DL/SCI/R34

DATA COLLECTION AND PROCESSING

Proceedings of the CCP4 Study Weekend,
29-30 January 1993

Compiled by
L. Sawyer, N. Isaacs and S. Bailey

SERC

DARESBURY LABORATORY
Daresbury, Warrington WA4 4AD

ISSN 0144-5677

DATA COLLECTION AND PROCESSING

Proceedings of the CCP4 Study Weekend
29-30 January 1993

Compiled by
Lindsay Sawyer, University of Edinburgh
Neil Isaacs, University of Glasgow
and
Susan Bailey, Daresbury Laboratory

SERC

DARESBURY LABORATORY
1993

CONTENTS

INTRODUCTION

The topic of this year's CCP4 Study Weekend, Data Collection and Processing, has not been covered in recent years and it was felt that the subject was well worth reviewing. The reasons for this are first that the need for the highest possible quality diffraction data cannot be overemphasised. Second, detector technology has undergone significant evolution in recent years with the advent of the imaging plate. Third, the strategies available, particularly with the successful use of the Laue method and the advantages of crystal cooling have changed the time taken to collect data and the number of crystals required so to do. Finally, better software has been developed which leads to the extraction of the maximum amount of information from the diffraction pattern.

We took the view that advances in protein purification and its assessment together with protein crystallisation, without which the crystallographer does not even pass go, should constitute one session not only for completeness but also for their intrinsic value. Low temperature methodology and optimising the available X-ray flux were also considered essential topics. The remainder of the meeting was designed to cover all of the important issues concerning data collection strategy, the factors which affect its quality and the methods (and programs) for extracting structure factors and their standard deviations from the raw observations.

Owing to a misunderstanding, one of the papers from last year's meeting on molecular replacement was omitted from the 1992 Proceedings and is thus included in this volume.

The meeting was organised and supported by the SERC Collaborative Computational Project in Protein Crystallography (CCP4). We wish to thank the invited speakers for their presentations and for their cooperation in providing manuscripts for these proceedings. We thank the Daresbury Laboratory and its Director, Professor A.J. Leadbetter, for the provision of organisational help and support; and in particular Pauline Shallcross, Val Matthews, Steve Quinn, Tony Buckley and Stuart Eyres who ensured that the meeting ran smoothly. In addition the proceedings owe much to the efforts of Mel Davies and his staff.

Lindsay Sawyer
Neil Isaacs
Sue Bailey

September 1993

Protein Purification and Analysis for Crystallographic Studies

by

David R. Thatcher

Zeneca Pharmaceuticals, Alderley Park, Macclesfield, SK10 4TG

A) Introduction

The use of purified protein preparations of extremely high levels of purity is a key prerequisite for the acquisition of precise and unambiguous data in structural analysis. The determination of these purity levels usually involves not only the analysis of heterogeneity [removal of >95% of the impurities derived from the original biomass], but also and perhaps more importantly for crystallographic studies the analysis of micro-heterogeneity [the removal of forms of the protein itself which have been either incorrectly modified during synthesis or degraded during isolation]. The accumulation of such defective molecules in the crystal lattice will lead to reduced diffraction and could be the cause of premature cessation of crystal growth (discussed in two recent reviews on the preparation of protein crystals for diffraction studies [1-2]).

This paper will focus on methods for monitoring such micro-heterogeneity.

(B) How Pure is Pure?

Purity is a relative term whose value is related to a particular application. In protein crystallography the relationship is not always clear although in general purer protein preparations give better crystals and there is an impetus to strive for higher and higher levels of purity. Recent advances in chemical analysis have outpaced developments in preparative separation technology. Impurities can be detected in the most highly purified of preparations and claims of homogeneity only indicate the inadequacy of the analytical techniques employed. The question therefore is not "How pure is a preparation ?" but "What are the levels and pattern of (micro)heterogeneity present ?", "How do theses vary with the purification methods employed ?" and "Can the same spectrum of impurities be reproduced as each new batch of material is prepared ?". In order to reduce the risk of failure at a later stage, it is crucial to employ a range of analytical techniques to probe purity and not, as has been common practice in the past, to rely solely on one or two techniques such as silver stained SDS-PAGE and reverse phase-hplc. These methods, although extremely sensitive to gross heterogeneity, provide little information on micro heterogeneity.

B.1. Detection of Contamination Derived Through Incomplete Removal of Biomass Constituents

The level of contamination of proteins and other macromolecules derived from the production system can be determined by a combination of well known and tested techniques (3-5) : silver stained SDS-PAGE (for proteinaceous contaminants which differ from the product by >1% in molecular weight), isoelectric focusing (for contaminants which differ in charge) and reverse phase hplc (for contaminants which differ in surface hydrophobicity).

B.2. Detection of Contaminants Derived Through Modification of Protein Structure.

In a purified preparation, the formation of varying amounts of chemically distinct species of the bulk of the protein present is due to :

(a) incomplete or inappropriate post translational modification during initial synthesis or

(b) chemical or enzymatic action during isolation and purification.

These types of chemical modification, their causes and detection are detailed in Tables 1 and 2. Most of the structural changes involve a change in molecular weight and/or overall net charge and can be detected by the use of mass spectrometry (MS), capillary electrophoresis (CE) and high resolution peptide mapping. Although these techniques were primarily developed to reveal problems in the production of protein therapeutics, advances in instrumentation have made the technology widely available . It is likely that in the future their application in the definition of quality for structural analysis will become routine.

B.2.*(a). Detection of Species Differing in Molecular Weight*

Most micro heterogeneity is formed by species which differ in molecular weight, from 1 to several hundred daltons. Until recently only gross changes in molecular weight (>2kD) could be detected directly. Developments in electrospray ionization mass spectrometry (ESI-MS) and matrix assisted laser desorption mass spectrometry (MALDS-MS) have brought the threshold of accuracy down significantly and have propelled mass spectrometry from the status of an esoteric specialist technology to a pre-eminent tool for detailed analysis of protein covalent structure (6-9). In particular, ESI-MS allows the determination of accurate mass values (>0.001% precision and therefore much less than the mass of a single amino acid residue in proteins of up to 100k). The method is able to analyse simple mixtures, resolving impurity series down to approx. 10% of the major species for a large protein. ESI-MS therefore provides evidence for the presence of proteolytic degradation, covalent modification through incomplete glycosylation or acylation and in addition detects mass increases due to chemical modifications such as oxidation. As the DNA/protein sequence is usually known, the theoretical mass of the protein can be predicted and compared to the ESI-MS result confirming or otherwise the sequence identity of the preparation. Recently it has been shown that the proportion of each of the charged species protonated on ESI can be dependent on the conformational state of the protein. ESI-MS can therefore, in a well characterized system, provide information on conformational stability of a preparation (10).

MALDS-MS is a technique of less mass accuracy but for most proteins can give enough information to detect proteolysis. The technique is capable of analysing more complex mixtures than ESI-MS and has extended the mass range of proteins amenable to analysis to >1000kD and to partially purified preparations. The method is therefore useful for the analysis of high molecular weight multi-subunit proteins such as antibodies. MALDS-MS is also capable of detecting other types of macromolecule such as detergents and carbohydrate and is therefore a useful scan for adventitious impurities which are otherwise difficult to measure.

Non-covalent irreversible association or aggregation occurs at varying frequency in many protein preparations. This type of modification is usually invisible to mass spectrometric techniques and is best estimated using high performance gel permeation chromatography or light scattering (11).

B.2.(b) *Detection of Charged Variants*

Deamidation leads to a unit charge increase in mass which cannot be detected by current methods of mass spectrometry. The generation of a negative charge however affects the overall charge of the molecule and may along with other types of chemical modification be detected by electrophoretic techniques. Control of deamidation is important for the structural chemist, as the reaction itself can result in isomerization (to isoaspartyl), racemization and peptide bond cleavage (12). The actual range of products formed is somewhat constrained by the conformation of the folded protein. In the case of proteins prepared by solubilizing and denaturing *E.coli* inclusion bodies, attention must be paid to deamidation levels in the final product. These processes involve opportunities for abnormal deamidation processes to occur prior to the completion of the refolding operation and before the tertiary structural constraints to such deamidation have been re-imposed.

In the past, charge variation has been most conveniently detected by isoelectric focusing. This method has many drawbacks : sensitivity is limited, the gels are difficult to quantitate and the method uses low molecular weight ampholytes to form the pH gradients. These ampholytes can bind to protein, modify their net charge and induce serious artefacts.

Recently capillary electrophoresis (CE) methods have become available. In CE a fused silica capillary is filled with electrolyte support buffer and placed between two buffer reservoirs containing high voltage electrodes. Proteins introduced at one end of the capillary migrate under the influence of the electric field to the other end of the column. The proteins only interact with the buffer and the electric field and the efficiency of resolution is theoretically only dependent on the magnitude of the field applied. As the capillaries are effective in dissipating the heat generated, high voltages can be applied and separation efficiencies of $>10^5$ theoretical plates can be achieved allowing the separation of species differing by ion mobilities of $< 10^{-6}cm^2/V*s$. In practice protein separations never reach such levels and efficiency is limited by peak tailing and adsorption to the silica capillary. For each particular protein a number of options are available for reducing these effects (13-14). However the method is capable of cleanly separating and quantitating single deamidation events in large proteins.

B.2.(c) *Detection of Modifications Which Do Not Result in Either Significant Changes in Molecular Weight or in Overall Net Charge*

Modifications such as the formation of mismatched disulphide bonding can be located by high resolution peptide mapping. Modern hplc instrumentation coupled with carefully controlled digestion conditions and high purity reagents have enabled the reproduction of precise peptide maps. The level of mismatched disulphide can be quantitated from the novel peptides produced. When a free thiol group is also present there is a greater chance of mismatched disulphide formation and in this case alkylation in combination with high resolution peptide mapping can be used to quantitate the level of mismatch (15).

This approach is also used to locate deamidation, phosphorylation etc. events after the separation and semi preparative isolation of the species involved..

B.2.(d) *Hyphenated Mass Spectrometric Techniques*

The ESI-MS method generates a series of multiple charged ions for every molecular protein species present. Although capable of resolving one or two series in a single sample, ESI-MS cannot distinguish and characterize several related species at low levels of contamination. As the samples are injected into the source in a liquid state in ESI-MS, the method is well suited to coupling on line to hplc and CE systems. Both configurations have been successfully exploited and in the case of CE/ESI-MS attomole levels of protein have been characterized (16). The sites and pattern of glycosylation of large molecules such as tissue plasminogen activator can be determined in a single experiment (17). It is theoretically now possible to gather high level information on the covalent structure of protein in a single micro crystal or fragment of crystal or in the contents of a single hanging drop.

B.3. Glycosylation : A Special Case

As a post translational modification, glycosylation is unique in that high levels of micro-heterogeneity are normal. Moreover the pattern of heterogeneity is possibly a functional significance. This inherent glycosylation based micro-heterogeneity could in certain cases pose a problem for the crystallographer and can be avoided by either genetically engineering out glycosylation sites, selecting a procaryotic production system or by enzymatically removing glycan groups during preparation. However in several instances the deglycosylated protein species have poor solubility and stability and production of carbohydrate free glycoproteins cannot be considered as a generic solution to the problem. It may be more prudent to work with the glycoprotein and in that case it may be necessary to establish the overall pattern of glycosylation in order to maximize the chances of obtaining consistent behaviour in crystallization trials. The glycosylation pattern should be characteristic of the type of glycoprotein under investigation (particularly important when certain recombinant systems are employed) and be similar from batch to batch.
Recent developments in instrumentation have made glycan analysis amenable to the non specialist laboratory. The glycan side chains of both N and O linked glycoproteins may be quantitatively released and automatically purified by controlled hydrazinolysis and ion exchange chromatography using the Oxford Glycosystems GlycoPrep (18). The carbohydrate fraction may then be analysed directly by size exclusion chromatography (using the OGS GlycoMap and by high pH ion exchange hplc (19), using a Dionex chromatograph. The identity of the oligosaccharides may be inferred from accurate mass data obtained by MALDS-MS or by fast atom bombardment mass spectrometry after conjugation to an amino benzoyl octyl ester function and enrichment by reverse phase hplc (20). These methods will in combination provide a detailed fingerprint of the level and extent of glycan heterogeneity.

When a fungal host such as yeast is used to produce a recombinant mammalian protein, any N-glycosylation sites will be modified with a large abnormal mannan group. Such large carbohydrate groups are highly heterogeneous and will have a marked effect on the physical properties of the molecule and may have to be engineered out. Even if no N-glycosylation sites are present in the sequence, low levels of O-glycosylated species may be formed and should be monitored in the final preparation.

C Choice of Production System

Many problems in crystallization were traditionally avoided by switching the source of protein to that of another but closely related species. The constraints of rational drug design make this option unattractive in the pharmaceutical industry where data on the human protein is of primary interest. Advances in genetic engineering have made it possible to express human proteins in a variety of recombinant organisms and cell lines. Each of these

systems has its own advantages although each system also has the potential for developing its own specific micro-heterogeneity.

The choice of host cell is largely dictated by the complexity of the protein under study and cost (Figure 1). Continued failure to obtain adequate crystals from one strain despite high levels of purity, may necessitate the evaluation of protein obtained from other recombinant sources. From the outset and if resources allow, the analysis of protein derived from at least two distinct recombinant sources should be attempted .

