interesting receptors, e.g.,

   1. Prealbumin (thyroid hormone binding protein, and model for the thyroid hormone receptor in the nucleus, Oxford, Univ. of Calif., San Francisco).

   2. Dihydrofolate reductase (methotrexate binding, Merck, Univ. of Calif., San Diego).

III. COMPUTATIONAL TOOLS

a. Structure determination and refinement

b. Conformational analysis

c. Molecular mechanics

d. Data base management and access

IV. COMPUTER GRAPHICS

a. Selective display of drugs and receptors

b. Interaction of drug with receptor

Displays available should include line drawing (calligraphic) and shaded color (raster) systems.

For communication and viewing of results, both for technical and non-technical audiences, a facility for the production of high quality films and slides is very desirable.

V. NRCC INVOLVEMENT

a. Support for determination, full refinement and conformation studies of relevant structures.

b. Data base support and access

1. Small molecules (Cambridge Crystallographic Data File)

2. Large molecules and receptor geometries -- including electron densities (Protein Data Bank)

3. Pharmacophoric pattern searching (Univ. of Calif., San Francisco; Wash. Univ., St. Louis; Texas A & M and various drug houses)

4. Structure/activity relationship tables (NIH Prophet System and many proprietary drug house files)

c. Computer graphics:

1. Transportable graphics software

2. Turn-key graphics and systems for export

3. High performance interactive three-dimensional graphics on site

4. High quality film production for presentation of results both to technical and non-technical audiences.

INFORMATION RETRIEVAL

INTRODUCTION

In recent years a number of databases have been created which consist of crystallographic information, much of it in forms which can be read by computers. The Information Retrieval Panel discussed the present status and future prospects of these compilations and studied ways to enhance their value to chemical research. This report of these discussions has a number of objectives:

- To identify the databases and indicate their current status, information content, and availability.
- To identify problem areas in their use.
- To identify areas in which changes are likely to occur.
- To identify areas in which NRCC can play a significant role in the use of databases, not only in crystallography but in other areas of chemistry, physics, and the material sciences.

CURRENT STATUS OF CRYSTALLOGRAPHIC DATABASES

There are seven crystallographic databases that are either publicly available or likely to be available. Details of their scope, size, etc., are contained in Table 1. A brief description of each database is included in the Abstracts at the end of this report. The information stored in them falls into the following categories:

1. Bibliographic citations
2. Chemical information -- compound names, molecular formulae, connectivity
3. Numerical results from single-crystal analyses
4. Results from the characterization of a material by powder methods
5. Textual comment

Not all databases are in the same state of development, nor are all of them publicly available at this time. Databases 1, 2, 3, 5 (Table 1) are available in machine-readable form; database 4, although highly developed, is not publicly available; database 6 is available only in printed form; database 7 is currently in the start-up phase.

TABLE I. SUMMARY DATA ON COMPUTERIZED DATABASES IN CRYSTALLOGRAPHY

NAME	ENTRIES (K) TOTAL	ENTRIES (K) INCREMENT PER YEAR	EXPORTABLE SIZE M BYTES	YEARS COVERED	CLASS[d]	EVALUATED	CONTACT
1. NBS CRYSTAL DATA FILE [a]	50	6	25	1913-1970	O,I,M	YES	A.D. MIGHELL, N.B.S. WASHINGTON, D.C., USA
2. POWDER DATA FILE [b]	32	2	17	1913 to PRESENT	O,I,M	YES	J.C.P.D.S., 1601 PARK LANE SWARTHMORE, PA., USA
3. CAMBRIDGE ORGANIC CRYSTAL DATA FILE	20	4	50	1935 to PRESENT	O	YES	O. KENNARD, CHEMICAL LAB, UNIVERSITY OF CAMBRIDGE, U.K.
4. METAL DATA FILE	4	0.5	f	1913 to PRESENT	M	YES	L.D. CALVERT, N.R.C., OTTAWA, CANADA
5. PROTEIN DATA BANK	0.1	0.02	20	e	BP	NO	T. F. KOETZLE, BROOKHAVEN NAT. LAB, UPTON, N.Y. USA
6. BOND INDEX [c] (BIDICS)	14	2	6	1969 to PRESENT	I (O)	NO	I.D. BROWN, MCMASTER UNIVERSITY HAMILTON, ONTARIO, CANADA
7. INORGANIC CRYSTAL STRUCTURE DATA BASE	0	0.5	f	STARTED IN 1977	I	YES	" " " " AND G. BERGERHOF, UNIV. OF BONN GERMANY (BRD)

A more complete survey of crystallographic files of all kinds is given in CODATA bulletin No. 25 (1977).
Files include, at least, space group, unit cell parameters, atomic parameters & references unless otherwise noted.

a) No atomic coordinates
b) d, I values only
c) Bond index and references only
d) I= inorganic (inc. minerals) O= organic (a few minerals) M= metals (some minerals) BP= biopolymers
e) Partial coverage; voluntary deposition since 1971
f) Not publicly available

The active participation of NIH/EPA/FDA with their efforts in CIS* is applauded. The unifying effort in coordination of data from each of the files (NBS, JCPDS, Cambridge) via the Chemical Abstracts Service registry numbers allows crystallographers and others to access these databases. It is felt that this work should be noted throughout the scientific community and that chemists interested in the continuation of such files should use the CIS systems. The cost is being partially underwritten by the U.S. Government. Thus wider utilization of the existing systems rather than duplication is to be encouraged.

A key factor of all numerical databases (except database 5 which is not derived from the primary literature) is the attention paid to critical evaluation and checking of information at input time and by cross-checking within and between databases. The stored material is therefore more reliable and self-consistent than the primary literature from which it is derived. A further important feature is the compression and systematization of the data as compared with the scattered and unsystematic form occurring in the primary literature.

Some databases (e.g., numbers 2, 3, 4 and 5) contain material which is not otherwise publicly available. It is expected that such archival functions of databases will increase during the next few years.

Since the information stored in different databases addresses different needs, there is no significant advantage in merging them into a single entity. Cross-references (pointers) between files should be stressed. There needs to be cooperation and the interchange of ideas between database compilers.

CURRENT USE OF CRYSTALLOGRAPHIC DATABASES

a. Rapid entry into the literature

b. Characterization of unknown materials

c. Preparation of indexes, handbooks and current-awareness listings

d. Use of search, retrieval and display software as a basis for research, e.g., into structural relationships, drug action, etc.

e. An archive for numerical data not available in the primary literature

f. As the source material for the production of educational tools.

---

*Chemical Information System, described by S. R. Heller, G. W. A. Milne and R. J. Feldman, Science 195, 253-259 (1977).

Despite promising high-technology approaches involving interactive graphics systems, the majority of crystallographic information retrieval is conducted by low technology methods. The machine-readable data is reduced to the form of books, cards or microfiche for supply to the user. Few small laboratories can justify the expense involved in bringing up more than one database on their local computer. To some extent this problem is being addressed by resource-sharing methods such as the Chemical Information System by batch processing at remote facilities, and by current-awareness services. At present, a, b and c above are well-developed services; but d, e and f leave room for considerable development.

PROBLEMS IN THE USE OF DATABASES

Service Uses of Crystallographic Data Bases

At present, many scientists who could benefit from exposure to the data in the various crystallographic compilations make no use of these resources. On the assumption that good science often depends on good information, and with the awareness that the data involved is available in its most reliable form in these data bases and sometimes not even available elsewhere, we stress the need to increase the use of crystallographic databases by the scientific community.

The data may be available in many forms: books, magnetic tapes, microfiche as well as in batch and interactive mode on various computers. Books have a generally good reception, but lag far behind the production of data, usually fail to include as many cross indices as may be available in other media and are passive. They allow browsing but depend strongly on the skill and persistance of the user to obtain specific information items. Microfiche save paper and are available with smaller lag-times; otherwise they share the disadvantages of books and are less convenient to use.

A useful form for data distribution is on a computer-readable medium. Dissemination on magnetic tape with suitable search and retrieval programs can satisfy the computer-wise scientist who has the time and energy to keep the data current and who has a sufficiently narrow focus to allow a small number of bases to meet his needs. For the less independent souls, or the ones in need of a broad range of bases beyond the capacity of their local facilities, a centrally maintained remotely accessible interactive or batch service is needed (e.g., CIS system). There are, however, barriers to the utilization of databases. These are highest for the case of remote use but are also substantial for batch work against the original data files.

Lack of Awareness. Although many scientists are aware of these compilations the message has not registered, and they remain unaware of what these databases can do for their research.

Lack of the Necessary Skills. Working with these bases requires that the user be confident that any lack of sophistication in

crystallography and/or computing will not detract from his ability to extract useful information.*

Cost. Even with interactive searching available at a very cost-effective rate, many scientists perceive the expenditure as excessive. Most funds for support of research are subject to restrictions or prior commitments, and often one has little money for this kind of discretionary expense.

Lack of Commonality. Each database is different and requires some training for effective use.

Difficulties in Browsing. Books allow the user to view a randomly chosen portion of the data but on computers this particular interactive mode is too expensive unless CPU and connect charges are small (e.g., in-house mini's).

Any effort by the database purveyors and other parties, such as the NRCC, to lower these barriers is highly desirable. The primary needs are for exposure, easier access, and education. Education as to the existence, use and coverage of the databases is needed. NRCC should attempt to pass such information on to its constituency and potential users. Ease of access is needed. NRCC should maintain current versions of such bases as The Cambridge Data File, The Protein Data Bank, The Powder Data File, and The Crystal Data File, and also the Inorganic File and the Metals File when these become available. In addition, the NRCC should maintain an account with the NIH/EPA Chemical Information System (CIS).

The CIS should satisfy many users, for whom tapes would merely be archival. However, the local copies of the tapes would permit research in any future direction. In such an environment, NRCC should attempt to provide users with mini-seminars in database use and capabilities, exposing increasing numbers of scientists to this material.

Lack of Exploitation of the Data Available in the Databases

Scientists responsible for building the databases have often desired an opportunity to do research with such sources of data. As they work with the data they become aware of many excellent research problems (of varying degrees of complexity) that could be investigated. In general, however, these compilers have been overwhelmed with merely handling the deluge of data, and consequently have had little chance to study or analyze the information themselves. Likewise, those writing the search and retrieval software are also fully occupied.

For several reasons it is now appropriate to start this research-oriented database activity. First, scientific research (and the resulting publications) derived using the database as a source will

* The extent of skill required to justify that confidence is the important question--Ed.

clearly demonstrate its creative research value (in contrast to the more mundane supportive research and retrieval value). This would help to demonstrate the value of the database in its own right. Second, scientific research at the data centers would make database building a more stimulating and rewarding task for the builders and help to elevate the prestige of the database operations. Third, the scientific research would undoubtedly lead to improvements in the processes of data capture, evaluation, and classification. A concrete example of such research is cited in the Appendix I of the Graphics Panel Report.

FUTURE ADVANCES

There are a number of areas where advances (technological, economic and in the databases themselves) will alleviate some of the problems identified above.

(a) More effective use of on-line databases (e.g., satellite communications).

(b) Improved technology for the transport of data (i.e., a replacement of the magnetic tape by a more compact and transportable storage device).

(c) An increase in the use of existing databases as archival depositories and the direct transfer of data in machine readable form from the author to the journal and to the database.

(d) Improved funding for database construction and use.

All of these require the impetus of an active user community. This can be brought about by increased awareness, which in turn can come from greater use.

It is also suggested that the range of existing databases could be expanded to include compilations of

(a) Magnetic structures.

(b) Measured and calculated electron densities (it may be preferable to store this information as structure factors). At present this information is only published in a rudimentary form, much of which could not possibly be recovered now.

(c) The contents of Volumes I and IV of the International Tables for X-Ray Crystallography may have a market in a machine-readable form.

However, we do not envisage a formal role for NRCC in establishing such databases.

INTERACTION WITH AREAS OUTSIDE CRYSTALLOGRAPHY

Crystallography provides the information on which most perceptions of structure in chemistry and biochemistry are based. In addition, it categorizes and identifies crystalline solids in the whole of materials science. Because of this breadth, we have a special mandate to interact with scientists in the areas of chemistry, physics, biochemistry, biology, metallurgy, geology, etc. Data must be transferred and a good way is by computer-readable databases which, incidentally, can provide hard copy for browsing.

RECOMMENDATIONS

Database activities fall into four basic areas:

1. Data capture, evaluation, and database construction.

2. Data dissemination, and the provision of access to databases.

3. Education of the scientific community as to the existence of the databases, their information content and accessibility.

4. Research using the data now in databases and research on the databases themselves to produce better management and retrieval systems.

The first two areas do not require NRCC support. These areas are currently supported by diverse groups and individuals. The last two areas, however, are "problem" areas in which the panel recommends that NRCC play a supportive role. Few chemists are aware of the databases or know how to access them. Therefore, the primary NRCC thrust should be directed to the area of education. Immediately following this educational activity, the panel would like to suggest that NRCC support research activities in connection with the data bases by supplying seed money. Table 2 summarizes the areas in which NRCC could play a role.

TABLE 2

RECOMMENDED NRCC INFORMATION RETRIEVAL ACTIVITIES

I. Education of Potential Users

1. Provision of access to databases at NRCC for local use by visitors, through CIS and direct files. Estimated cost $10 K/year.

2. Establishment of tutorials and workshops for broad scientific groups, e.g.,

   a. Workshop at ACS/ACA Hawaii meetings (1979) in collaboration with CIS. Estimated cost $2 K/workshop

   b. Collaboration with AAAS in education effort for scientists around the U. S. Estimated cost $6 K/year.

3. NRCC should join CODATA. Estimated cost $0.1 K/year.

II. Research on Data

1. Support of personnel to examine correlations among data. Estimated cost $20 K/year;

2. Use of interactive graphics with databases. Estimated cost not known. See, however, the graphics report.

III. Research on Databases on Retrieval Techniques

1. Collaborative Projects among database compilers and users (e.g., on retrieval techniques). Estimated cost $20 K/year;

2. Investigation of standardized nomenclature and data descriptions to allow proper intercollation of data from disparate sources. Estimated cost not known.

ABSTRACTS

PHASING

Overview of Current Developments in Phasing Techniques 49

Non-Crystallographic Symmetry as a Phasing Tool. 51

The Relationship Between Isomorphous Replacement and Direct Methods 52

Finding Molecular Fragments: Demands on NRCC. 53

Developments Which I, as a User, Would Like to See in Phasing/ Structure Solution 54

Toward Making More Efficient Use of Molecular Structure Information in Solving Crystal Structures 55

Solution of the Diffraction Phase Problem by Molecular Packing Analysis 56

NUMERICAL REFINEMENT

Algorithms and Software of Optimization 57

Conjugate Gradients for Linear Least Squares 59

Some Questions, Prejudices, and Ideas 62

Crystallographic Refinement of Macromolecules 63

Precision Crystal Structure Analysis and AB-Initio Molecular Orbital Quantum Mechanics 64

A Robust/Resistant Algorithm for Crystal Structure Refinement. . . . 66

Stereochemically Restrained Crystallographic Least-Squares Refinement of Macromolecule Structures 67

GRAPHICS

Bilder: A Manipulative Graphics Program for Biopolymers 68

Feasibility and Usefulness of an Efficient, General, Distributable Electron Density Fitting Program for Cheap Graphics Terminals 71

A Suggestion for the Role of NRCC in Providing A General, Transportable Graphics Capability to the Chemistry Community 72

A Reading List of Books on Computer Graphics, Programming Style and Structural Programming 73

A Position About NRCC and Computer Graphics hardware 74

Status and Plans for Molecular Graphics 75

The MMS-X System 77

Molecular Modeling--Biographics Laboratory 79

INFORMATION RETRIEVAL

Description of the JCPDS Power Diffraction File 82

The Cambridge Crystallographic Database: Search, Retrieval, and Display 84

The Metal Data File 85

Structural and Crystallographic Data for Biological Macromolecules . 86

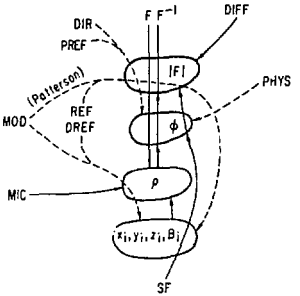
NBS Crystal Data File 88

Bidics 89

Inorganic Crystal Structure Data Base 90

OVERVIEW OF CURRENT DEVELOPMENTS IN PHASING TECHNIQUES. David Sayre, IBM Research Center, Yorktown Heights, NY 10598.

PHASING OF DIFFRACTION DATA (SCHEMATIC)



1. The figure shows the network of methods existing today for phasing diffraction patterns and solving structures.

2. Some of the methods fall wholly or in part in the scope of other panels. Main examples: REF (refinement) and DREF (Diamond-type refinement), panel 3; MOD (model-building), panel 4.

3. The situation as a whole is very active today in terms of major new developments. Examples: PHYS

(physical phasing), use of synchrotron radiation in anomalous dispersion phasing, Post's technique; DIFF (diffraction), small-sample crystal growth techniques, dynamical theory for microcrystal ED; (F and F^{-1} (Fourier and inverse Fourier transformation), Winograd superfast algorithms; DIR (direct methods), higher-order structure invariants and seminvariants, MAGLIN; PREF (phase refinement), use of non-crystallographic symmetry in large proteins and viruses, least-squares and matrix methods in smaller proteins; MIC (microscopy), near-atomic resolution of inorganic materials, use with electron DIFF as protein structure-solving technique.

4. Due in part to these developments, phasing techniques continue generally adequate to needs of crystallographers despite continual growth in number and complexity of structures undertaken.

5. U.S. is making a very large contribution to these developments, but we lag Europe in training of future researchers through workshops, schools, special conferences, etc. (e.g. Prague, Twente, Erice).

NON-CRYSTALLOGRAPHIC SYMMETRY AS A PHASING TOOL. Patrick Argos, Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907.

Just as computers have grown in magnitude from the mid-1950's so has the size of biologically significant structures examined by means of X-ray diffraction. Examples include chromatin, small plant viruses and even human polyoma virus. The great plethora of information required to determine such complex tertiary structures naturally follows. The typical isomorphous replacement technique used to phase structure factors necessary to calculate Fourier electron densities also finds itself a victim of the larger molecules. The positions of heavy atoms are more difficult to locate; their lessened comparative scattering ability contributes less information for the phase solution. The concomitant computing tasks involving the large data sets quite naturally become formidable.