D. : Choice of Recovery and Purification Strategy

The methods of protein purification have been extensively reviewed elsewhere (21-23). For most soluble proteins at reasonable expression levels, a combination of the methods of ion exchange, hydrophobic and gel permeation chromatography are adequate for the removal of >95% host cell impurities. Membrane proteins and other proteins present in trace quantities usually require enrichment by immunoaffinity chromatography. Immunoaffinity and other affinity systems which exploit ligand binding, although the most highly selective separation steps available for purification, are relatively inefficient at reducing protein micro-heterogeneity. These steps should not therefore be used in isolation and not as the last step in a purification sequence.

There are two emerging trends in protein purification both of which are directed at reducing micro heterogeneity:

(a) Reduction in the opportunity for modification to occur after synthesis by direct recovery of the protein product from crude extracts (e.g.. by the use of fluidized bed chromatographic separations in cell cultures or homogenates).
This approach is dependent on the availability of cheap highly selective media which can operate in the presence of large amounts of non proteinaceous macromolecules such as nucleic acids and lipopolysacccharides.

(b) Application of high performance preparative hplc methods for resolution of multiple forms of a protein.
The physical properties of many modified species of protein found in purified preparations of proteins are so similar that increasing the selectivity of chromatographic media alone is unlikely to provide a preparative solution. Only the application of high performance methods which exploit small differences in properties are likely to be successful. These preparative methods should ideally be based on and developed in conjunction with high grade micro bore analytical data. Once the pattern and level of heterogeneity have been established then a rational separation strategy can be designed. Improved resolution in column chromatography can only be obtained by reduction in bead size of the matrix. At the moment this stands at 5μm bead size for 1 - 5 mg separations.

E. : Summary

There are no generic answers to problems caused by micro-heterogeneity, either in their analysis or in their solution through the application of preparative separation methods. Whether or not significant micro-heterogeneity will accumulate during the production of a protein preparation and pose a problem in crystallography is a question which is dependent

on the particular properties of the protein under investigation, the nature of the source material and the recovery and purification strategies adopted. Consequently, separation strategies need to be developed in conjunction with the analysis of micro-heterogeneity.

References

(1) P.W. Weber (1991) Protein Crystallization, Adv. Prot. Chem. 41, 1-36.

(2) S.P. Wood (1990) Purification for Crystallography *in Protein Purification: Applications*, pp 45-59, (Harris and Angal Eds), IRL Press.

(3) R.A. Oliver (1989) HPLC of Macromolecules, IRL Press

(4) Janson J.C. and Ryden, L. (1990) Electrophoresis in Gels VCH Publ., New York

(5) J.E. Shrively (1986) Methods of Protein Microcharacterization, Humana Press.

(6) M.J. Geisow (1992) Mass measurement at High Molecular Weight. Trends in Biotech. 10, 432-441.

(7) F. Hillenkamp and M. Kraus (1991) Matrix assisted laser desorption ionization mass spectrometry of Biopolymers. Anal. Chem. 63, 1193A

(8) K. Biemann (1992) Mass Spectrometry of Peptides and Proteins. Ann.Rev. Biochem. 61, 977-1010.

(9) S.A. Carr, M.E. Hemling, M.F. Bean and G.D. Roberts (1991) Integration of Mass Spectrometry into Analytical Biotechnology Anal.. Chem. 63, 2802-2824.

(10) U.A. Mirza, S.L. Cohen and B.T. Chait (1993) Heat Induced Conformational Changes in Proteins Studied by Electrospray Ionization Mass Spectrometry . Anal.Chem. 65, 1-6

(11) C.A. Schein (1991) Physical Methods And Models For The Study Of Protein Aggregation *in ACS Symposium 470 : Protein Folding* Eds. Georgiou and De Barnardez-Clark, ACS, Washington

(12) D.T.-Y. Liu (1992) Deamidation : A Source of Micro-heterogeneity in Pharmaceutical Proteins. Trends. Biotech. (10) 364-369.

(13) A.G. Ewing, R.A. Wallinford,and T.M. Olefirowicz (1989) Capillary Electrophoresis. Anal. Chem. 61, 292A

(14) B. Karger (1992) Capillary Electrophoresis. Curr. Opinions. Biotechnol. 3, 59-64

(15) A. Hitchcock and D.R.Thatcher (1993) in Protein Folding ed. R.H. Pain Oxford University Press

(16) J.W.Wahl, D.R. Goodlett, H.R. Udseth, R.D. Smith (1992) Attomole Level Capillary Electrophoresis-mass spectrometry Protein Analysis Using 5 mm i.d. Capillaries, Anal. Chem. 64, 3194

(17) V.L. Ling, A.W. Guzzetta, E.Canova-Davis, J.T. Stults, W.S. Hancock T.R. Covey and B.I. Shushan (1991) Characterisation of the Tryptic Map of Recombinant DNA Derived Tissue Plasminogen Activator by High Performance Liquid Chromatography-Electrospray Ionization Mass Spectrometry. Anal. Chem. 63, 2909-2915

(18)Oxford Glycosystems Ltd., Unit 4, Hitching Court, Abingdon, OX14 1RG, U.K.

(19) R. Reid Townsend and M. R. Hardy (1991) Analysis of Glycoprotein Oligosaccharides using High pH Anion Exchange Chromatography. Glycobiol. 1, 139-147

(20)L. Poulter and A.L. Burlingame (1990) Desorption Mass Spectrometry of Oligosaccharides Coupled with Hydrophobic Chromophores. Meth. Enzymol. 193, 661-689

(21) R. Scopes (1985) Protein Purification, Springer Verlag

(22) E.V.L. Harris and S. Angal (1989) Protein Purification Methods, I.R.L. Press

(23) S. Wheelwright (1991) Protein Purification : Design and Scale Up of Downstream Processing, Hauser.

Table 1

Origins of Microheterogeneity : Variability in Levels of Post-Translational Modification

TYPE	OCCURENCE	STRUCTURAL CHANGE	DETECTION METHODOLOGY
Incomplete removal of signal peptide	Micobial secretion systems	Molecular weight increased by approx 2kD	SDS-PAGE, MS
Incomplete removal of initiator methionine or deformylation of N-formyl methionine	Intracellular expression in E.coli	Molecular weight increased by 131D	EIS-MS, IEF, CE, High Res. Peptide Mapping
Incomplete or Mixed Disulphide Formation	E.coli Intra and extra cellular expression	Presence of inappropriately bonded thiol	Total thiol content, High Res. Peptide Mapping
Unexpected Glycosylation Patterns	Molecular Weight Distribution Changes, Charge forms may change New Glycosidic Linkages Formed	Use of Eucaryotic recombinant Host Cell Systems such as Baculovirus, Yeasts	MALDS-MS, CE, High Res.Glycan and Peptide Mapping
Incomplete or Inappropriate Acylation	Eucaryotic Cell Lines	Addition of Palmitoyl, Myristoyl, farnesyl, geranyl etc. groups	EIS-MS, High Res. Peptide Mapping
Phosphorylation	Eucaryotic Intracellular Expression	Molecular Weight Increases by 95D, More Electronegative	EIS-MS , CE High Res. Peptide Mapping

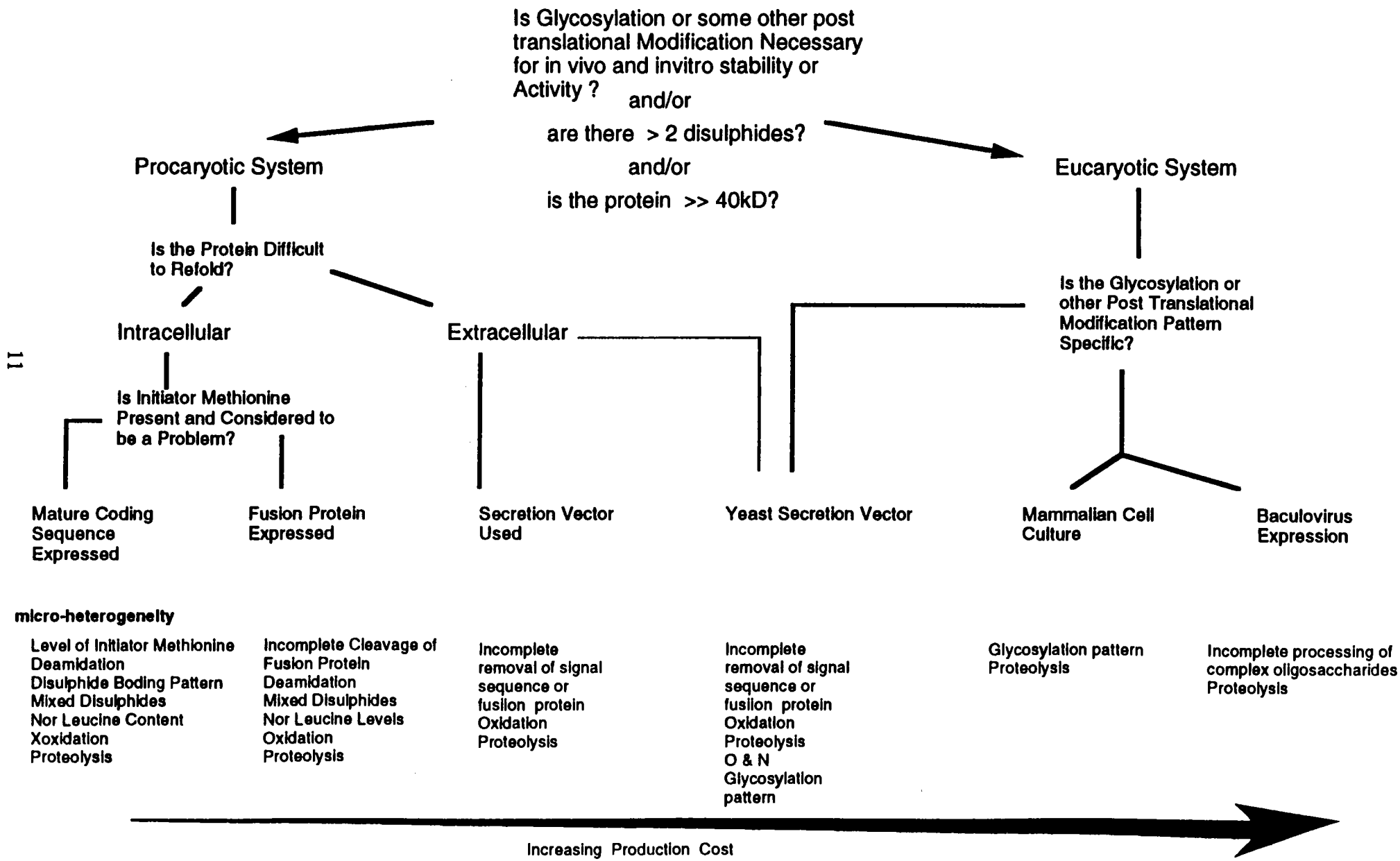
Table 2

Origins of Microheterogeneity : Covelent Modification During Processing

TYPE	OCCURENCE	STRUCTURAL CHANGE	DETECTION METHODOLOGY
Deamidation	Exposure to pH extremes, Exposure to unfolding conditions at high pH	Formation of a non native carboxyl group, isomerization, peptide bond cleavage	CE, High Res. Peptide Mapping
Proteolysis	Degradation by enzymes before high resolution purification steps	Decreases in covalent Molecular Weight	MALDS-MS, EIS-MS, SDS-PAGE (often not detected by hplc under native conditions)
Oxidation	Extracellular expression sytems, exposure of process stream to air particularly the use of metal catalysed air oxidation of disulphides	Increase in Covalent Molecular Weight	EIS-MS, Amino acid analysis (for small proteins)
Modification by Buffer Components e.g. :			
Mixed Disulphides	By products of refolding of inclusion body proteins particularly when a free thiol is present	Increase in Molecular Weight	EIS-MS, CE
Carbamylation	Use of urea buffers at high pH		

Figure 1

Decision Tree for The Selection of a Production System for a Human Recombinant Protein



NOVEL TRENDS IN PROTEIN AND NUCLEIC ACID CRYSTALLIZATION: BIOCHEMICAL AND PHYSICO-CHEMICAL ASPECTS

by

RICHARD GIEGE, ANNE THEOBALD-DIETRICH & BERNARD LORBER

UPR "Structure des Macromolécules Biologiques et Mécanismes de Reconnaissance",

Institut de Biologie Moléculaire et Cellulaire du Centre National de la Recherche Scientifique,

15 rue René Descartes, F-67084 Strasbourg-Cedex, France.

1. INTRODUCTION

The multiparametric nature of crystallization processes and the need of understanding the hierarchy of the parameters involved in the different steps of crystal growth are now well accepted facts among crystal growers of biological macromolecules [1-3]. Preparation of suitable crystals, however, often remains the limiting factor in structural biology, and some important classes of macromolecules, such as hydrophobic membrane proteins [4] and especially ribonucleic acids [3, 5], still are difficult or even reluctant to crystallize.

In this paper we discuss structural properties of macromolecules that favour their crystallization and new research trends intended to overcome failures in crystal growth experiments. We emphasize the importance of protein homogeneity and present methods valuable to evaluate this parameter. As an example we show how macromolecular impurities can impair crystallization of lysozyme. The particular requirements for nucleic acid crystallization, including the new trends for their preparation, are outlined and experiments on transfer RNAs are given. We also show how parameters such as temperature or pH can be used to control supersaturation or to uncouple nucleation and growth. Finally, we propose improved strategies for crystallization and discuss the need of instrumental developments for a better control of crystal growth.

2. CONTROLLING THE HOMOGENEITY OF MACROMOLECULAR SAMPLES

Although crystallization is often used as a method to purify molecules, obtaining monocrystals for X-ray diffraction studies in most cases requires samples of highest purity. The concept of purity (in terms of lack of impurities and of structural heterogeneities) is of particular importance in the case of biological macromolecules [6] and can become dramatic when heterogeneities concern domains of the macromolecules involved in packing contacts. In what follows, the detrimental effects of structural heterogeneity for protein crystallization, and consequently the need of molecular homogeneity, is discussed in the light of recent experimental evidences.

Sequence homogeneity of proteins and nucleic acids is best analyzed by electrophoretic methods. For proteins, isoelectricfocusing (IEF) methods are the best adapted and are sensitive to very subtle charge changes not seen by SDS-gel electrophoresis [3]. For instance in the case of aminoacyl-tRNA synthetases, IEF revealed batch dependent variability of samples due to minute proteolytic degradations which could even be revealed from proteins recovered from crystals [3, 6]. Sequence heterogeneities due to variations in post-translational modifications also can hamper crystallization [3]. Genetic engineering methods that permit to design and overproduce active cores of proteins in appropriate host cells, combined with modern purification and handling methods are presently of great usefulness to improve quality of samples and to overcome some of the above drawbacks.

Conformational homogeneity of samples is more difficult to reach. This factor has been shown to be of primary importance to obtain crystals. Light scattering measurements on

several proteins crystallizing under different salt conditions have shown that protein samples leading to crystal growth remain monodisperse even in concentrated solutions up to the supersaturation limit in the phase diagram defining the crystalline and soluble states of these proteins [7-9]; conversely when a protein is in a solvent that does not lead to crystal growth, it starts to aggregate much before reaching precipitation [7, 8] (Figure 1). This suggests that nucleation, which occurs under supersaturated conditions, is initiated by the association of homogeneous protein particles and that non-specific associations leading to early aggregation prevent nucleation. The extreme case of unspecific self-association corresponds to the strong aggregation of membrane proteins in aqueous solvents. It that case the hydrophobic effect can be overcome by addition of adequate non-ionic detergents [2, 3]. For soluble proteins, their more or less pronounced tendency to aggregate in certain solvents remains essentially unexplained and reflects our ignorance of the precise physico-chemical rules governing protein solvatation and of the effects of salts or other small molecules on this process. For instance, the fact that hen egg white lysozyme aggregates in ammonium sulfate solutions (in which it does not crystallize [7]) and remains perfectly monodisperse in NaCl solutions (in which it easily crystallizes [7]) is not yet explained. The same is true for jack bean concanavalin A where the opposite situation is observed [7].

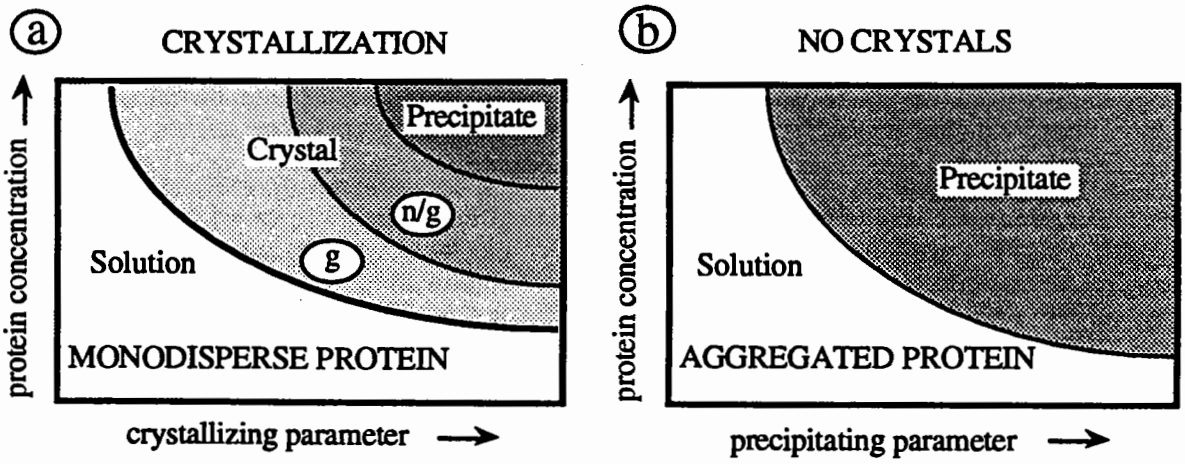


Figure 1: Schematic diagrams of the various states of proteins in pre-crystallization and supersaturated conditions as deduced from light scattering studies on model proteins [7]. The two panels represent a theoretical solution/crystal phase diagram (a) and a precipitation phase diagram (b) in which protein solubility is represented as a function of crystallizing (precipitating) parameter (it can be a salt, temperature or pH). In (a) the "crystal" domain is divided in two regions that show where nucleation (n) and growth (g) can occur. Notice that higher supersaturation is needed for nucleation than for growth, and that the higher the supersaturation, the more nuclei are formed and the faster they appear and grow. In (b), aggregates increase in size when the conditions approach the precipitating line.

Beside self-aggregation, flexibility of macromolecules is another major source of conformational heterogeneity. It can be an intrinsic property of macromolecules and is often required for functional necessity. Conformational changes arise during the functioning of enzymes and can be particularly dramatic in systems involving interactions between macromolecules. In complexes between tRNAs and aminoacyl-tRNA synthetases it has been explicitly shown that the tRNA undergoes important conformational changes [10, 11]. Thus, for macromolecules reluctant to crystallize in their free state, crystallization may be facilitated in the presence of their ligands that will freeze a biological significant conformation. An alternative solution to render macromolecules structurally more stable (or more globular) consists to prepare active minimum structures or domains. This was first exemplified by the crystallization of an active monomeric proteolytic fragment of methionyl-tRNA synthetase [12]. In present days, design of active protein fragments by genetic engineering facilitates this approach.

It is clear that the control of the conformational homogeneity of macromolecules will facilitate their crystallization. Diagnostics of concentrated protein solutions by light scattering is a convenient way to select crystallization solvents. Furthermore, addition of ligands or additives that stabilize conformations has also to be explored. However, because of the lack of enough theoretical background on protein solubility it is also advised to choose solution compositions as simple as possible, in order to avoid uncontrolled effects of buffer and solvent molecules (or their contaminants). From that point of view, if possible, it could be advisable to try crystallization assays just in water. From another point of view, use of material originating from organisms resistant to extreme conditions should lead to easier crystallizations because macromolecules from these organisms are more stable. That is actually the case for proteins or macromolecular assemblies (*e.g.* ribosome) isolated from extreme thermophiles [3].

3. THE PARTICULAR CASE OF NUCLEIC ACIDS

Most native nucleic acids, except small RNAs such as tRNAs, are not suitable to crystallization because of their great size, structural flexibility, and multi-domain nature. In the case of tRNAs, stabilization of the nucleic acid structure could be achieved by addition of spermine, a positively charged polyamine, and by magnesium ions [3, 4]. These additives were also found very useful when crystallization of synthetic polynucleotide fragments became technically feasible. For crystallization of nucleo-protein complexes, polyamine molecules are not needed. Likely, stabilization of the nucleic acids is achieved by the interacting protein molecules. Interestingly, such complexes could be crystallized with ammonium sulfate as the precipitant [3], despite the fact that salts are known to dissociate complexes between nucleic acids and proteins. Most probably the decrease of the electrostatic interactions provoked by the salt is compensated by the establishment of hydrophobic interactions favoured in the presence of high concentrations of ammonium sulfate [11].

Chemical synthesis of short nucleic acid fragments corresponding to biologically significant sequences (*e.g.* helix fragments recognized by DNA binding proteins, or containing particular structural features) has solved the preparative problem in the case of DNA [3]. The situation is much more delicate in the case of RNA molecules that present more structural potentials, are chemically much more fragile than DNAs due to the free ribose 2'-OH residues [13], and consequently are difficult to prepare in crystallization amenable forms. However different new methodological breakthrough will contribute to easier crystallization of RNAs in near future:

(i) Chemical synthesis of RNA fragments on automated DNA synthesizers is developing rapidly [14].

(ii) Macromolecular engineering methods permit to prepare RNA sequences of any structure using *in vitro* transcriptional systems and synthetic genes under the control of DNA dependent RNA polymerases [3, 15].

(iii) Structural mapping with chemical and enzymatic probes, combined with computer modeling, is a powerful method to define the structural domains within large RNA molecules [16, 17]. These domains may be prepared by the above *in vitro* synthetic procedures. Interesting systems which should be amenable to crystallization are catalytic RNAs including the active core of introns, tRNA-like domains from viral or messenger RNAs, fragments of ribosomal or viral RNAs. Of particular interest will be the structures containing pseudoknotted foldings.

We have used *in vitro* transcriptional methods to prepare $tRNA^{Asp}$ molecules from yeast for crystallization purposes. Synthetic genes under the control of the promoter of phage T7 polymerase were transcribed by the phage polymerase. The T7 system is particularly convenient because the large quantities of polymerase needed for preparative transcriptions can easily be purified in the laboratory from an overproducing recombinant bacterial strain [18]. For an optimal use of this methodology, however, several technical problems have to be considered. Transcriptions require sequences starting with 5'-G residues and yield RNA molecules with a 5'-triphosphate end, which amount can be reduced but not completely eliminated when transcriptions are conducted in the presence of high concentrations of GMP. Transcripts very often do not terminate correctly at their 3'-end and are contaminated with

molecules having one or two additional residues. They have to be separated from the transcription mixtures containing the polymerase, the template DNA, and the precursor nucleotides. This is easily done by HPLC methods; however fractionation of the heterogeneous transcript population by chromatographic methods is not feasible and can only be done by preparative PAGE on thick gels. This step is tedious and limits the preparations to rather small amounts of RNA. It may be skipped if 3'-heterogeneity of RNAs is not detrimental for crystallization. For $tRNA^{Asp}$ we have prepared several mg of pure, 3'-homogeneous, and active molecules (Table 1). These molecules do not contain the post-transcriptional modifications present in tRNAs isolated from cells and present a more flexible conformation [19]. Thus they represent interesting models for understanding the structural role of base modifications in tRNAs.

Crystallization attempts of the $tRNA^{Asp}$ transcripts in the presence of aspartyl-tRNA synthetase produced crystals of the cubic space group under conditions slightly different from those yielding the cubic crystals with wild-type modified $tRNA^{Asp}$ [20, 21] (compare conditions in Table 1). These crystals only diffract moderately (6 to 8 Å resolution). Under the conditions yielding the high diffracting complex crystals [15], the complex with the unmodified tRNA transcript crystallizes also in the cubic space group, suggesting a slightly different conformation of both types of complexes. This different behaviour is also reflected by different solubility properties of the two complexes. Experiments are underway, including micro gravity assays in the EURECA platform, in order to improve the diffraction quality of crystals with the transcripts or to find another crystal form.

Table 1. Preparation of tRNA transcripts and their co-crystallization with an aminoacyl-tRNA synthetase.

Transcription and Purification :

(1) Preparation of a synthetic $tRNA^{Asp}$ gene under the control of T7 promoter (may be cloned for DNA amplification).

(2) A typical large scale transcription was conducted in a 4 ml medium (containing 5 µg of single stranded tRNA gene, originated from 300 µg plasmid, and 5600 units of T7 RNA polymerase).

(3) HPLC fractionation (on a preparative Bio-Sil TSK-250 column, Bio-Rad) of RNA transcripts (complete and abortive fragments) from DNA template and precursor nucleotides (recovery of 3.5 mg RNA).

(4) PAGE separation of active $tRNA^{Asp}$ transcripts from molecules terminating with one or two additional nucleotides and abortive fragments (on 40 x 30 x 0.2 cm^3 urea gels). After 2 gel runs (each gel loaded with 1.75 mg nucleic acid) and recovery of the tRNA bands by electroelution, 490 µg of tRNA were obtained (100-fold transcription of the synthetic gene).

Crystallization :

(5) Vapour diffusion co-crystallization of transcripts with yeast aspartyl-tRNA synthetase (AspRS) in 10 µl drops (10 mg/ml AspRS, 6 and 8 mg/ml transcript RNA, 40 mM Tris-Maleate pH 6.0, 5 mM $MgCl_2$, 25% ammonium sulfate) equilibrated against a buffered reservoir containing 46, 48 or 50% ammonium sulfate as precipitating agent. Two different stoichiometries were tested: 1/3 and 1/4 (AspRS/transcript).

(6) Crystals of cubic morphology (I432) appear after 8 days.

(7) Conditions to grow cubic crystals of wild-type modified $tRNA^{Asp}$ complexed with AspRS [20, 21] (drops of 30 µl containing 3 to 5 mg/ml AspRS, 1.2 to 2 mg/ml $tRNA^{Asp}$, 45 mM Tris-HCl pH 7.8, 0.18 mM EDTA, 0.18 mM DTE, 36% ammonium sulfate) differ from conditions in (5). Here, stoichiometry between AspRS and tRNA was 1/2.