The utilization of non-crystallographic symmetry (molecular replacement) displayed by subunit aggregation in macromolecules has provided an important supplement or substitute for the isomorphous replacement phasing technique. Five areas of molecular replacement application can be enumerated: (1) the determination of an unknown structure from similar but known structures, (2) solution of a completely unknown structure which possesses non-crystallographic symmetry, (3) deconvolution of heavy atom difference Pattersons, (4) phase extension, and (5) ab initio phasing.

The first requirement of molecular replacement, the determination of the orientation of the rotational axes, is normally achieved with relative ease especially given the Crowther fast rotation function. The second requirement, the determination of translational parameters, is frequently found to be difficult in all but special cases and still presents formidable unsolved problems both theoretical and computational when no structural information is available.

The most widely used application of molecular replacement has been in the solution of unknown structures from similar but known structures; for these cases heavy atom derivative data are non-essential. A knowledge of the orientation of the non-crystallographic symmetry axes can provide vectors to deconvolute automatically heavy atom difference Pattersons. Computational limits still exist when the symmetry is so high that thousands of vectors have to be calculated in the Patterson search for each postulated heavy atom site.

Real-space averaging has been used primarily for phase improvement at a given resolution in conjunction with isomorphous replacement; in this form molecular replacement has had substantial success. However, severe computational problems and ill-understood constraints on the solution still exist. Presently there are only two programming systems to effect molecular replacement cycles for large complex structures. Minimum core requirements approximate 150K words and at least hundreds of minutes of calculations are required for the larger molecules. Presently it would thus seem advantageous to establish computing centers for molecular replacement applications from several research groups. Alternatively computer programs could be designed and made available to match core and time limitations of a particular research center.

Phase extension by molecular replacement has only been used on an already known protein structure and shown to be viable. It must now be applied to an unknown structure; further, the degree of extension possible per cycle must be accurately determined and improved. Ab initio phasing by molecular replacement is still beyond the horizon both in real and reciprocal space.

THF RELATIONSHIP BETWEEN ISOMORPHOUS REPLACEMENT AND DIRECT METHODS.

E. A. Green, Medical Foundation of Buffalo, Inc., Buffalo, NY 14203.

It is well known that current *ab initio* direct methods are ineffective in solving macromolecular crystal structures; isomorphous replacement, on the other hand, is the method of choice. However, once the values of a body of native phases are assumed, it is feasible to use a number of different direct method techniques for phase extension and refinement. In view of these diverse applications, one might assume that the underlying assumptions in the two methods are radically different. It will be shown that the two methods are special cases of a more general system based on the use of prior structural information. An example will also be given to make plausible the notion that if properly combined, a joint system of isomorphous replacement-direct methods yields more phase information than either system alone can provide.

Stripped to bare essentials, isomorphous replacement methods obtain phase information because of their ability to exploit the relations among the structure factors associated with a single reciprocal lattice vector but corresponding to related structures (native, derivative, heavy atom); in short, correlations of one structure factor in different cryscals. In contrast, direct methods assume a single crystal, native for example, and obtain phase information by exploiting relations among structure factors having different, but related, reciprocal lattice vectors, $\underline{h}$, $\underline{k}$, $\underline{\ell}$ .... It would seem reasonable to consider correlations among different structure factors in different crystals, e.g. $F_{\underline{h}}$, $D_{\underline{k}}$, $H_{\underline{\ell}}$, which is shown in the figure. This last case represents the total interaction between the isomorphous replacement and direct methods.

$$
\begin{array}{ccccc}
F_{\underline{h}}, & F_{\underline{k}}, & F_{\underline{\ell}}, & \ldots, & F_{\underline{m}} \\
D_{\underline{h}_1}, & D_{\underline{k}_1}, & D_{\underline{\ell}_1}, & \ldots, & D_{\underline{m}_1} \\
H_{\underline{h}_1}, & H_{\underline{k}_1}, & H_{\underline{\ell}_1}, & \ldots, & H_{\underline{m}_1} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
D_{\underline{h}_n}, & D_{\underline{k}_n}, & D_{\underline{\ell}_n}, & \ldots, & D_{\underline{m}_n} \\
H_{\underline{h}_n}, & H_{\underline{k}_n}, & H_{\underline{\ell}_n}, & \ldots, & H_{\underline{m}_n}
\end{array}
$$

The structure factors $F_{\underline{h}}$, $D_{\underline{h}_n}$, $H_{\underline{h}_n}$ are from the respective native, *nth* derivative and heavy atom structures.

FINDING MOLECULAR FRAGMENTS: DEMANDS ON NRCC. Eaton Edward Lattman, Department of Biophysics, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA.

Protein molecules and domains often belong to a family in which all the members bear a certain structural similarity. The crystal structure of one member can often be determined directly by using another, known structure in the family. Typically, one attempts to assemble a (computer) model of the crystal, in which the test molecule and its symmetry mates are positioned and oriented in a unit cell so as best to account for the observed diffraction pattern. Such models can generally be refined to eliminate the bias introduced by the test structure. In the simplest case, one molecule in the crystallographic asymmetric unit, only six variables specify the model completely: three angles to determine the molecular orientation, and three translations to fix its position. The six values are often determined by search procedures. Fortunately, the process can be carried out as a sequence of two three-dimensional searches, rather than as a single six-dimensional one. A number of widely circulated computer programs allow such searches to be carried out quite routinely. When successful they provide a great saving in cost and labor compared with isomorphous replacement.

As the difficulty of the problem increases only modestly, however both the amount of computation and the variety of search procedures called for increase very rapidly, often becoming prohibitive for an individual crystallographer. The immunoglobin fragment Fab, for example, is composed of two domains whose tertiary structures are fairly well conserved from molecule to molecule. The domains are linked by a flexible region, however, so that their orientations and positions have considerable independence. In a structure under study by Barry Wishner there are two fragments or four domains in the asymmetric unit. The effect of this is drastically to reduce the signal to noise ratio in the search correlation functions: four weakened peaks must be found, and they must all be right. In practice this means that a great number of wrong answers must be examined before a correct one is found. Results from several different search protocols may also have to be considered. Nevertheless, the search procedures are often much quicker and cheaper than isomorphous replacement.

NRCC can provide a number of services that will decrease the cost and encourage the use of these calculations.

(1) Almost all search programs use Fourier transforms extensively. For any given array length special purpose routines much faster than the normal FFT can be written, if careful attention is paid to factoring and to the timing of machine instructions. A library of routines for common array lengths would be valuable. Proper hardware FFT capability would do the same.

(2) Many search procedures use the finely sampled Fourier transform of the test molecule. NRCC could maintain a library of such transforms, either for central use or for distribution.

(3) NRCC could encourage the establishment of a documented library of search programs by offering special rates to those implementing a procedure for the first time.

(4) Handholding and tutorial services would also be useful. Search procedures have the reputation (possibly fostered by the high priests of the discipline) of being difficult and confusing to use. Documented test problems for the various programs for example would be very helpful in boosting crystallographers over the start-up potential barrier.

DEVELOPMENTS WHICH I, AS A USER, WOULD LIKE TO SEE IN PHASING/STRUCTURE SOLUTION.

Mary Frances Richardson, Department of Chemistry, Brock University, St. Catharines, Ontario. L25 3A1, Canada.

Technique for solving the following problems more or less routinely would be useful:

1. Disorder, particularly if it involves heavy atom location
2. Psendosymmetry, especially the types involving psendotranslation along an axis or molecular locations so that he apparent space group is of higher symmetry than the true one.
3. Nonrigid molecules with 100 - 200 nonhydrogen atoms and no heavy atoms, e.g. peptides with 10 - 20 amino acid residues.

The NRCC could give the following types of support to crystallographers:

1. Act as a distribution centre for programs similar to QCPE
2. Develop, or encourage development, of a system of crystallographic programs for use on NRCC computers. Users could "rent" the programs and buy computer time for solving structures their own computers couldn't handle.
3. Sponsor tutorial sessions of a week or so if a program system as visualized above were available. Practical experience in solving structures could be coupled with lectures on theory. These could be either elementary, with focus on standard techniques, or advanced, where new developments are discussed. People who took the tutorial sessions might be encouraged to bring along an unsolved structure.
4. Compile a list of structures (and associated structure factor data) which were difficult to solve by standard techniques. These could provide a testing ground for future programs, and aid in evaluating the efficacy of a given technique for solving structures.

TOWARD MAKING MORE EFFICIENT USE OF MOLECULAR STRUCTURE INFORMATION IN SOLVING CRYSTAL STRUCTURES.* W. E. Thiessen, Chemistry Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830.

A class of crystallographic problems of great biochemical interest involves molecules of 1000-2000 daltons whose crystals may include considerable solvent (usually disordered) and give X-ray diffraction patterns which die out above $\sin\theta/\lambda$ of 0.4 to 0.5.

In a large fraction of such problems the molecular connectivity is largely (often completely) known and the desired information has to do with the molecular conformation. In other words, we have available at the outset a detailed model containing predictable bond distances and angles but with some conformation angles to be determined.

How can we best utilize this wealth of prior information in solving structures? Some beginnings have been made — various orientation and translation search schemes exist for rigid molecular fragments, and MULTAN-78 utilizes structural information in normalizing structure factors and computing weights and phases for the triplets. As we search for more effective ways to utilize known structural information as input into the phasing process, we shouldn't forget the traditional method of using molecular information in direct methods: recognizing the structure in the E-map. The nature of the class of problems under discussion here is such that this pattern recognition may be more complicated than simply plotting the largest peaks in whatever approximation to the electron-density map has been produced by the structure-solving method employed.

---

*Research sponsored by the Division of Materials Sciences, U. S. Department of Energy, under contract W-7405-eng-26 with the Union Carbide Corporation.

SOLUTION OF THE DIFFRACTION PHASE PROBLEM BY MOLECULAR PACKING ANALYSIS. D. E. Williams, Department of Chemistry, University of Louisville, Louisville, Kentucky, 40208.

This method is most useful in cases where the molecules have a more or less predictable rigid structure or have a limited number of possible intramolecular bond rotations. For one rigid molecule in the asymmetric unit there are only six variables (rotations and translations), and calculations may be very fast. For hydrocarbens symplified repulsion-only quadratic nonbonded potentials are effective for the purpose of obtaining a trial model which can be used as input for structure factor refinement methods. The lattice energy minimizations may be carried out on a rather coarse translation-rotation grid, since the convergence range is usually large. Many false minima are usually present in the nonbonded energy surface. When reasonably accurate nonbonded potentials are used, the global minimum is usually sufficiently close to the observed structure to allow structure factor refinement. For nonhydrocarbon molecules more work is needed to well define the nonbonded potential energy functions. These functions should be extended to also include hydrogen bonding interactions.

ALGORITHMS AND SOFTWARE FOR OPTIMIZATION. Margaret H. Wright, Department of Operations Research, Stanford University, Stanford, California 94305.

The problem to be considered is:

$$\text{P1:} \qquad \text{minimize} \quad f(x), \quad x \in E^n$$

$$\text{subject to} \quad c_i(x) = 0, \quad i = 1, 2, \ldots, m',$$

$$c_i(x) \geq 0, \quad i = m' + 1, \ldots, m,$$

where f and $\{c_i\}$ are general nonlinear functions. Efficient theoretical algorithms for P1 are typically based on iteratively approximating the general problem by a simple sub-problem whose solution can be readily computed -- for example, Newton's method in the unconstrained case uses a positive definite quadratic approximation to $f(x)$. For any practical implementation, it is necessary to introduce "safeguards", to allow for the possibility that the simplifying assumptions may not be reasonable.

Effective algorithms and software are available today for the unconstrained problem. If the constraints of P1 are linear, techniques are also available for directly reducing the dimensionality of the minimization sub-problem, even for the large sparse case.

With nonlinear constraints, algorithm design is more complicated. Early approaches, such as penalty function methods, are known to be inefficient and to generate a sequence of increasingly ill-conditioned sub-problems. Several promising algorithms today rely on local linearization of the optimality conditions, and on using the limiting properties of penalty function methods in a numerically stable manner.

References:

1. Dixon, L., ed. (1976). Optimization in Action, Academic Press.
2. Gill, P. and W. Murray, ed. (1974). Numerical Methods for Constrained Optimization, Academic Press.
3. Jacobs, D., ed. (1977). The State of the Art in Numerical Analysis, Academic Press.
4. Murtagh, B. and M. Saunders (1978). "Large-Scale Linearly Constrained Optimization", Mathematical Programming.

CONJUGATE GRADIENTS FOR LINEAR LEAST-SQUARES. Eric Grosse, Computer Science Department, Stanford University, Stanford, California 94305.

After linearization, the least squares refinement problem may be stated as:

$$\text{minimize} \quad \sum_h w_h (F_h^{obs} - f_h^{calc}|_{x'} - \sum_i \frac{\partial f}{\partial x_i}|_{x'} \, \delta x_i)^2 =: \|J\delta - f\|_2^2.$$

To improve the data/parameter ratio in macromolecular refinement, stereochemical restraints may be added, giving the problem:

$$\text{minimize} \quad \|\begin{bmatrix} J \\ \rho C \end{bmatrix}\delta - \begin{bmatrix} f \\ d \end{bmatrix}\|_2 =: \|A\delta - b\|_2.$$

The minimum is attained if $A^T A\delta = A^T b$, but explicitly forming the normal matrix can cause problems. Because $A^T A = J^T J + \rho^2 C^T C$, data may be lost due to rounding if ρ is large. Define the condition number κ^2 as the ratio of the largest to the smallest eigenvalue of $A^T A$. Errors from rounding $A^T A$ or, if the R value is not very small, errors in b can be magnified by κ^2 when computing δ. Common sources of ill-conditioning are: unfortunate choices of parameter units, dependency in the model, and strong weights on some of the equations. Changes of variables -- for example, scaling so that $A^T A$ has equal diagonal elements -- can improve conditioning. The conjugate gradient method solves the normal equations iteratively without needing to explicitly form $A^T A$. One step of the method is given by

$$a_i \quad := \|A^T r_i\|^2 / \|Ap_i\|^2$$

$$\delta_{i+1} := \delta_i + a_i p_i$$

$$r_{i+1} := r_i - a_i \, Ap_i$$

$$b_i \quad := \|A^T r_{i+1}\|^2 / \|A^T r_i\|^2$$

$$p_{i+1} := A^T r_{i+1} + b_i p_i \, .$$

Here the vectors $A^T r_i$ are gradients of the function $\frac{1}{2}\|A\delta_i - b\|_2^2$ and they satisfy the conjugacy property $(A^T r_i)^T A^T A (A^T r_j) = 0$ for $i \neq j$. From such orthogonality properties, it follows that δ_i exactly solves a nearby problem if $\|A^T r_i\|_2$ is small. With the starting values $p_0 := A^T b$ and $\delta_0 := 0$, $\delta_{i+1} = Q_i(A^T A)A^T b$ for some polynomial Q_i which happens to minimize the weighted residual

$$\|\delta_* - \delta_{i+1}\|_{A^T A} = \|[I + Q_i(A^T A)A^T A]\delta_*\|_{A^T A} \, ,$$

where δ_* is the exact solution and $\|\delta\|_{A^T A}$ means $\delta^T A^T A \delta$. By expanding δ_* in eigenvectors of $A^T A$ and using particular (necessarily optimal) polynomials, estimates of the convergence rate can be obtained. For example, if $C\delta_* = 0$ (that is, the shift is orthogonal to the constraints) then a conjugate gradient method is not bothered by the large eigenvalues corresponding to constraints.

Conjugate gradients is based on the tridiagonalization process of Lanczos. If V_k is the matrix whose columns are normalized copies of $p_0, \ldots, p_k$, then $(V_k V_k^T)A^T A(V_k V_k^T) = (V_k V_k^T)$, so $V_k V_k^T$ is an estimate of $(A^T A)^{-1}$. Note that the estimate of diagonal element σ_{ii} is $\sum_j v_{ij}^2$, so the estimates can be easily updated. How rapidly they converge is unknown.

References:

1. Bjorck, A. (1976). "Methods for Sparse Linear Least-Squares Computations", in Sparse Matrix Computations, pp. 177-199, Academic Press.
2. Chandra, R. (1978). "Conjugate Gradient Methods for Partial Differential Equations", Research Report 129, Department of Computer Science, Yale University.
3. Elfving, T. (1978). "On the Conjugate Gradient Method for Solving Linear Least-Squares Problems", Report LITH-MAT-R-1978-3, Department of Mathematics, Linkoping Univer
4. Murray, W. (1971). "Analytical Expressions for the Eigenvalues and Eigenvectors of the Hessian Matrices of Barrier and Penalty Functions", J. Opt. Th. Appl., Vol. 7, pp. 189-196.
5. Paige, C.C. and M.A. Saunders (1978). "A Bidiagonalization Algorithm for Sparse Linear Equations and Least-Squares Problems", manuscript, to appear.

SOME QUESTIONS, PREJUDICES, AND IDEAS ABOUT REFINEMENT, DOUBTLESS INCLUDING CONSIDERABLE CONSEQUENCES OF IGNORANCE, ARE LISTED WITH THE HOPE THAT THEY MAY WARRANT FURTHER EXAMINATION. Verner Schomaker

Major questions: How try to be sure the structure is correct, the treatment (fixed assumptions of the model, such as iso's vs aniso's) adequate, the refinement converged? How facilitate the refinement and make it less expensive? How assess the reliability and accuracy of the results? How make the most of data that are limited in number or precision? How make the best use of a priori information?

In the early stages of refinement, which perhaps reach to the very end if the GOF, honestly estimated, remains large, the (random) experimental error may be inconsequential. Then manipulate the weights to facilitate the refinement by reducing the off-diagonal terms. This means simplifying the peak shapes of the corresponding difference map, the least-squares and the differential-difference-map formulations being equivalent or almost equivalent (the off-diagonal terms correspond to the ripples that surround any non-zero interaction in the difference map and imply corrections to the shifts inferred for neighboring interactions). And if the error weights are replaced by a simple function, it should be possible somehow to estimate the elements of the normal-equations matrix by methods much less expensive than the usual straightforward term-by-term summation.

The essential issue, however, is the non-linearity, apparently in restrained refinements as much as in the unrestrained. To find a better way than mere iteration! One observation: The lowest-order correction to the linear result, when terms quadratic and higher in the parameters are included, can be seen to be of the order of BVV, where B represents the matrix of quadratic terms and V the vector of residuals to be fitted. This correction is large if the curvature (B) is large and the residuals are large, whether because large parameter shifts are required or because the total error, random and systematic, is large. It is to be feared, therefore, that both noise in the data and inadequacies of the treatment will aggravate the effects of non-linearity.