Important to emphasize here is the high tendency of RNA molecules to undergo hydrolytic processes, either catalyzed by traces of nucleases, or by contaminating metal ions, or simply by alkaline type hydrolysis in water. The rate and amount of these hydrolyses are generally low, but can become important during crystallization assays that can last for long periods. Such degradations are favoured in certain sequences of the Pyrimidine-A type (*e.g.* [23]), especially when they are present in flexible loop regions [13]. It is probably this intrinsic chemical fragility of RNAs, combined with the difficulty of purifying these molecules which account for the low number of crystallographic studies done on this family of macromolecules.

4. THE INFLUENCE OF IMPURITIES ON CRYSTAL GROWTH

While it is a common laboratory observation that improvements in the purification of macromolecules improve their crystallization, little is known about the mechanisms by which impurities affect crystallization, and also on the effects they have on crystal morphologies and qualities. Here we show how minute amounts of contaminating proteins affect the solubility and crystal growth of hen egg-white lysozyme [24]. First, it was found that various lysozyme batches behave differently in well defined crystallization experiments. They showed different solubility behaviours and phase diagrams. Furthermore, under similar salt conditions these batches yield crystals with different growth morphologies. Careful biochemical analysis of the batches reveals different contaminant patterns (the amount of foreign proteins does not exceed a few percent in the most contaminated samples). Crystals with the most regular tetragonal morphologies (characteristic of lysozyme) were obtained with the purest samples; small crystals with frequent twinning grew from the contaminated samples. Experiments in which apparently pure lysozyme was contaminated on purpose by ovalbumin or serum albumin yielded as anticipated crystals with bad morphologies.

Noticeable is the fact that such variability in crystallization experiments due to minute contaminants was found with a protein, lysozyme, believed to crystallize readily without problems. Similar effects were observed by others, either induced by macromolecular contaminants (*e.g.* [25, 26]) or by small molecules present in the crystallization solutions [27].

5. TEMPERATURE AND pH AS VERSATILE FACTORS FOR ACTIVE CONTROL OF SUPERSATURATION

Crystal growers generally modify protein solubilities in crystallization assays by changing the concentration of precipitating agents (salts, organic solvents or PEG). However, other physico-chemical parameters such as temperature and pH can achieve the same goal but are seldom used in a controlled way. The reasons for that are not clear and rely probably to laboratory practices and to seemingly more convenient handling procedures with chemical precipitating agents. Temperature and pH changes, however, occur frequently in current crystallization trials because temperature generally is never well defined and undergoes fluctuations in cold rooms or when "room temperature" experiments are done. Similarly, pH in crystallization assays may be subjected to large variations in vapour diffusion methods when ammonium sulfate is the precipitating agent and when the reservoir has a different pH than the drops containing the macromolecules. This is due to the volatile ammonia, and it was demonstrated that the pH of the reservoir dictates the pH of the drop during equilibration [28, 29]. Temperature and pH fluctuations in crystallization trials may however be useful for a rapid screening of conditions, because in such a way a larger part of the parameter matrix will be assayed and thus crystals should be obtained more easily. But on the other hand such uncontrolled screenings cause non-reproducibility.

A rationale use of temperature and pH for reaching and changing supersaturation during crystallization experiments has advantages as compared to the use of precipitating agents. For instance in vapour diffusion methods, the kinetics for reaching concentration equilibrium between reservoir and drop are rather slow and can take more than one week for equilibrations in the presence of PEG [3, 30]. The kinetics of pH changes are much more rapid (hours instead of days) [29], and temperature equilibration is even faster. Thus changes of conditions can be done much more quickly and in principle without perturbing otherwise

the thermodynamics of the systems (*e.g.* no opening of crystallization boxes required for temperature changes). However, adjusting supersaturation by varying these parameters, especially temperature, requires adapted crystallization reactors. For this purpose we have developed a versatile crystallization chamber in which temperature is controlled by Peltier effect and kinetics of growth monitored by time-lapse video-microscopy [31].

To test the above ideas we used model proteins for which the phase diagrams were known. As predicted by the theory, it was possible to induce nucleation or modify growth characteristics of crystals by pH [28, 32] or temperature [31] changes.

6. IMPROVED CRYSTALLIZATION STRATEGIES

Provided macromolecules are of good biochemical quality, a convenient crystallization strategy may be divided in two stages. First, a wide screening of parameters should permit to define conditions where crystals appear. Use of statistical [3, 33] and automated [3] methods may facilitate this search. Initial diagnostics of the protein solutions in pre-crystallization conditions by light scattering [3, 7] (or by other physical methods sensitive to macromolecule heterogeneities) should permit to eliminate those solvent conditions leading to strong protein aggregations and thus to reduce the number of parameters to be screened.

In a second stage, conditions have to be refined in order to grow mono-crystals of large size and of good diffraction quality. While diffraction quality of crystals is not yet predictable, it is in principle possible to control crystal habit and size. However, improvement of crystal morphologies often is accompanied by improvements of internal crystalline order as the result of less perturbations during growth and less crystal poisoning by impurities. Such improvements should best be obtained in experiments conducted in crystallization set-ups where parameters can be controlled and monitored (especially supersaturation and the growth kinetics). Thus supersaturation may be reduced by temperature or pH variations as soon as nuclei (or small crystals) are detected by optical systems. In such a way the number of crystals within one assay should be reduced and their size increased. Moreover, because the physico-chemical conditions will remain unperturbed, it is expected that the growth of the crystals will not be perturbed as well. Systems permitting such active control of the growth process (e.g. [3, 31]) unfortunately are not yet widely used in crystallography laboratories but it is expected that they will be widespread in future in multireactor versions. This will however require instrumental developments, because the present systems are all laboratory prototypes not commercially available.

The aim of a crystal grower would be to predict characteristics of the desired crystals. Different routes may be imagined for engineering crystals of biological macromolecules. Beside exploring the physics of crystal growth, chemistry related approaches present interesting potentials. If adequate structural features of the crystal building blocks are known, their crystalline packing assembly should be understood, as in the case of particular DNA helical fragments [34, 35], and as a consequence the crystallizability of related structures should become predictable. Advantage can be taken of structural complementarity which may exist between a macromolecule and an appropriate solid matrix; this complementarity may induce epitaxial crystal growth of the macromolecules [36]. From another point of view, crystal growth of proteins should be modified in a controlled way by additives able to interact with these proteins in the crystal lattice, as was demonstrated in the small molecule field [37]. Finally three-dimensional crystal-like lattices may be constructed by self-assembly of macromolecular building blocks [38]. This last possibility was discussed for the self-assembly of DNA sequences, and a first application leading to the synthesis of a DNA cube was reported [39].

Acknowledgements: Our own experiments discussed in this paper were supported by grants from CNES, CNRS, Université Louis Pasteur in Strasbourg, and the Human Frontier Science Program. We thank S. Candau, D. Moras, and their colleagues for collaboration in some of the reported crystallogenesis studies.

REFERENCES

1. A. McPherson, *The Preparation and Analysis of Protein Crystals,* John Wiley & Sons, New York (1982).

2. R. Giegé and V. Mikol, Trends Biotech. *7* (1989) 277-282.

3. A. Ducruix and R. Giegé (eds) *Crystallization of Nucleic Acids and Proteins. A Practical Approach.* IRL Press at Oxford University Press (1992).

4. H. Michel (ed) *Crystallization of Membrane Proteins.* CRC Press, Boca Raton (1991).

5. A.-C. Dock, B. Lorber, D. Moras, G. Pixa, J.-C. Thierry and R. Giegé, Biochimie *66* (1984) 179-201.

6. R. Giegé, A.-C. Dock, D. Kern, B. Lorber, J.-C. Thierry and D. Moras, J. Crystal Growth *76* (1986) 554-561.

7. V. Mikol, E. Hirsch and R. Giegé, J. Mol. Biol. *213* (1990) 187-195.

8. V. Mikol, P. Vincendon, G. Eriani, E. Hirsch and R. Giegé, J. Crystal Growth *110* (1991) 195-200.

9. M. Skouri, M. Delsanti, J.-P. Munch, B. Lorber and R. Giegé, FEBS Lett. *295* (1991) 84-88.

10. M. Ruff, S. Krishnaswamy, M. Boeglin, A. Poterzman, A. Mitschler, A. Podjarny, B. Rees, J.-C. Thierry and D. Moras, Science *252* (1991) 1682-1689.

11. R. Giegé, J.D. Puglisi and C. Florentz, Prog. Nucleic Acid Res. Mol. Biol. *45* (1993) in press.

12. J.-P. Waller, J.-L. Risler, C. Monteilhet and C. Zelwer, FEBS Lett. *16* (1971) 186-188.

13. A.-C. Dock-Bregeon and D. Moras, Cold Spring Harbor Symp. Quant. Biol. *52* (1987) 113-121

14. N. Usman and R. Cedergren, Trends Biochem. Sci. *17* (1992) 334-339.

15. J.R. Wyatt, M. Chastain and J.D. Puglisi, BioTechniques *11* (1991) 764-769

16. C. Ehresmann, F. Baudin, M. Mougel, P. Romby, J.-P. Ebel and B. Ehresmann, Nucleic Acids Res. *15* (1987) 9109-9128.

17. E. Westhof, P. Romby, C. Ehresmann and B. Ehresmann. (1990) In *Theoretical Biochemistry and Molecular Biophysics* (D. Beveridge and R. Lavery, eds) pp. 399-409. Adenine Press, Guilderland, NY, USA.

18. P. Davenloo, A. Rosenberg, J. Dunn, and F.W. Studier, Proc. Natl. Acad. Sci. USA *81* (1984) 2035-2039.

19. V. Perret, A. Garcia, J. Puglisi, H. Grosjean, J.-P. Ebel, C. Florentz and R. Giegé, Biochimie *72* (1990) 735-744.

20. R. Giegé, B. Lorber, J.-P. Ebel, D. Moras and J.-C. Thierry, C. R. Acad. Sci. Paris, D-2, *291* (1980) 393-396.

21. B. Lorber, R. Giegé, J.-P. Ebel, C. Berthet, J.-C. Thierry and D. Moras, J. Biol. Chem. *258* (1983) 8429-8435.

22. M. Ruff, J. Cavarelli, V. Mikol, B. Lorber, A. Mitschler, R. Giegé, J.-C. Thierry and D. Moras, J. Mol. Biol. *201* (1988) 235-236.

23. P. Romby, D. Moras, M. Bergdoll, P. Dumas, V.V. Vlassov, E. Westhof, J.-P. Ebel and R. Giegé, J. Mol. Biol. *184* (1985) 455-471.

24. B. Lorber, M. Skouri, J.-P. Munch and R. Giegé, J. Crystal Growth (1993) in press.

25. C.W. Carter, J. Crystal Growth *90* (1988) 168-179.

26. C. Abergel, M.P. Nesa and J. Fontecilla-Camps, J. Crystal Growth *110* (1991) 11-19.

27. F.J. Jurnak, J. Crystal Growth *76* (1986) 577-582.

28. V. Mikol, J.-L. Rodeau and R. Giegé, J. Appl. Cryst. *22* (1989) 155-161.

29. J.-L. Rodeau, V. Mikol, R. Giegé and P. Lutun, J. Appl. Cryst. *24* (1991) 135-141.

30. V. Mikol, J.-L. Rodeau and R. Giegé, Anal. Biochem. *186* (1990) 332-339.

31. B. Lorber and R. Giegé, J. Crystal Growth *122* (1992) 168-175.

32. V. Mikol and R. Giegé, J. Crystal Growth *97* (1989) 324-332.

33. C.W. Carter Jr and C.W. Carter, J. Biol. Chem. *254* (1979) 12219-12223.

34. Y. Timsit and D. Moras, J. Mol. Biol. *221* (1991) 979-940.

35. Y. Timsit, E. Vilbois and D. Moras, Nature *354* (1991) 167-170.

36. A. McPherson and P. Schlichta, Science *239* (1988) 385-387.

37. I. Weissbuch, L. Addadi, L. Lahav and L. Leiserowitz, Science *253* (1991) 637-645.

38. N.C. Seeman, DNA and Cell Biol. *10* (1991) 475-486.

39. J. Chen and N.C. Seeman, Nature, *350* (1991) 631-633.

Crystallisation of Membrane Proteins

Gerry McDermott

Department of Chemistry
Glasgow University
Glasgow G12 8QQ.

Introduction

Integral membrane proteins are those which span the lipid bilayer of a membrane at least once. For many years it was presumed that such proteins would not be suitable for X-ray crystallographic analysis. It was argued that randomly oriented detergent molecules, necessary for solubilisation, would prevent the formation of an ordered crystal lattice. This hypothesis was proven wrong in 1980 with reports of low resolution diffraction from single crystals of bacteriorhodopsin (Michel & Osterhelt, 1980) and porin from *E. coli* (Garavito *et al.* 1980).

The observation of high resolution diffraction from crystals of the photosynthetic reaction centre from *Rhodopseudomonas viridis* (Michel, 1982), resulted in the first crystallographic structure of a membrane protein (Deisenhofer *et al.* 1985). Although several reports have been made of membrane protein crystallisation, in only a few cases have these resulted in a crystallographic structure being established. Indeed, membrane proteins account for less than 1% of the protein structures available. The preponderant obstacles are: obtaining "good" quality crystals; that is crystals of sufficient size and order such that they diffract X-rays to high resolution and the procurement isomorphous heavy atom derivatives. Consequently, crystallographic studies on membrane proteins have been considered to be *"more art than science"* (R.M. garavito. 1990).

This paper is a brief explanation of some of the reasons behind this statement, illustrated with a case, in which some of these difficulties have been overcome.

Membrane proteins

The major complication in producing crystals of membrane proteins, as opposed to soluble proteins, can be attributed to their location in the membrane phospholipid bilayer. Part of the protein is embedded in the quasi-solid bilayer in tight association with lipids; and is accordingly hydrophobic. The extramembranous regions, exposed to an aqueous environment, are hydrophilic (Figure 1). In order that crystallisation trials can be conducted the membrane must be disrupted: the protein isolated, and the resultant hydrophobic region solubilised.

Disruption of the membrane can be achived by the use of several types of reagent; Organic Solvents, Chaotropic Agents and Amphipathic Detergents. The latter is most frequently used, as the first two methods tend to result in a loss of protein integrity. Amphipathic detergents have also been sucessfully used to solubilise the hydrophobic

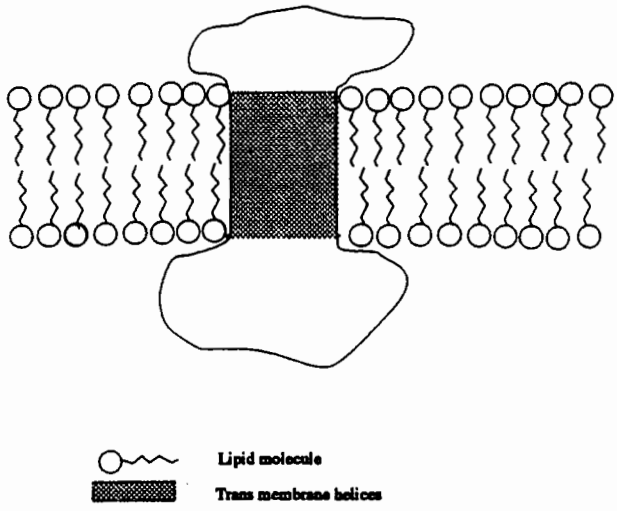


Figure 1:

region. To date, all reported membrane protein crystallisations have been achieved via detergent solubilisation methods.

Detergents and Crystallisation

Detergents are amphiphilic molecules, consisting of a polar head group and a hydrophilic hydrocarbon tail. Most of the detergents used for membrane protein crystallisation have been commonly used by biochemists for many years. Several classes of detergent have been sucessfully used, the common factors between them being;

- They are non-ionic or Zwitterionic at the pH used,
- The maximum hydrocarbon chain length is 12 carbon atoms,
- The have a high Critical Micelle Concentration.

Detergents have the particular physical property that above a certain molar concentration (the critical micelle concentration) they undergo a basic phase transition. Detergent monomers self-associate to form micelles. These are essentially spherical aggregates (around 50Å in diameter): the hydrophobic tail groups adopt an entropically advantageous position, by forming the core of the micelle. The polar head groups constitute the micelle surface and interact with water molecules. Amphipathic molecules such as membrane proteins, can be incorporated into detergent micelles to form protein-detergent "mixed micelles".

It is these protein-detergent micelles, with detergent constituting between $40 \rightarrow 60\%$ (by weight), which are used in crystallisation experiments. The detergent used for crystallisation must maintain protein integrity, whilst allowing the protein-protein contacts necessary for nucleation and crystal growth. In most cases this is a different detergent to the one used for initial membrane disruption and solubilisation. The choice of detergent

can be quite critical; in at least one case varying the lenth of the hydrocarbon chain by a single carbon atom precludes crystal growth.

To facilitate protein-protein contacts the physical properties of the detergent used may have to be altered by the addition of "small amphiphilic molecules". These interact directly with micelles affecting critical micelle concentration, micelle size and phase transitions. In the case of the reaction centre from *Rps. viridis* crystals could only be grown in the presence of heptane-triol. This lead Michel to propose the "small amphiphile concept": that small amphiphiles somehow induce crystallisation. To date this is the only protein for which the presence of of such additives, is an absolute requirement for crystal growth.

The classical method of inducing crystallisation is to drive a protein solution to a state of reduced solubility. This is achieved by increasing the concentration of a precipitating agent, or by altering some physical property such as the pH. This procedure, essentially an empirical trial and error method, has been utilised sucessfully in the crystallisation of detergent solubilised membrane proteins. The additional parameters such as choice of detergent and amphiphiles: increase the number of variables which must be permuted in crystallisation trials. This results in the probability of determining successful crystallisation conditions being reduced, and the optimisation of crystal growth parameters is accordingly complex.

Bacterial Photosynthetic Membrane Proteins

Phototropic purple bacteria are found in polluted lakes and rivers in the North of America. Under anaerobic conditions they have the ability to synthesis photosynthetic apparatus. The apparatus consists of two types of membrane-bound protein-pigment complexes:

- **Reaction Centres:** where light induced charge separation is carried out
- **Light Harvesting Complexes:** which capture light energy and transfer it to the Reaction Centre.

The X-ray crystallographic structure of the Reaction Centre, from several strains of purple bacterium, has been elucidated.
In Glasgow we have been concentrating our efforts on the crystallisation, and structure determination, of the antenna Light Harvesting Complexes (LH II).

The driving forces behind this work are two-fold:

- To further the understanding of the mechanism involved in bacterial photosynthesis.
- To use the experience gained as an aid in crystallising other membrane proteins.

The LH II complex is composed of two apoproteins; α and β, with molecular weights of 5.8 KDa and 4.6 KDa. Each apoprotein has stoichiometrically bound bacteriochlorophyll *a* and carotenoid molecules. A model has been proposed (Zuber, 1986) with the smallest stuctural unit an $\alpha_2\beta_2$ unit, binding six bacteriochlorophyll *a* and three carotenoid molecules. Hydropathy analysis has shown that the apoproteins contain a hydrophobic region of between 20 and 25 residues. It has been proposed (Brunisholtz *et*

al. 1986) that this would form an α helical membrane spanning region. This has been confirmed by far UV circular dichroism studies (Cogdell & Scheer, 1985).

A conserved Histidine on each of the apoproteins has been proposed as a binding site for bacteriochlorophyll *a*. The position of the carotenoid molecules have not yet been determined.

The LH II complex from purple bacteria is an ideal candidate for use in crystallisation trials for several reasons; the protein can be produced in large quantities, it is very stable and can be assayed readily for purity and integrity. The presence of specifically bound pigment molecules give rise to a distinctive visible/near Infra-Red spectra both *in vivo* and *in vitro*. The carotenoid molecules are responsible for absorption in the 400 to 540 nm range. For each $\alpha_2\beta_2$ unit two monomeric bacteriochlorophylls lead to absorption at 800nm. Four bacteriochlorophylls, in two interacting pairs account for the absorption at around 850 nm. The exact wavelenths at which absoption occurs depends upon the bacterial growth conditions, and the enviromental conditions encountered during isolation of the complex. A spectroscopic record is maintained of the protein throughout purification.

Crystallisation of LH II complex from *Rps. acidophila*

Crystals of LH II suitable for X-ray analysis were first grown in Glasgow in 1989. Diffraction to a resolution of 3.5Å was observed using a synchrotron source (Papiz *et al.* 1989) The initial optimism was short lived. In conducting a search for heavy atom derivatives it was found that approximately 4% of the crystals displayed significant diffraction. This hampered progress as there was no way of pre-determining which crystals would show diffraction. Optical examination proved fruitless, LH II crystals are opaque thus birefringent crystal properties could not be used to judge the degree of crystalline order. The external morphology of all the crystals appeared equally well defined under the microscope.

The only way forward was to expose the crystal to the X-ray beam, soak it in heavy atom solution and then re-expose. This was far from ideal from a radiation damage perspective, as well as being very time consuming. The task of solving this problem seemed to be Herculean, did the problem lie in; the initial solubilisation, the purification method , the choice of detergent , the crystallisation or post -crystallisation manipulation? The impression gained from obtaining crystals, which under the microscope appeared to be of very high quality, was that that the answer lay either in the purification, or post-crystallisation handling. The protein purification was refinened in order that any heterogeneity would be eliminated.

Optimising protein purification

The ratio of absorption at 270:850 nm can be used as an indication of the purity of the complex. An decrease in this ratio means protein is present which has no bound bacteriochlorophyll *i.e.* more contaminants or denatured complex. Initially a ratio of around 1:2.7 was considered to be reasonable for the purpose of crystallisation trials.

Using a modified purification protocol [1] resulted in increasing this ratio to around 1:3.7

Linear Dichroism

At this time the opportunity arose to carry out some Linear Dichroism measurements on LH II microcrystals at the laboratory of W. Mantele, in the University of Freiberg. It was hoped this study would yield information on the orientation of the pigment molecules with respect to the crystal lattice. Additionally spectral assays could be performed on crystals of LH II.

The technique of linear dichroism involves orienting a crystal in a beam of plane polarised light and recording an absorption spectrum. The crystal is then rotated by 90° and another spectrum recorded. All of the possible orientations of transition dipoles are present in solution, hence both of these absorptions will be equal. In a system of ordered molecules such as a crystal, provided the symmetry is reasonably low: a discrete number of orientations of transition dipoles will be found. The absorption of plane polarised light will be non-isotropic. Linear Dichroism spectroscopy had previously been carried out on crystals of LH II (A. Hawthornethwaite, personal communication), but these crystals had been grown using a different precipitant. At this time there was a requirement to carry out such measurements on crystals which had been grown using similar conditions to those used for X-ray analysis. This would allow a complete spectroscopic study of the complex; from *in vitro*, through purification, crystallisation and to the crystal form to be used for structure determination. In the case of LH II linear dichroism measurements can be used as an indicator of protein integrity (by comparison of solution and crystal spectra) and also as a measure of crystallinity.

Linear Dichroism Results

Crystals of LH II were grown, using the protocol which produced crystals for X-ray analysis. Initially the spectra recorded from crystals showed a high level of dichroism (Figure 2) and no discernable difference in either peak height or position from a solution spectra (Figure 3). A series of solution spectra were recorded, where the concentration of various components present during crystallisation were increased. The aim being to determine if any of these had a deleterious effect on protein integrity. The concentrations of these components could be increased, to a level greater than that encountered during crystallisation, with no adverse effects on the spectra. Linear Dichroism spectra were also recorded, and then re recorded several days later, again with no ill-effects.

The only avenue which remained was to treat some of the crystals with some of the "artificial mother liquor" which was used in mounting crystals and as a solvating agent for "heavy atom" compounds. Initially this had no effect, but when the spectra was re-recorded some 30 minutes later (Figure 4) it was obvious that both crystal and protein integrity were being lost. The significant absorption at 780nm indicating that the some

[1] increasing the solubilisation time, substituting a molecular-sieve column for an ion-exchange and increasing the stringency when selecting fractions from columns.

of the bacteriochlorophylls were now "free". The presence of dichroism suggested that the crystal still retained some degree of crystallinity. When the spectra was re-recorded an hour later (Figure 5) most of the bacteriochlorophylls were found to be "free" and absorption at 680 nm indicated the presence of oxidised bacteriochlorophyll.

Conclusion

The deterioration in crystallinity on the addition of "artificial mother liquor " was commensurate with the poor diffraction propeties of the LH II crystals. This problem was remedied by judicious alteration of the detergent concentration in the "artificial mother liquor ". The optimal detergent concentration was found to be approximately one half of the original estimate. Subsequent diffraction experiments showed a marked increase in the consistency of diffraction properties.