When to go over to error weights, how to make the transition, how to deal with apparent systematic error, how (in the first place) to design the data collection for best eventual results? If the counts obtained are indeed precious, all should be properly used, including peak counts less than the estimated background.

Finally, some comment about R tests et al. They make nonsense if the absolute weights are known and the primary conclusion has to be that the odds for such poor agreement are infinitesimal, that the models are both wrong. (But, strictly, the counting weights aren't known; they also must eventually be refined, mustn't they?)

CRYSTALLOGRAPHIC REFINEMENT OF MACROMOLECULES.

Keith D. Watenpaugh, Department of Biological Structure SM-20, University of Washington, Seattle, Washington 98195.

While the least-squares refinement of small structures has become relatively routine when high resolution data have been collected on good crystals, there are still many instances in which refinement is not routine. Refinements of macromolecules have proved universally difficult. The small protein rubredoxin has been refined by difference syntheses and unconstrained least-squares to a conventional R-factor of 0.12 for the 1.2 Å sphere of data and 0.09 for the 2.0 Å sphere. The success of the refinement is in large part due to the data extending to atomic resolution and the overdetermination (11,000 observations and about 550 atoms). Estimated standard deviations of many of the atomic positions are 0.05 Å or better. Regularization of bond lengths and angles in the model does not significantly increase the R factors. Rubredoxin is not an atypical protein in the sense that the B parameters are large, a number of the atoms are disordered and there are regions of non-discrete solvent.

Other macromolecules are being refined by both real space and reciprocal space methods with data sets of 2.0 Å resolution or poorer. At this resolution, constraints or restraints must be applied to maintain physically reasonable models and a degree of overdetermination. In general, these refinements have failed to yield R factors below 0.20 to 0.25 when using 2.0 Å resolution data sets. The data are clearly better than this and the experience with rubredoxin indicates that more correct models are capable of lower R factors. Reasons for the termination of refinements at these relatively high R factors must be critically evaluated with regard to new refinement techniques, effects of constraints, model errors and their accuracy.

PRECISION CRYSTAL STRUCTURE ANALYSIS AND AB-INITIO MOLECULAR ORBITAL QUANTUM MECHANICS

G. A. Jeffrey, Department of Crystallography, University of Pittsburgh, Pittsburgh, Pennsylvania 15260 USA

One of the interesting interfaces that is developing rapidly between Chemistry and Crystallography is that between ab-initio type Quantum Mechanics and precision Crystal Structure Analysis. Due to the advances in nunber-crunching computing power provided by the new virtual memory "midi" computers, it is possible to study the same molecule by both methodologies at a level of detail that permits quantitative comparisons between theory and experiment rarely possible in the Molecular Sciences. A first-row element molecule, such as erythritol, $C_4H_{10}O_4$, can be examined by the RHF molecular orbital method with 431-G (now) or 631-G* (by 1980) basis sets to provide optimized molecular dimensions (bond length, valence angles and torsion angles), the electron density distribution, and such related properties as the Mulliken populations, dipole moments and perhaps, in the future, optical rotation. The same molecule can be studied in the crystalline state at temperatures from close to the melting point to 10°K by modern X-ray and neutron diffraction methods to provide molecular dimensions and electron density distributions (c.f., Jeffrey, Pople, Binkley and Vishveshwara, JACS, 100, 373, 1978). The theoretical and experimental molecular dimensions and electron densities have comparable precision, i.e. standard deviations of less than 1 p. in distances, 0.5° in angles, and 0.1 e $1 Å^3$ in electron densities. Yet the results are not strictly comparable. The theory refers to with the isolated molecule at rest, while the experiment deals with a molecule undergoing thermal and zero-point energy motion in a crystal field. These differences in molecular state and environment will produce differences in molecular geometry and in electron density distributions that are significant in terms of the experimental standard deviations. If we consider that the "definitive" ground-state chemical molecule is the isolated molecule at rest, we can ask the question, how close to "definitive" are the molecules we observe in the gas (ED and micro-wave), liquid (NMR) or crystalline states (crystallography)? To establish a common base we can extrapolate the experimental results to those of an isolated molecule at rest, or extrapolate the theoretical results to a molecule undergoing thermal motion in the observed crystal field. At present, we have adequate systems of programs for the theoretical calculations of the molecule at rest, e.g. Gaussian 70, 74 and 77 of Pople's methodology (Ditchfield, Hehre and Pople, J. Chem. Phys. 54, 724, 1971), and comparable methods of others. The program systems for the data reduction, phase determination and refinement for X-ray or neutron crystal structure determinations for small molecules are extremely well-tested and documented. R-factors of 2% between symmetry equivalent and calculated and observed structure amplitudes can be routinely obtained and lower levels of disagreement might be obtained were special efforts made. What we do not have are well-tested and well-documented programs for the application of the thermal motion and the crystal field effects.

It is common practice in high precision X-ray neutron crystal structure analyses to apply the rigid-body thermal motion analyses of Schomaker and Trueblood (Acta Cryst. B24, 63, 1968), followed by the corrections to bond lengths of Busing and Levy (Acta Cryst. 17, 142, 1964). These programs are used world-wide in all high precision structure analyses, but they are inadequate for this purpose. We need the same type of high quality programs which do not necessarily require either the rigid-body approximation or the harmonic motion approximation.

The subject of crystal field distortions received much attention a few years ago, directed towards predicting lattice vibrations and molecular geometry from "standard" molecular dimensions, unit cell dimensions, space group symmetry and semi-empirical intra- and intermolecular potential functions (e.g. Busing, 1972 WMIN,

unpublished work; Huler and Warshel, Acta Cryst. B30, 1822, 1974; Warshel, Huber, Rabinvitch and Shakkard, J. Mol. Struct. 23, 173, 1974; Caillet and Claverie, Acta Cryst. A31, 448, 1975). These programs exist in several varieties with different approximations. They are poorly documented and difficult to put into operation, especially if there is a change of computer hardware. Here the primary need is for some centralized documentation, critical comparison and evaluation of the existing programs.

A ROBUST/RESISTANT ALGORITHM FOR CRYSTAL STRUCTURE REFINEMENT.

E. Prince, National Measurement Laboratory, National Bureau of Standards, Washington, DC 20234, and W. L. Nicholson, Battelle Pacific, Northwest Laboratories, Richland, WA 99352.

If a set of data points has a Gaussian error distribution and each point is weighted in proportion to the reciprocal of the variance of its error distribution, the method of least-squares will provide estimates of the parameters in a model with minimum variances. A fitting technique is "robust" if it gives estimates with variances near the minimum for a variety of error distributions and weighting schemes and "resistant" if it is insensitive to small subsets of the data. Least-squares is neither robust nor resistant. In crystallography the conditions for least-squares to give the best results are rarely, if ever, met. Tukey[1] has suggested modifying the weights such that $w' = w(1 - 2x^2 + x^4)$ for $|x| \leq 1$, $w' = 0$ for $|x| > 1$, where $x = (|F_o| - |F_c|)/k\sigma$, σ is the standard deviation of F_o as estimated from some empirical formula, and k is a measure of scale. We have studied the behavior of this procedure, with k set equal to 9 times the median value of $\Delta F/\sigma$ for the previous refinement cycle. Experience with a number of real data sets has shown the procedure to be resistant, while a study using synthetic data suggests that it is also robust.

1. J. W. Tukey in Critical Evaluation of Chemical and Physical Structural Information, p3, Nat. Acad. Sci. (1974).

STEREOCHEMICALLY RESTRAINED CRYSTALLOGRAPHIC LEAST-SQUARES REFINEMENT OF MACROMOLECULE STRUCTURES. Wayne Hendrickson, Naval Research Laboratory.

A new procedure has been developed for the crystallographic refinement of atomic models of proteins and other biological macromolecules. It optimizes the match of the calculated diffraction pattern of the molecular crystal with the observed structure factor amplitudes while simultaneously incorporating, in an appropriately weighted manner, any prior information about the stereochemistry of the molecular components. The function minimized in the least-squares procedure is a sum of terms of several kinds. These include terms pertaining to structure factors, bonding distances, planes, chiral centers, non-bonded contacts, thermal factors and occupancy factors. Only those matrix elements that are non-zero for the stereochemical restraint contributions are retained in the normal equations. The conjugate gradient method is then used to solve this system of linear equations.

The procedure is rapidly convergent and relatively economical. It has been applied in this laboratory to carp calcium-binding protein, snake venom neurotoxin, lamprey hemoglobin, myohemerythrin and yeast phenylalanyl-tRNA. The procedure has also been implemented in a number of other laboratories for refinements of other structures.

BILDER: A MANIPULATIVE GRAPHICS PROGRAM FOR BIOPOLYMERS. R. Diamond, MRC Lab of Molecular Biology, Hills Rd., Cambridge, England.

This program has been developed on a PDP 11/50 with 48K memory, RP04 disc and E&S Picture System I with tablet.