In our experience the over-riding factors in obtaining good diffraction from membrane protein crystals are:

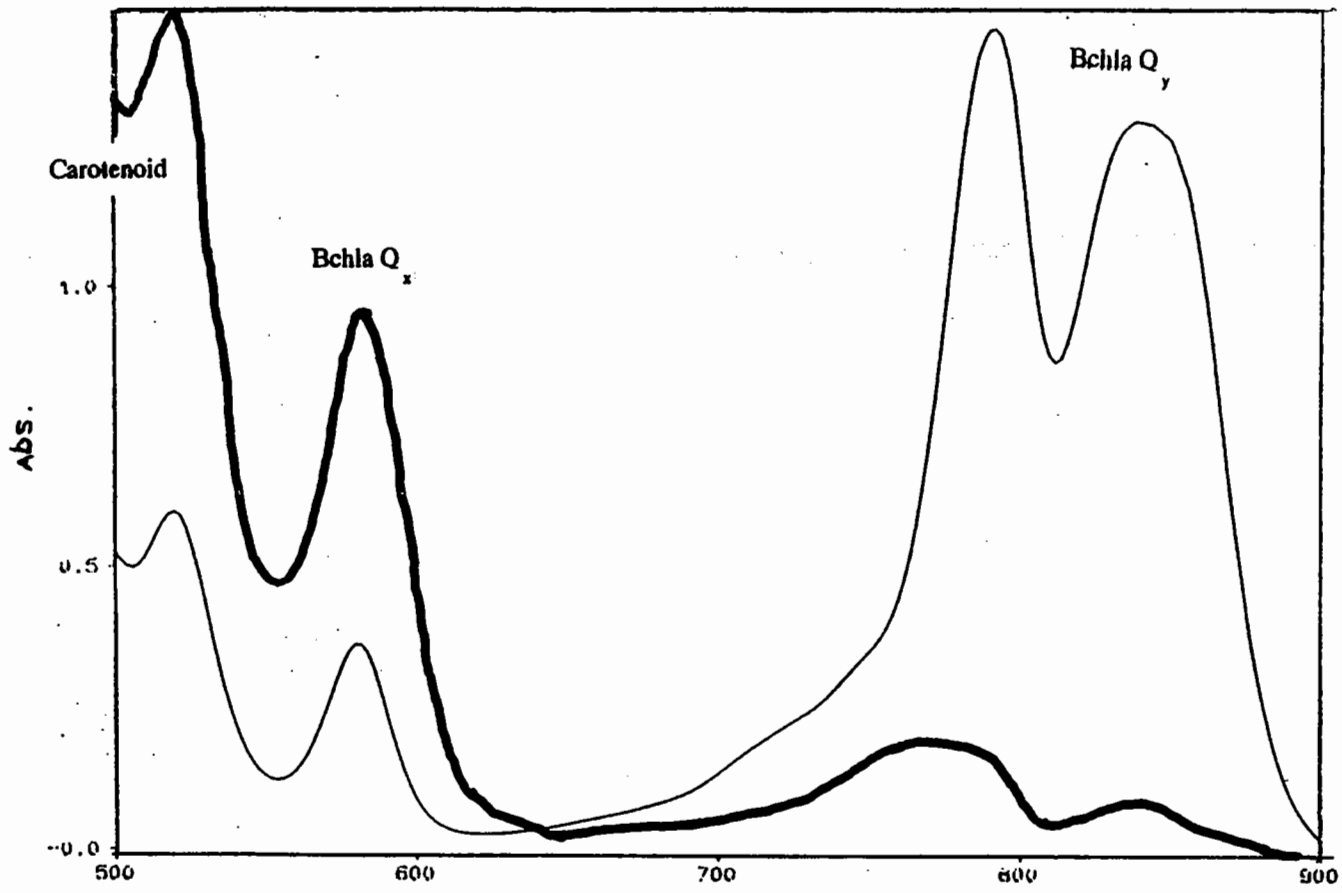
- The protein-detergent micelle which is used in crystallisation trials should be as homogeneous as possible. This is best achieved by optimising the solubilisation protocol and having more than one purification step that is based on "size". For example our initial purification step uses sucrose gradient centrifugation and the last step is a molecular sieve column on FPLC.
- During purification only the highest purity protein should be retained from each step ie. from FPLC molecular sieve only fractions adjacent to the peak maxima are retained.
- The post crystallisation methods would seem to be as important as anything that precedes.

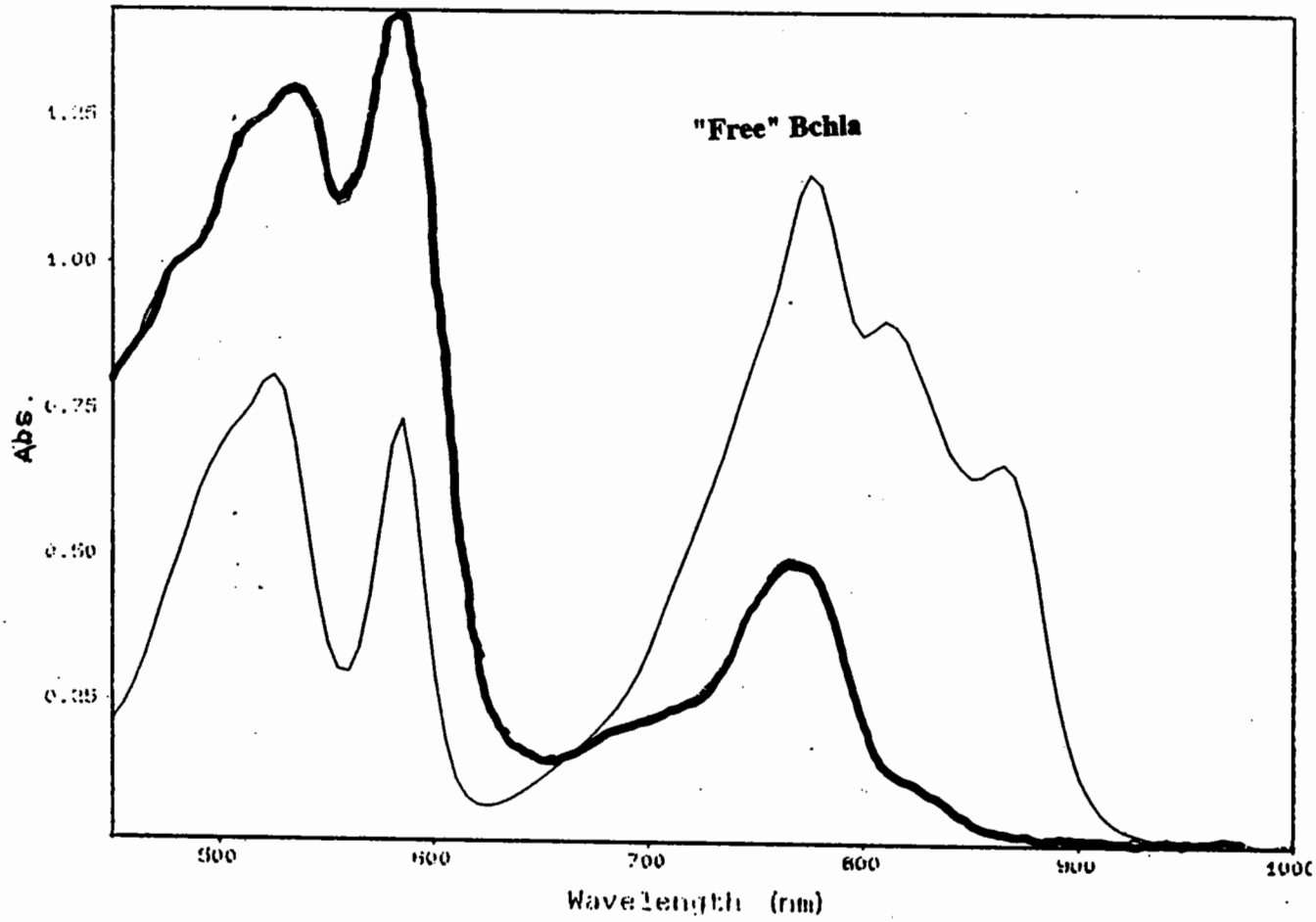
Epilogue

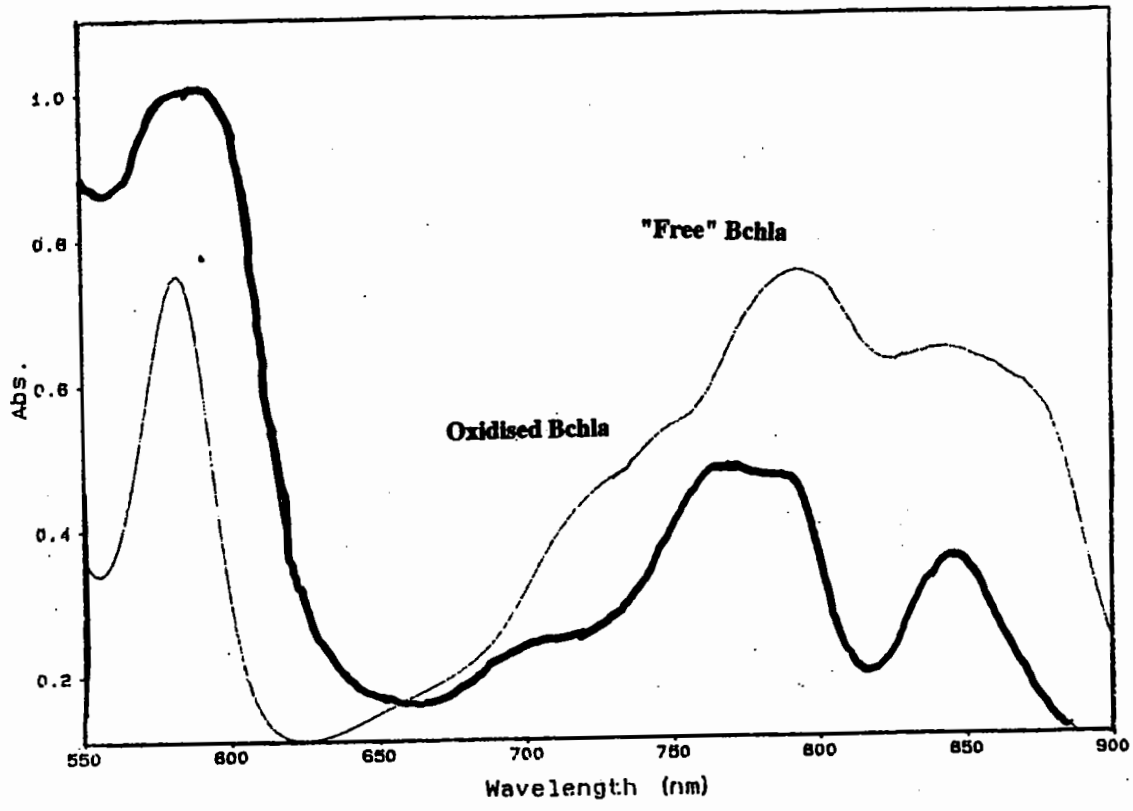
The LH II crystals now show diffraction to around 2.5Å using a synchrotron source. It is now possible to soak LH II crystals in moderate concentrations (around 5 mM) of "heavy atom" solutions for as long as desired with no detrimental effects.

Acknowledgements

The SERC Membrane Initiative for funding this project. The British Council for funding the trips to Freiberg. All at Freiberg for making life there so pleasurable. And the Glasgow mob; "Big" Neil, RJC, Steve,The Mong, The Pixie, Marie, Adrian, The two Daves and previous workers on the project; Anna Lawless and Marjo Thunnissen.
Special thanks to the old boy.... Cheers, Faither!


Carotenoid
Bchla Q x
Bchla Q y
Abs.
1.0
0.5
-0.0
500
600
700
800
900


"Free" Bchla
Abs.
1.00
0.75
0.50
0.25
500
600
700
800
900
1000
Wavelength (nm)



References

Michel, H., and Oesterhelt,D. (1980) *Proc. Nat. Acad. Sci.*, **77**, 1283-1285
Garavito, R. M., and Rosenbuch, J. P. (1980) *J. Cell Biol.* **86**, 327-329
Deisenhofet et al. (1985) *Nature* **318**, 618-624
Garavito, R. M. (1990) *Methods; a companion to Methods in Enzymeology.* **Vol. 1**, 57-69
Zuber, H., (1986) *Molecular Biology of membrane bound complexes in phototropic bacteria.* Plenum Press.
Bruinisholtz R. A., (1986) *Biochimica et biophysica Acta.* **849**, 295-303.
Cogdell, R. J. & Scheer, H. (1985) *Photochem. Photobiol.* **42**, 669-689.
Papiz, M. Z. *et al.* (1989) *J. Mol. Biol.* **209**, 833-835.

Some Practical Details of Data Collection at 100 K

S.J. Gamblin * & D.W. Rodgers.

* Howard Hughes Medical Institute and
Dept. of Biochemistry and Molecular Biology (B.M.B.)
Harvard University
Fairchild Building, 7 Divinity Avenue, Cambridge, MA

1.) INTRODUCTION.

That crystals of biological molecules are sensitive to X-ray irradiation is a fact of which most of us are all too aware. The advantages of preventing X-ray induced damage to crystals of biological samples are potentially immense and have been noted elsewhere (1). Having once solved the trivial practical problems of low temperature data collection, enormously improved data quality is possible for many challenging problems. Although many schemes for low temperature data collection have been shown to work well we shall, in general, describe the practices that are routinely used in this laboratory. In outline the procedure may be summarised as follows. The desired crystal is introduced to a new liquor containing some cryoprotective agent such as glycerol. The crystal is suspended in a thin film of this liquor inside a fibre loop. The loop containing the crystal is then mounted on the X-ray camera and rapidly cooled by a stream of N_2 gas at around 100K. At this temperature the film of liquid is transformed into a glass and the crystal is rendered relatively immortal. This article will address some of the practical details of this kind of low temperature data collection that have been adopted, developed and exploited at BMB.

2.) CRYOPROTECTANT.

Solutions of cryoprotectant should be made up with the appropriate crystal mounting buffer. The concentration of cryoprotectant should be at least as high as will afford flash freezing of a thin film of the liquor alone without ice formation. Table 1. shows both the variety and concentration ranges of cryoprotectants that have proved useful at BMB.

To some extent the choice of cryoprotectant will depend on the mother liquor that the crystal is grown in. Crystals grown from high salt will often require a high salt concentration in the cryo buffer to prevent dissolution (although it has been shown that high salt can usually be exchanged for a range of organic solvents (2)). Under these circumstances the highest concentrations of cryoprotectant attainable may be determined by solubility. In practice we have found that protein crystals in 65% Ammonium sulphate can be frozen very adequately with the addition of 15% glycerol.

Table 1.

cryoprotectants	concentration (w/v)
glycerol	13-25%
ethylene glycol	11-30%
PEG 400	25-35%
xylitol	22%
(2R,3R)-butane 2,3-diol	8%
erythritol	11%
glucose	25%
MPD	28% + 5% PEG 8K

Crystals may well grow in liquors that are already mildly or fully cryogenic. In this light it may be

considered prudent to include some cryogenic regimes in standard crystallisation screening procedures. Having chosen a suitable cryoprotectant the crystals must now be introduced to this liquor. This process may be carried out in a number of ways.

1) growth in cryoprotectant

2) direct transfer of crystal -- serial steps of increasing concentration or directly into final concentration.

3) dialysis -- again serial or direct.

4) exchange of liquor using flow cell and gradient maker.

All other things being equal, the first procedure is very satisfactory. It is quite reasonable to mount crystals directly into small loops (ca. 500μ diameter) from crystallisation drops of just a few μl. This procedure avoids many of the trials of determining what are good harvest conditions for a particular crystal.
Ideally we would like to use a cryoprotective regime in which the crystals would remain stable for some period of time. In practice we have seen many examples of good data collected from crystals in liquors in which they are not stable. Fortunately the rate of crystal dissolution, disorder and decay is often slower than the rate at which cryogenic properties are transferred to the crystal. Indeed the half rate of diffusion of small molecules into a 250μ crystal is on the order of a few minutes (3).
If cryoprotectant is dialysed into the crystals (especially if done in the cold room) then it may take up to12 hrs for equilibration. If conditions are not readily found where the crystals are stable over these time periods then the faster procedures described earlier should be tried. It is generally held that smaller crystals are more suitable for low temperature data collection than their more monolithic counterparts. Diffusion of small molecules and the rate of heat loss during flash freezing may be significantly faster for smaller crystals. Oftentimes crystal cracking due to changes in mother liquor is less problematic in the case of smaller crystals.

3.) CRYSTAL HANDLING & MOUNTING.

There are many ways of mounting a crystal. The use of loops (4) has proved extremely succesful. Suspending a crystal in a fine film of its mother liquor inside a fibre loop provides it with a very kind and gentle environment. The subsequent rate of flash freezing in a cold N_2 stream is also very good because of the small amount of non-crystal material which needs to be brought to 100K. Once frozen the loop material itself and the associated liquid film produce a very low amount of background scattering of X-rays. This enhances both signal/noise and eliminates blind regions during data collection.

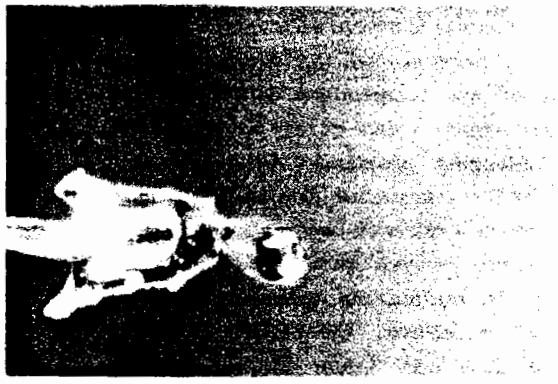


Figure 1.

Loops can be made from fine wire, glass and a range of thin fibres. Rayon fibres are very suitable for loops with diameters from 200 to 800 μ. Loops much bigger than this tend to fold over on themselves and so glass is often more suitable for larger crystal specimens. Loop size is generally chosen such that the crystal just fits inside the loop.

Most cryo-liquors will readily form a thin film across the loop in which the crystal is to be suspended. It is convenient to mount the crystal from a 3-well glass depression plate. The crystal is first gently pulsed with liquid from a pippette in order to free it from the glass surface and lift it into the body of the liquor. Then, using the loop itself, the crystal is wafted through the liquid until it reaches the meniscus. At this point the crystal will

often be held by surface tension at the air/liquid interface. The bottom edge of the loop is used to pull the crystal from the liquid into the thin film which forms across the loop. By lifting the loop perpendicularly through the meniscus very thin films are readily achieved. The crystal is now suspended inside the loop. The loop must then be transfered directly to the X-ray camera for flash freezing. It is highly desirable that this step be achieved in just a few seconds. We have sometimes found it necessary to humidify the air around the X-ray camera when working with very high salt crystals or when the air is very dry. In order to both facilitate fast transfer of crystals onto the X-ray camera and subsequent storage of these frozen crystals it is useful to mount the fibre loop on a steel cap. Figure 2 shows this arrangement. The arc slide of the goniostat has a small locating pin at its centre and the top surface of the slide has a piece of self-adhesive magnetic strip attached. The steel cap will then be held onto the goniostat by the magnet and the bevelled pin locates its position. The steel cap itself has an allen-head screw relieved into the shank of the cap about half way along its length. This screw allows adjustment of the brass pin which sits inside the steel cap. The brass pin is used to hold the wire/loop arrangement shown in Figure 1. The fine rayon loop is attached to the wire pin by epoxy and the distance from the centre of the loop to the base of the cap should be about 21mm for Supper arcs.

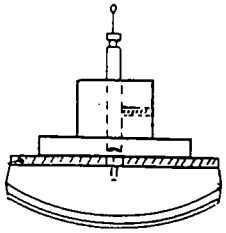


Figure 2.

During the course of the crystal transfer step the stream of cold N_2 gas from the cryosystem is temporarily blocked off. This is done with some shutter device such as a plastic ruler. When the crystal has been put in place on the magnetic mount the shutter is removed and the gas stream flows over the crystal rapidly cooling it to around 100K. A few practical points are worth bearing in mind. The end of the N_2 gas nozzle should be 8-14 mm from the crystal position and the crystal should sit reasonably well in the centre of the cold gas flow. Any adjustment to the nozzle position should be checked beforehand by monitoring the temperature at the crystal position with a fine thermocouple. The operating temperature range of the cryo-system should be between 90 and 110K and stable over a period of a few days to within a few K. In order to maintain stable temperatures at the crystal position and help prevent ice formation, the cold gas stream should be as free from turbulence as possible. The use of an outer coaxial stream of dry warm N_2 is useful both to protect the trajectory of the the cold gas flow and prevent ice formation at the nozzle and crystal positions. The flow rate of this outer stream is best set visually . At the best flow rate very little turbulence can be seen near the crystal postion. Ideally the loop containing the crystal would be mounted parallel with the cold gas stream to prevent disruption at the edge of the gas stream. If this condition cannot be met (because of the need to set a particular crystal orientation for example) then the pin or wire supporting the loop should be kept as small as possible at the point at which it bridges the edge of the cold gas stream. If the loop is mounted poorly then the disruption at the edge of the gas stream will make it very likely that the crystal will soon gather a layer of ice across its surface.

4.) DATA COLLECTION.

Crystals with rather asymmetric unit cell axes and/or high point group symmetry will often need to be mounted in a particular orientation with respect to the X-ray camera geometry. With character and dexterity it is possible to entice a crystal to enter a fibre loop in a particular orientation. Occasionally, however, adjustments of the crystal setting will be desireable. Supper now

supplies goniostats with an asymmetric bottom arc capable of nearly 70° of offset. These goniostats are also particularly useful for saving frozen crystals (next section).

Since N_2 gas at 100K is much denser than room temperature air it is more important than ever to minimise the collimator to beamstop distance. To this end it is useful to fit the collimator with a telescoping cap. The crystal can then be mounted in the N_2 stream and then the collimator cap translated along the collimator so that the aperature of the cap is almost touching the edge of the N_2 stream. At CHESS, Tom Irving has designed such a cap with an ion gauge inside. This enables necessary adjustments of the camera table relative to the X-ray beam position without interfering with the diffraction experiment.

If a frozen crystal is to remain on the X-ray camera for an extended period of time it is often necessary to build a tent around the crystal which can be flushed with dry N_2. In this way it is possible to keep the crystal and goniostat ice free even under very humid conditions. In the case of the Xentronics detectors used in the lab, the whole of the camera and detector are enclosed in a plastic bag which is sealed up against the top of the generator and the anode housing. In other cases a simple plastic box with a fine mylar window can be fitted just around the crystal-goniostat housing. These smaller tents are much easier to flush quickly with dry N_2 and it is not important to be concerned with creating good seals. They do, however, require a small stream of dry N_2 gas flow across the surface of the mylar window to prevent the buildup of condensation.

5.) CRYSTAL STORAGE & TRANSPORT.

We routinely freeze crystals and check their diffraction in the laboratory before data collection at the synchrotron. To do so requires a simple and reliable way of transfering and storing frozen crystals. The procedure used by us requires both the large arc goniostat and the magnetic mounting procedure described earlier. The large arc of the goniostat is placed vertically and the arc slide adjusted so that the frozen crystal is sitting close to the vertical with the magnetic mount sitting uppermost. A cryo-vial filled with LN_2 is then placed just under the crystal. In a single, unhesitating, action the cryo-vial is raised so that the crystal is submerged in the LN_2 in the vial. With a slight sideways twisting motion the steel cap holding thee loop-pin arrangement is displaced from its magnetic mount, dropping slightly until its base rests against the top edge of the vial.

This tube with the crystal assembly sitting in the top is then submerged under LN_2 to refill any liquid nitrogen that boiled off during the harvesting procedure. The vial assembly can then be clipped into a standard aluminium cell culture cane which can be kept in a long-term storage dewar. These dewars can be easily transported including air freight. Restoring the frozen crystal onto the X-ray camera is a straightforward reversal of this process.

6.) SOME RESULTS.

Increased mosaicity is one of the most frequently discussed topics regarding frozen data collection. The table below summarises the results obtained in this laboratory over the past year.

Table 2.

Survey of cryo results at BMB	
number crystal systems frozen	21
successful	19
range of unit cell size	
47x47x104A -- 168x162x636A	
range of solvent content	40 to 80%
mosaic increase on freezing	
none	7
<50%	6
50-100%	5
>100%	1

There are just a few cases which show a substantial increase in mosaicity. There are two main consequences. Increased mosaicity reduces the signal/noise ratio of the diffraction amplitudes, especially important at higher resolution. In most of the relevant cases shown in the table this effect was more than compensated for by collecting longer exposure times on the frozen sample. If the crystal has large unit cell dimensions then a large mosaicity may lead to spot overlap problems at higher Bragg angles. In our experience only virus crystals showed increases in mosaicity which made data collection unfeasible.

In general, flash freezing of protein crystals is a process which maintains the integrity of the crystal as it would be at the point of its first encounter with an X-ray photon. Since the crystal can be kept in this condition for a very long time there is often an apparent improvement in the diffraction limit of the crystal. Usually static disorder will determine how well the crystal diffracts. Freezing does not improve this. In some cases, however, dynamic disorder may be significant and in these cases cooling/freezing procedures have been shown to improve diffraction (2). Protein models refined against frozen data tend to show lower B-factors and more ordered solvent than their room temperature counterparts.

Frozen crystals may be regarded as immortal with respect to data collection using a laboratory X-ray source and so data merging from several crystals is not an issue. However some of the crystals we have studied require long exposures on synchrotron sources. In these cases we have found that crystals have a rather finite lifespan, which at CHESS F1 station amounts to about 6hrs of exposure time. In this particular case it was necessary to merge more than 10 crystals for the final data set. For all of these crystals the overall statistics for the data reduction was comparable to the merging R-factor within any one crystal taken individually. In a related sense heavy metal derivative data has been collected in the laboratory which shows good isomorphism to its frozen native crystal counterpart. Phases have been calculated from such data and structures solved. Frozen crystals notwithstanding, we still fully concur with Phil Evans' dictum that the three most important shortcomings of heavy metal derivatives are lack of isomorphism, lack of isomorphism and lack of isomorphism (5).

Of course the unit cell dimensions of crystals usually shrink somewhat on freezing; the important point is that these changes tend to be rather reproducible between crystals. Indeed the use of both frozen and unfrozen data sets from the same crystals with different cell dimensions has shown to be valuable in real space averaging procedures (6). The possibility of deliberately producing frozen crystals with different cell dimensions by using alternative freezing regimes for this purpose remains a tantalising prospect.

7.) REFERENCES.

1. Hope, H., et al. Acta Crystallogr. (1989) **B45** 190-199.

2. Petsko, G.A, J. Mol. Biol (1975) **96**, 381-392.

3. Ray,W.J., *pers. comm.*

4. Teng,T., J. Appl. Cryst. (1990) **23**, 387-391.

5. Evans, P. Daresbury Study Weekend on Phasing.

6. Bullough, P.A. & Hughson, F., *pers. comm.*

ACKNOWLEDGEMENTS.

We should like to thank all members, past and present, of the Harrison/Wiley laboratories whose work has aided in the development and use of many of the techniques described here. The work was supported in part by NIH grants CA-13202 and GM-39589 (to S. C. Harrison) and by a grant from AMFAR.

X-RAY COLLIMATION AND GENERATION

U. W. Arndt, MRC Laboratory of Molecular Biology
Hills Road, Cambridge CB2 2QH

ABSTRACT

The desirable characteristics of the primary X-ray beam for diffractometry are discussed with particular reference to data collection from single crystals of proteins and viruses. The parameters discussed are the cross-sectional area of the beam and the crossfire at the sample and the X-ray wavelength and bandwidth. It is shown that, compared with conventional systems, it is possible to increase the X-ray intensity at the sample and to reduce the X-ray tube power by large factors by employing a high-brilliance microfocus source and focusing collimators. The design is described of such a system which is in progress of construction.

The Primary Beam.

In designing an X-ray diffraction experiment we are concerned with the following parameters of the beam incident on the sample:

the shape and diameter of the beam,
the cross-fire, i.e. the conveyance or divergence at the sample,
the X-ray wavelength, λ, and
the bandwidth. $\frac{\delta\lambda}{\lambda}$.

These quantities are determined by the collimation of the beam.