The program is intended as an aid to all those operations which have to be performed before the first cycle of coordinate refinement, or between cycles. It can develop a molecular list ab initio, or modify an existing one and can do so with or without the maintenance of chain continuity. Modifications possible include sequence changes (including those involving deletions and insertions) and still produce a continuous resulting chain. The independent variables are dihedral angles plus a sub-set of inter-bond angles, the coordinates being treated throughout as dependent variables. Hence bond-lengths are conserved at the values generated in building or at the input values in the case of a list being modified. The angular variables may be controlled explicitly by engaging up to three angles at a time, in any combination, and turning them and either accepting or not accepting the result. (The core coordinates are not affected until the accept command is executed) Alternatively, the shape may be controlled by specifying some combination of target positions for atoms and/or target values for angles, these targets being then data for the optimiser. The optimiser uses least squares principles and a system of nested constraints, nested to a depth of four hierarchies. The first hierarchy contains target positions for atoms at the end of the active region (molten zone) which have to be reached if chain continuity is to be maintained. Optimisation against this small amount of data generally produces a rank-deficient normal matrix which is then projected down into its determinate subspace before

inversion, and the degrees of freedom available to remaining hierarchies are then expressed in terms of (n-r) orthonormal vectors of dimension n where n is the order and r the rank. All positional target data are placed in the third hierarchy (which does most of the work). Angular target information may be placed in the second, third (default) or fourth. If placed in the second they specify shape, taking absolute precedence over positional data, and this may be used to position a group subject to the puckering of a ring being as stated. Angular targets in the third hierarchy compete with positional data in a least squares sense and hence adopt elastic properties. Angular targets in the fourth hierarchy act as defaults in the event that the given positional targets fail to specify shape.

The coordinates are held in a linear list which is thus not suitable for framework structures. Portions of the chain not in cores are held in front and back spools on disc. Spooling may go either way along the chain and all the software is fully bipolar. Residues may be removed from or added to either end, main chain parameters may move structure on either side of themselves, and so on.

The electron density map is pre-contoured and divided into bricks, usually about 5A cubes. Each brick may contain up to 20 elements which are independently switchable of display time. Typically, one element comprises contours at one level on one set of planes, but may consist of any fixed vectorial information and they may have various sources. Thus a map and a difference map may be made available simultaneously.

The program has developed an ab initio interpretation of the TMV protein in Cambridge, and a group from Uppsala has used it to interpret their map of thioredoxin.

The program was developed under DOS, but it has recently been installed at NIH Bethesda, Md., under RSX II M on an 11/70 using a Picture System II, although, at the time of writing, this implementation is more limited in scope and speed than the original.

FEASIBILITY AND USEFULNESS OF AN EFFICIENT, GENERAL, DISTRIBUTABLE ELECTRON DENSITY FITTING PROGRAM FOR CHEAP GRAPHICS TERMINALS. J. C. Hanson, Biology Dept. Brookhaven National Laboratory, Upton, NY 11973.

Several groups now have storage tube terminals which they use to monitor the progress of protein refinement and for simple modifications to the structure with respect to the density function. An example is the program described by Hanson, et. al [1] which was written at Johns Hopkins University. It was written for a PDP 11/40 with 2 disks, a storage tube display and a hard copy device. I feel that there is a need for a new program which could be more readily distributed, was more general, was easier to use , and made more efficient use of direct access disk. This program would provide wider access to graphics as an aid to refinement. However, it could not replace the more elaborate systems where a large amount of refitting was necessary.

---

1 J. C. Hanson, W. M. Ringle, B. C. Wishner, L. C. Andrews & W. E. Love, ACA Abstracts, Aug, 1977.

A SUGGESTION FOR THE ROLE OF NRCC IN PROVIDING A GENERAL, TRANSPORTABLE GRAPHICS CAPABILITY TO THE CHEMISTRY COMMUNITY. Wm. Johnston, LBL.

NRCC could provide a general purpose, transportable, device independant, scientific graphics system to the chemistry community. Such a system would provide general scientific graphics capibility in the area of two and three dimensional line graphs, bar charts, polar graphs, general coordinate transformation capibility, display of two and three dimensional scaler fields as contour plots, grey level plots and surface plots, two and three dimensional vector fields and complex surface display. This would run at LBL and could provide the basis of a color movie making system. Such a system would interface to the graphic device drivers through a SIGGRAF standard interface, and would therefore be capable of operating a wide variety of non-interactive and interactive devices.

In the large, this system, already exists at LBL and is called Integrated Data Display System (IDDS). NRCC could support future development in areas of specific interest to chemists and assist in completing the task of transforming the existing coding to a machine independent FORTRAN. The design philosophy of this system, which is a subroutine library, has included the concepts that the user should have to perform a minimum interaction with the system in order to get "first try" results, but that the potential for a great deal of control exists for "custom" work. It is believed that the user should be provided with many different types of graphic display techniques with which to examine his/her data. A number of graphic related utilities are provided for data transformation, smoothing, interpolation and extrapolation. Finally the design philosophy includes the concept that the code should be readily accessible to the user and that it be well documented internally so that the user may adapt, modify and specialize the coding as he wishes.

A READING LIST OF BOOKS ON COMPUTER GRAPHICS, PROGRAMMING STYLE AND STRUCTURED PROGRAMMING. Robert Langridge, Computer Graphics Laboratory, Dept. of Pharmaceutical Chemistry, University of California, San Francisco, Calif. 94143.

In response to several requests made after my invited lecture on computer graphics at the NRCC workshop on Tuesday, June 20th, this is a list of useful books on computer graphics, programming style and structured programming. All of these books have the additional advantage of being very well written. These are personal choices which I have used as recommended texts for several courses on computer programming and computer graphics, but others may make different selections. All are quoted from memory (mine, not the computer's) so neither I nor the NRCC can accept responsibility for errors and omissions.

Computer Graphics

An Introduction to Interactive Computer Graphics, W.M. Newman & R.F. Sproull, McGraw Hill, N.Y., 1973 $14.95

By far the best presently available text on computer graphics. Should be even better with the new edition promised in July 1978. Other books are in the works, but even from the proofs I have seen, Newman & Sproull still seems the choice.

Programming Style

B. Kernighan & P.J. Plauger, The Elements of Programming Style, Addison Wesley, 1975 $5.95

Modelled on Strunk & White's "little book" "The Elements of Style", this is a stylishly written book about programming style. The examples are in FORTRAN or PL/1, & in spite of this (or perhaps because of it?) illustrate many practical points, which, if observed, will make your code not only easier to read, but easier to debug.

B. Kernighan & P.J. Plauger, Software Tools, Addison Wesley, 1976 $8.95

This is a very practical book on how to develop software tools, which although illustrated with examples from text editing, searching, sorting & the like, provides standards which are applicable to any software development. RATFOR (Rational Fortran) is used throughout, the text was set directly from the computer (using the Bell Labs UNIX system);a tape of the programs as they appear in the book is available (for $25.00 I believe)

Structured Programming

An Introduction to Structured Programming, N. Wirth, Prentice-Hall, 1973 $12.95; Data Structures + Algorithms = Programs, N. Wirth, Prentice-Hall, 1976 $14.95

Wirth, from Zurich, was one of the first proponents of structured programming, designed PASCAL, a language which encourages structured programming, & writes English better than most of us.

Structured Programming, O-J Dahl, A Dykstra & C.A.R. Hoare, ? 1969

Read the section by Dykstra, who set the discussion of structured programming going with a publication in The Communications of the ACM in 1968

A POSITION ABOUT NRCC AND COMPUTER GRAPHICS HARD[illegible], D. L. Vickers, Computer Graphics Group, Lawrence Livermore Laboratory, P.O. Box 808, Livermore, California 94550

Computer graphics is still rather volatile, especially graphics hardware, and it is therefore somewhat presumptuous to make firm recommendations as to what hardware crystallographers and physical scientists should get for future use. Nevertheless, I see three main areas which should be given serious attention by NRCC:

1. Film Recorder. There should be a single, high quality, color film recorder available to the national community for the purpose of producing movies, slides, and publishable quality prints. It should accept 7- and 9-track tapes, floppy disks, and cartridge disks with data in either vector or raster format and should produce black and white and color 16 mm and 35 mm film and 4" x 5" color negatives or positives. A standard minicomputer is needed to drive the recorder; film processors would also be nice, though this could be done commercially at first.

2. Highly Interactive System. There should be some small number, say 6, very highly interactive systems at strategic sites throughout the nation. These should be available, by appointment to those who need to really visualize and interact with their data. They should (initially) be calligraphic displays with a 3-D pointing method of interacting with the picture and with the the ability to produce a real-time stereoscopic picture. Color displays would be nice. Output should also include tapes and/or disks necessary to be compatible with the film recorder.

3. Moderately Interactive System. These should be ubiquitous, inexpensive, off-the-shelf, systems such that every serious scientist would have one close at hand. For now they might be based on storage tube displays, though I feel a technology more inducive to future expansion would be video. As frame buffers continue to drop in price, video will also support color at little extra cost. These small systems might look like a "personal" or "home" computer with a keyboard and floppy disk I/O. Where several scientists are working together, several small systems may be tied together to share larger resources. NRCC could develop a prototype with software and encourage individuals to purchase their own hardware.

STATUS AND PLANS FOR MOLECULAR GRAPHICS, W.V. WRIGHT, Department of Computer Science, University of North Carolina, Chapel Hill, North Carolina, 27514

The GRIP-75 molecular graphics system, developed by the Department of Computer Science at the University of North Carolina, Chapel Hill, has been used by 19 groups of chemists, biochemists, and crystallographers from 13 institutions. Results obtained in part with this system have appeared in 12 papers in the chemical literature.

This resource includes a team of computer scientists who are informed about macromolecule structure computations and dedicated to collaborating with chemists in developing tools and techniques.

We propose to develop this disciplinary resource further with three main objectives:

1. To continually develop a high-powered computer graphic system specialized for molecular structure studies and using the most advanced techniques available. We see this as a trailblazer system for the discipline, where new methods and ideas can be quickly tested for feasibility and exposed to many chemists in pilot use. The most fruitful techniques can then be built into smaller systems.

   A research effort will study the seeing and manipulation of 3-D images.

   A second research effort will seek to improve algorithms for calculating densities, contours,

idealized geometries, collisions, energies, and refinements.

2. To provide a consistently available service center where chemists may come to solve molecular structure problems. We expect this to serve both chemists who do not have a molecular graphics system and those who need more power for a particular problem than their own system provides.
3. To develop machine-independent, exportable, and configurable software, of product quality in construction and documentation, which others can readily install on their own systems.

THE MMS-X SYSTEM. C. D. Barry, Computer Systems Laboratory, Washington University, St Louis.

MMS-X is a relatively inexpensive and replicable computer graphics system tailored to the task of displayed macromolecules and performing interactive electron density map fitting. It is a combination of a Texas Instrument 980B minicomputer with 48K words of memory and a two mega-word cartridge disc unit and special hardware designed and fabricated by the Computer Systems Laboratory (CSL) to perform transformations and provide line drawing graphics. The control circuitry for the special hardware (the molecule processor unit) has been embedded in the I/O structure of the minicomputer and provides an integrated system in which the hardware operates directly on data in the computer memory. Integration at the software level is attained through use of Fortran callable subroutines which employ online assembly code to establish and execute the "parameter blocks" which provide the function codes and data specification for the special hardware operations. Man-machine interactions utilize a selection of knobs, switches, buttons and 2 joysticks which form a molecule control panel as well as a Beehive terminal which supports a question and answer style of interaction and provides alpha numeric read out of data on a frame by frame basis.

The basic system capabilities have been planned to support 3D presentations of a typical protein (5-10 thousand short vectors per frame) and to support the 3D display and real time interactive fitting of electron density maps on a fragment by fragment basis; each fragment represents 50 to 100 atoms--the density being displayed as contours on 3 orthogonal sets of planes. (Chicken wire or basket weave) while the model manipulation supports simultaneous variation of up to six variables. However, the system has been designed to facilitate expansion of both the hardware and the software in response to the emerging needs of the user community, hardware extension being accomplished by use of macromodules (digital computer modules also designed and fabricated by CSL) which can be used to build special processing units for integration in the same manner as the coordinate transformer and display units. Likewise, while the design is for a system with stand-alone capabilities, loose coupling to a host computer is feasible and is being implemented for the PDP11 family as well as a couple of main frame machines (Amdahl and CDC 6500).

Productive research in science requires the development and use of tools specialized to the needs of particular research groups. MMS-X is an example of such a tool for use by the crystallographic community providing a high-performance graphic display tightly coupled to a modular processing structure supporting convenient and rapid man-machine interaction. Creation of a national collaborative program provides the mechanism by which such tools and methods can evolve, be evaluated and exported for use by a larger community, and through which a better understanding of research needs at the natonal level can be obtained.

By virtue of its extensible design, new processing equipment, specialized to the needs of researchers more broadly interested in molecular structure may be conveniently added to the basic MMS-X. The scientific environment that can exist with multiple groups having their own copies of the same extensible system allows a kind of collaborative development that provides benefits to all members of the community. Each group's innovations, whether in methodology, hardware or software can be shared and extended or adapted to fit the needs of others through the national collaborative research program.

The major aims of our program are to:

- Establish an active, national MMS-X community interested in the collaborative development of computer algorithms, software and hardward for research on molecular structure.

- Encourage identification and development of new uses for interactive, high-performance, line-drawing graphics.

Supplementary aims are to obtain an evaluation of our efforts in this area and to foster the development and use of modular computer systems by the scientific community.

MOLECULAR MODELING--BIOGRAPHICS LABORATORY, E. Meyer.

A modestly configured mini computer system (POP11/40, 28k words memory, RK05 disc, 9 track magtape) is linked to a 3-d (three dimensional) Vector General line drawing graphics display with rapid hardware functions (rotation, translation, scaling, intensity depth viewing). An interlocking system of computer programs has been written to provide the following functions:

1) Micromolecular modeling (program FIT): fast, in core contouring, electron density modeling as contours and chain tracing, manipulative amino acid modeling, storage and recall of molecular fragments. Data base generality permits interactive modeling of substrate-receptorsite interactions (cf. Fig. 1,2) with the same program.

2) Structural retrieval and display: Input of coordinates of small structures or micromolecules from the Protein Data Bank, extraction of pertinent sections of macromolecules on the basis of amino acid sequence or proximity to a target group (e.g., a 10A scoop about HIS 57).

3) Interactive display (program PIXMOV): simultaneous intensity control of up to 10 picture files; global rotation, translation, scaling, left and right or top and bottom stereo, interactive control of two 3-d pointers. The time ordered sequence of interactive display commands may be recorded digitally and played back for previewing picture sequences. Acceptable sequences may be committed to film (slides, movies) with the computer controlling the frame count for proper exposure.

4) System design features: database generality permits the input and manipulation of any chosen sequence of atoms (i.e., prosthetic groups, cofactors), modest hardware configuration and high level programming language make the programs readily exportable.

5) Results: two macromolecules have been fit in production mode: the monoclinic form of hen-egg-white lysozyme (2.5Å) and stoph. nuclease (~1.5Å).

Fig. 1. Backbone (C,N,O) atoms of alcohol dehydnegenare (ADH, Branden, et al.) with the substrate molecule, cyclohexanol, docked in the highlighted active site cavity.

Fig. 2. Zoom stereo view with identical orientation to Fig. 1. NAD^+ and cyclohexanol are positioned relative to the central Zn ion.

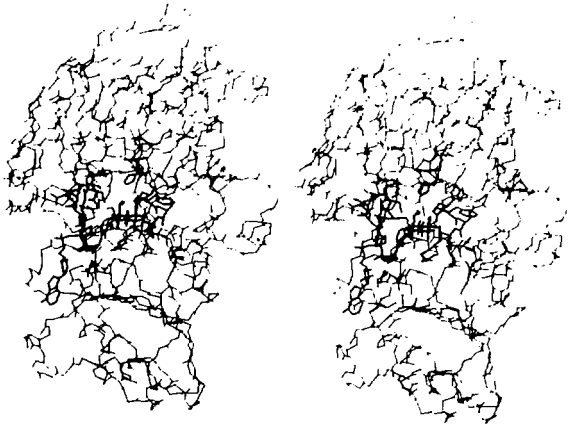


Fig. 1. Backbone (C,N.O) atoms of alcohol dehydnegenare (ADH, Eranden, et al.) with the substrate molecule, cyclohexanol, docked in the highlighted active site cavity.

E. Meyer, Biophysics Lab
Texas A & M

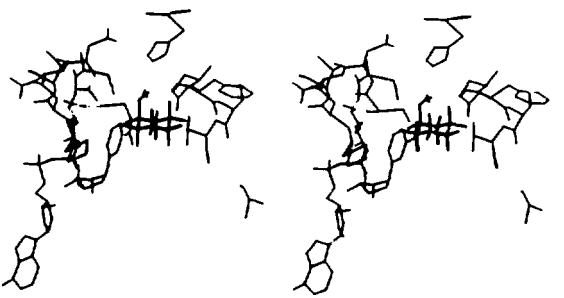


Fig. 2. Zoom stereo view with identical orientation to Fig. 1. NAD^+ and cyclohexanol are positioned relative to the central Zn ion.

E. Meyer, Biophysics Lab
Texas A & M

DESCRIPTION OF THE JCPDS POWER DIFFRACTION FILE. Gerald G. Johnson, Jr.

The JCPDS has been collecting, reviewing, publishing and distributing various indexes of powder diffraction data from the open literature and from research supported by JCPDS Grants. The primary purpose of this data is the identification and characterization of a particular cyrstallographic phase. A secondary and lesser used purpose, is a bibliographic file on the method of preparation or other informational material. The primary produced data is on specially formated 3 x 5 inch cards for human viewing. The data base (with softbase) in magnetic tape format has been available since 1966 and consists of approximately 16.8 M bytes comprising chemical (and mineral) names, chemical formula, alloy phase designation, a quality of data indicator, and all the different data of the interplanar spacings relative intensities for that phase. In the Fall of 1978 the JCPDS file will be a member of CIS and the public can use the JCPDS file to identify an unknown crystallenephase via on-line interactive computing.

Distribution of the data base in book and microfiche format is also available. Special subfiles have been produced--(1) the metals and alloys (~5000 phases), (2) frequently encountered materials (~2500 phases), (3) mineral subfile (~3000 phases), (4) NBS produced subfile (~1000 phases) and (5) organic subfiles (~9000 phases).

The inorganic file consists of approximately 2400 phases. This file (both books and magnetic tape) is updated by the editorial and scientific staff with 2000 new phases added yearly.

Currently the JCPDS has distributed their data base (tapes) plus appropriate search/match software to many sites. It is common for industrial analytical laboratories to identify an unknown crystaline sample as follows: prepare powder→ record powder pattern→use search/match procedures with JCPDS powder data base.

After approximately 5 years usage by the user community (mostly industrial scientists rather than trained crystallographers) the data is re-reviewed to correct any errors found in the initial publication.

The file is currently being extensively computer verified for correlation between patterns, and improvement in the quality of the data.* A future project for the JCPDS is to have available on magnetic tape additional information (crystallographic, bibliographic, optical, and additional chemical preparation data).

---

* New file data is checked by combining single crystal and powder data evaluation programs (i.e., the same program 'AIDS' is used to evaluate data entering into the JCPDS-powder file and the NBS single crystal files).

THE CAMBRIDGE CRYSTALLOGRAPHIC DATABASE: SEARCH, RETRIEVAL AND DISPLAY. Frank H. Allen, Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England.

The Database currently holds information on 20,483 crystal structures of compounds which contain 'organic' carbon: Bibliographic (BIB, 3.5 Mbytes), Chemical Connectivity (CONN, 4.5 Mb), Evaluated Numeric Data (DATA, 20 Mb). The Database is available worldwide through 12 National Affiliated Centres. Search queries are normally posed to the BIB and CONN files (e.g searches for chemical name or formula fragments; structural or substructural fragments). The results of searches (hits) may be then used to retrieve lists of BIB citations and /or the relevant DATA. The latter is used to generate molecular geometry tables and a range of illustrations. A variety of Search, Retrieval and Display (SRD) systems have been successfully applied to the Database. The Cambridge System operates in batch mode, using sequential search techniques on BIB and CONN. Some 30-40 separate queries may be processed in one pass. Facilities are available for the systematic tabulation of geometry and for the preparation of a wide range of mono, stereo and 'space-filling' illustrations. The system has been implemented on IBM, CDC, PDP, Univac, Hitachi and Burroughs machines. The NIH Interactive System (CSSR in UK) operates on hierarchically organised inverted files, is fully interactive and ideally suited to networking (via PDP host). Response is faster than the Cambridge System but geometry facilities less extensive. The French DARC System provides interactive CONN searching and high quality chemical diagrams. The Japanese TOOL-IR System is operational as are variants of the Cambridge System in Italy and New Zealand. It is planned to speed up the Cambridge SRD system by the introduction of BIB and CONN screen records.

THE METAL DATA FILE. L.D. Calvert, National Research Council of Canada, Ottawa, K1A 0R9.

The Metal Data File (MDF) is a private file in the Chemistry Division of the National Research Council of Canada (N.R.C.C.!). It contains the relevant data for the ~4000 intermetallic compounds with structures which have been refined in the period 1913-1976; 1977 data are being compiled and processed. The file is resident on a PDP-8e and is operated by programs in Fortran written by A.C. Larson, E.J. Gabe, Y. Wang and F.L. Lee. It is based on data originally compiled by A.C. Larson and D.T. Cromer at Los Alamos, W.B. Pearson at Waterloo and L.D. Calvert in Ottawa. The file structure is similar to that of the Cambridge Organic Crystal Data File. Emphasis has been placed on data integrity and evaluation. Input is kept simple by using the general space group interpreting routine of A.C. Larson (Acta Cryst. (1969), A25, 51) which operates directly on the Hermann-Mauguin symbol and free format numerical data input, which is later formatted under program control. Data are checked by diagnostic programs which check for transcription errors and for data inconsistencies. The formula is computed from the structure and compared with that given; the cell volume is compared with that expected for metallic cubic close packing; the density and volume per atom are also calculated. The major check is by interatomic distances which are selected by the criterion of Frank and Kasper (Acta Cryst. (1958) 11, 184), with appropriate allowance for atomic radii and cell volumes, and then calculated and compared with the sum of atomic radii as well as those given; unusual values are flagged. Binary substitution can be handled in one normal ATOM card but ternary substitution requires extra cards.

The MDF is used to check the data appearing in the Metals section of Structure Reports and also to calculate peak intensity (i.e. full profile) powder patterns for the P.D.F. of the J.C.P.D.S. In addition the relevant current data on metals is sent to Crystal Data published by NBS/JCPDS. It is planned that retrospective data on metals will be sent to Crystal Data and patterns now missing from the P.D.F. or requiring checking will be calculated from older data.

STRUCTURAL AND CRYSTALLOGRAPHIC DATA FOR BIOLOGICAL MACROMOLECULES. G. J. B. Williams, F. C. Bernstein and T. F. Koetzle, Department of Chemistry, Brookhaven National Laboratory, Upton, New York 11973.

The Protein Data Bank[1] is the internationally recognized repository and distribution center for the results of crystallographic studies on biopolymers. Two categories of information are available from the Bank; (i) atomic coordinate sets in a uniform format and (ii) structure factor data and other non-standard entries. The first category, which is perhaps of the most interest to the scientific community, is made up of about 100 entries comprising some 10^7 characters of information. The Bank depends upon depositions by scientists active in the field of macromolecule crystallography, and in most cases, these data are not available in the open literature.

After consultation with depositors and other investigators in the field the Bank has developed a file structure that allows extensive general commentary as well as specific footnotes to be carried with the coordinates.[2] In addition, many entries for protein structures contain, in a structured computer-sensible style, descriptions of the secondary structures as provided by the depositors. Because many of the actual coordinate sets held by the Bank are neither published per se nor explicitly cited in the literature the inclusion of descriptive information with the coordinates is of the utmost importance. We believe that this feature is responsible for the widespread acceptance of the Bank by the originators of the structural information.

The facility with which retrieval systems are commonly used for the extraction and manipulation of numeric information from many data bases suggests that on-line modes of interaction with the Protein Data Bank may soon become important. In this case however special care must be taken to represent the depositors' work faithfully. No single representation (numeric, textual or graphic) is entirely adequate and it would seem that unless an interactive system explicitly recognizes the need for an integrated presentation, there may be no effective substitute for a detailed study of the original entry.

Copies of the Protein Data Bank's files are available on magnetic tape from BNL, Cambridge (U.K.), Melbourne and Tokyo. Collection, standardization and entry of data into the Bank are carried out at BNL with support from the National Science Foundation (Grant PCM75-18956).

---

[1]The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, published in J. Mol. Biol. 112, 535 (1977) and reprinted in Eur. J. Biochem. 80, 319 (1977) and Arch. Biochem. Biophys. 185, 584 (1978).

[2]Protein Data Bank File Record Formats (Feb. 78). Copies are available from the authors.

NBS CRYSTAL DATA FILE. Alan Mighell

The NBS Crystal Data File contains approximately 50,000 entries. The file covers the inorganic literature to 1970, and organic literature through 1974. The data are compiled and critically evaluated in collaboration with the Cambridge Data Centre, the Metals Data Center, and individual scientists in selected disciplines. New entries are being added to the data base at the rate of approximately 6,000/year and a concerted effort is being made to bring the file up to date.

A primary use for the data is in the identification and characterization of materials. Thus, for a typical entry, the file contains bibliographic crystal and chemical data such as the unit cell, space group, Z, density, chemical name and formula. Data in the file are linked via the crystal data "determinative number" to entries in volumes 1, 2, 3, 4 of the 3rd Edition of the Crystal Data Determinative Tables (Vols. 3 and 4 are in press), and entries in the Crystal Data Space Group Tables. Key data elements from the file are being incorporated into the NIH/EPA Chemical Information System and, therefore, the data will soon be available to the public by on-line interactive computing. Also in the Fall of 1978, it is planned to issue (tape or printed publication) and new product called the Crystallochemical Identification Tables. In these tables all materials are classified by reduced cell so that an unknown substance can be simply identified by the sequence: any primitive cell of the lattice → reduced cell → table look-up.

BIDICS. I. D. Brown

BIDICS - Bond Index to the Determinations of Inorganic Crystal Structures - is a bibliographic retreival file based on the bond types present in a crystal. It is prepared by a direct search of about 60 journals and covers all crystal structures except those that contain only pure organic compounds. The computer file is used primarly for printing the annual bibliography but it has also been designed for computer searching. At present the file exists only in a archival format and no search programs have yet been written. It currently contains entries for 14,000 papers covering the years 1969 - 1977. Computer based searches could be mde, inter alia, on bond types, accuracy, radiation used, mineral names, authors' names and title. Some CAS numbers are also included. Queries (including proposals for implementing the computer file) should be addressed to I.D. Brown,

INORGANIC CRYSTAL STRUCTURE DATABASE. I. David Brown

In 1977 three groups (Bergerhoff, University of Bonn; Dittmar, Zintl T.H. Darmstadt and Brown, McMaster University, Hamilton, Canada) started work on an inorganic crystal structure database similar in structure, but complementary in content, to the Cambridge and Metals Data Files. Initially we are concentrating on searching the literature, inputing and checking the crystal structure data published since January, 1977. We expect each year to cover the 500-600 fully reported structures that are not included either in MDF or the Cambridge File. Eventually we intend to merge ICSD with BIDICS to produce a database containing both bibliographic and structural data but initially the two files will be kept separage. The file will become publically available when it has reached a useful size. Queries may be addressed to any of the contributing laboratories.

☆ GPO 690-762/177