In the interests of economy we should use the lowest-power source of X-rays sufficient for our experiment and should, therefore, design the collimation for the maximum efficiency of utilisation of the source.

The optimum parameters vary with the type of diffraction experiment, as does the need for optimisation. Among the most challenging X-ray diffraction experiments is the data collection from small single crystals of macromolecular substances. In the following we shall, therefore, discuss such measurements; here the diffraction spot intensities are low and the X-ray background is relatively high.

In the interests of maximum spot-to-background ratio, it is important that in direct space the primary beam diameter be no larger than the specimen crystal; in many cases we shall want the beam to be smaller than the crystal. In reciprocal space we should minimise the illuminated volume. The X-ray background is proportional to this volume to which is proportional the total number of diffraction spots recorded in the experiment: this number increases with the rotation range during an exposure, with the cross-fire of the beam and with the bandwidth $\frac{d\lambda}{\lambda}$. There are, of course, experiments where larger rotation ranges or bandwidths are used deliberately, for example in data collection by the rotation method, especially using image plates, and in Laue methods, but these experiments necessarily produce a poorer spot-to-background ratio than an 'ideal' experiment. A reduction in the cross-fire and in the bandwidth usually reduces the intensity of the diffraction pattern, but this may not matter with very powerful (SR) sources. The upper limit to these quantities is given by the need to resolve neighbouring reflections.

When deciding on the beam cross-fire it may be necessary to consider separately the plane containing the crystal rotation axis and the incident beam and the equatorial plane which is perpendicular to the rotation axis.

In four-circle diffractometry with a point detector all Bragg reflections are brought into the equatorial plane and the crystal is turned through its reflecting range by rotation about an axis (omega) perpendicular to this plane; the beam should thus be as nearly parallel as possible in this plane to produce sharp reflection profiles but the cross-fire in the perpendicular plane can be considerably greater. In diffractometry with an area detector reflections are measured out of the equatorial plane so that an isometrically parallel beam is desirable. In the screenless rotation method, using film or image plate, the rotation range for a given exposure is usually considerably greater than the reflection range so that we shall aim at a small cross-fire parallel to the rotation axis and allow a larger value in the plane perpendicular to it to obtain a more intense pattern. Note, however, that the number of partially-recorded reflections increases with the cross-fire in the equatorial plane.

The bandwidth should be matched to the mosaic spread of the sample so as not to broaden reflection widths unduly. Typical protein crystals have mosaic spreads between 5×10^{-4} and 5×10^{-3} radians. In this connection it should be noted that although the relative wavelength separation of the Kα doublet is more than twice as great for MoKα than for CuKα radiation the angular dispersion of a spot of a given spacing is very similar for the two characteristic radiations.

For many years all data collection from macromolecular crystals was carried out with an X-ray wavelength of about 1.5Å, largely because of the relative inefficiency of

photographic emulsions and of gas-filled radiation detectors for harder X-rays. The tendency today is to use wavelengths below 1Å at synchrotron beam lines and, perhaps, to go to MoKα radiation (λ = 0.71Å) in the laboratory. The advantages and disadvantages of shorter wavelengths are shown in Table 1.

Table 1

Use of Harder X-rays

Advantages

Smaller angle of incidence on detector, hence smaller parallax

Greater distance from sample for same minimum d, hence better spot-to-background ratio

Lower absorption corrections (see Table II)

Less radiation damage (?)

Smaller polarisation correction

Larger sample-to-detector distance, leading to a better spot-to-background ratio

Conventional sources: difference between mosaic and perfect monochromator crystals is smaller

Synchrotron radiation: variation of correction with distance from orbital plane is smaller; smaller differences between equatorial and meridional reflection corrections (see Fig. 1).

Disadvantages

Lower efficiency of some detectors

Lower intensity, - but note (i) no λ dependence of reflectivity of monochromator (ii) intensity of sample reflection $\alpha \lambda^2$ not λ^3

Less favourable emission spectrum for SR at bending magnets (but the number of characteristic X-ray photons per unit power from an X-ray tube target is about the same for different targets)

Table II

Variation of Absorption Correction for a Typical Protein Crystal

(From J. R. Helliwell, 1992)

λ	(Å)	1.743	1.488	1.040	0.620
Variation	(%)	±38%	±10%	±2%	±0.5%

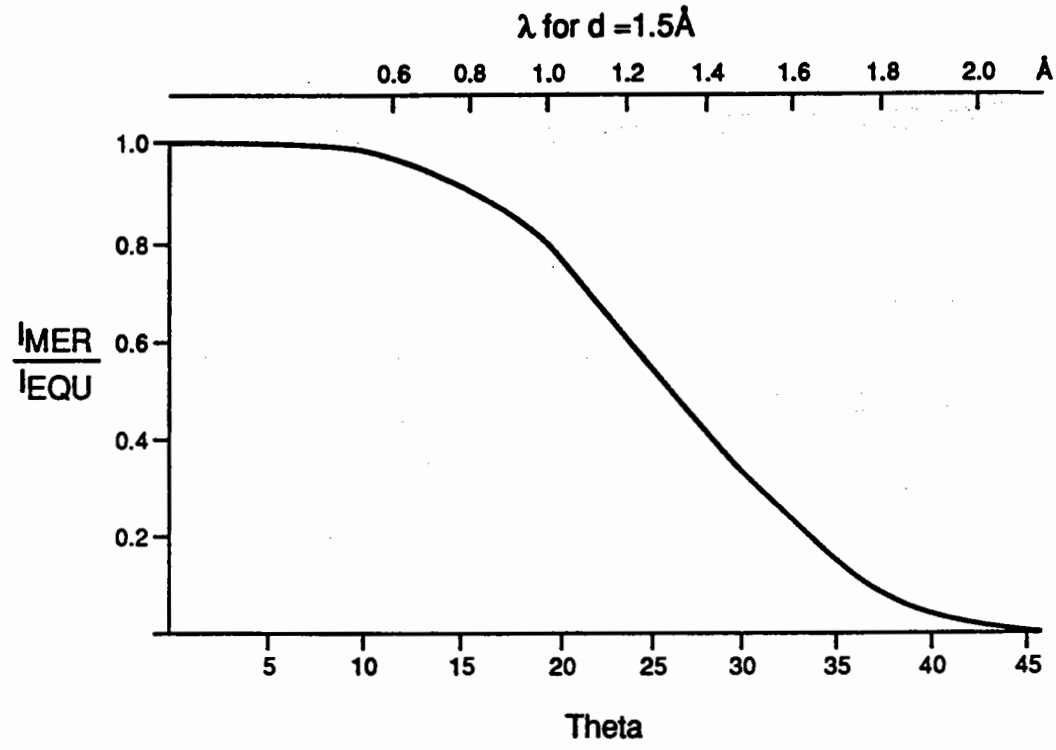


Figure 1. Ratio of the intensities of X-ray reflections in the meridional and in the equatorial plane as a function of Bragg angle. The equatorial plane is assumed to coincide with the orbital plane of the storage ring and the perfect-crystal monochromator disperses in the equatorial plane. Also marked are the wavelengths which correspond to a spacing of 1.5Å.

Collimation.

A collimator may be regarded as a device for producing an image of the X-ray source. This image may be at infinity (production of a parallel beam) or at a finite distance, it may be at unit magnification (e.g. with a pin-hole collimator), or it may be larger or smaller than the source, in which case the magnification may be different in two perpendicular planes containing the X-ray beam.

In all cases Liouville's Theorem applies and we can write

$$f\,\Delta\theta = w\,\Delta\phi \qquad (1)$$

where f is the size of the source and w that of the image, $\Delta\theta$ is the planar angle of collection at the source and $\Delta\phi$ the cross-fire (convergence or divergence) of the collimated beam. Appropriate suffixes may be given to the quantities in equation (1) for two orthogonal planes in systems which are not circularly symmetrical.

We have seen above that w and $\Delta\phi$ are fixed by the requirements of the experiment, so that the product $f\,\Delta\theta$ is also fixed. We can distinguish between three types of collimators.

1. $\Delta\phi = \Delta\theta$ - pin-hole collimation or curved crystal monochromators.

2. $\Delta\theta \neq \Delta\phi$ (Usually for single-crystal diffractometry $\Delta\theta > \Delta\phi$), f and $\Delta\theta$ variable within limits, to produce beams with the desired cross-fire for a given w, e.g. by the use of focusing mirrors.

3. $\Delta\theta$, $\Delta\phi$ fixed by the properties of the collimating element (e.g. an unsymmetrically cut perfect crystal), in which case equation (1) defines the size of the focus which can be 'seen' by the sample of size w.

We shall consider mainly collimators of the second type. For maximum utilisation of the source we want to make $\Delta\theta >> \Delta\phi$. It then follows that $f << w$ and that the distance between the collimating element and the sample must be much greater than that between the source and the sample (magnification of the image). A practical limit is set by the maximum desirable distance between collimator and sample, (perhaps 1m, even with an evacuated beam pipe).

The use of a microfocus source and a magnifying collimator brings with it an increase in the photonflux at the sample and a reduction of the X-ray tube power. The maximum power P which can be dissipated in the target of an X-ray tube is approximately proportional to the linear dimensions of the focal spot (Müller, 1927; Grider, Wright and Ausburn, 1986), so that

$$P = Kf \qquad (2)$$

For a stationary copper target K is approximately 500 watts mm^{-1}.

The intensity of a given diffraction pattern is proportional to the product of the X-ray tube power and of the solid angle of collection,

$$I = CP\ \Delta\theta^2 \tag{3}$$

where C is a constant for a given crystal diffraction pattern. Combining (1), (2) and (3)

$$\begin{aligned} I &= CKf \left(\tfrac{w\Delta\phi}{f}\right)^2 \\ &= CK\ (w\ \Delta\phi)^2\ f^{-1} \end{aligned} \tag{4}$$

For a given pattern with w and $\Delta\phi$ fixed I is thus proportional to f^{-1} and P to f.

Typical problems of data collection from proteins require values of w of about 0.3 mm and of $\Delta\phi$ of 10^{-3} rad. We shall see in the next section that the maximum value of $\Delta\theta$ which can be achieved with practically feasible collimations is between 3×10^{-2} and 5×10^{-2} rad, requiring source sizes between 10μm and 3 μm and promising a 20- to 150-fold increase in intensity coupled with a 20- to 50-fold reduction in X-ray tube power as compared with a conventional system for which f = 300 μm. The system which we are constructing and which is described below is intended to achieve gains of this order.

Practical Focusing Collimators

The following methods have been used for focusing X-rays.

1. Specular reflection by mirrors curved in one plane (Kirkpatrick & Baez, 1948; Franks, 1955).
2. Specular reflection by toroidal mirrors (Elliott, 1950 and a large number of applications for softer X-rays in X-ray astronomy and microscopy).
3. Reflection by graded-spacing crystals (Smither, 1982; Knapp and Smither, 1986).
4. Reflection by artificial crystals (multi-layers) (see, for example, the review by Barbee, 1986), with a prospect of using curved graded-spacing mirrors (Philip, Rivoira, Lepêtre and Rasigni, 1988).
5. The use of bundles of curved X-ray guide tubes (Kumakhov and Komarov, 1990); producing too large a cross-fire for single-crystal studies.

The first two methods have been discussed by Arndt, 1990, but note that this paper considered only parabolic and paraboloidal mirrors and neglected more promising elliptical and ellipsoidal mirrors. We are at present attempting to produce small diameter (~1 mm) gold-coated ellipsoidal mirrors with the dimensions shown in Table III (Hudec, *et al.* 1992).

It can be shown that with an ellipsoidal mirror one can achieve planar angles of collection at the source of up to three times the critical angle for total external reflection, θ_c. (For a gold or platinum surface $\theta_c \equiv 10^{-2}$ rad for 8 keV X-rays). With two successive reflections at pairs of mirrors flat in one plane and elliptically curved in the other plane, (similar to a double Kirkpatrick-Baez or Franks mirror pair), the maximum planar angle of collection is about $2\theta_c$; this latter arrangement lends itself to being upgraded later by using a multi-layer instead of a gold surface for which glancing angles of incidence some 2.5 greater are possible.

Table III

Ellipsoidal Mirror For 8keV X-rays

Length	26 mm
Smallest Dia	0.394 mm
Largest Dia	0.730 mm
Length of Major Axis	600 mm
Magnification	x30
Solid Angle of Collection	8.9×10^{-4} sterad
Surface	Au ($\theta_c = 10^{-2}$ rad.)
RMS Surface Roughness	< 1.5 nm
Source to Mirror	10 mm
Source Dimensions	20 μm x 200 μm

It should be noted that for any curved surface, (other than a cylindrical one for which $\Delta\theta = \Delta\phi$), the angle of deflection at the mirror 2θ varies along the length of the mirror; for Bragg reflection to occur it is thus necessary for the interplanar spacing of a natural or artificial crystal to vary with distance along the mirror in the appropriate fashion.

It must be remembered that the critical angle is proportional to the X-ray wavelength. The solid angle of collection at the source is thus proportional to λ^2 and specularly-reflecting focusing collimators thus rapidly become less effective for wavelengths shorter than 1.5Å. It is for these shorter wavelengths, therefore, that multi-layer mirrors are particularly important.

The practical problem in producing X-ray mirrors, whether coated with a specular or a multi-layer reflecting surface, lies in the fact that it requires a very high degree of surface smoothness for near-unity reflectivity (RMS deviation from flatness <10Å).

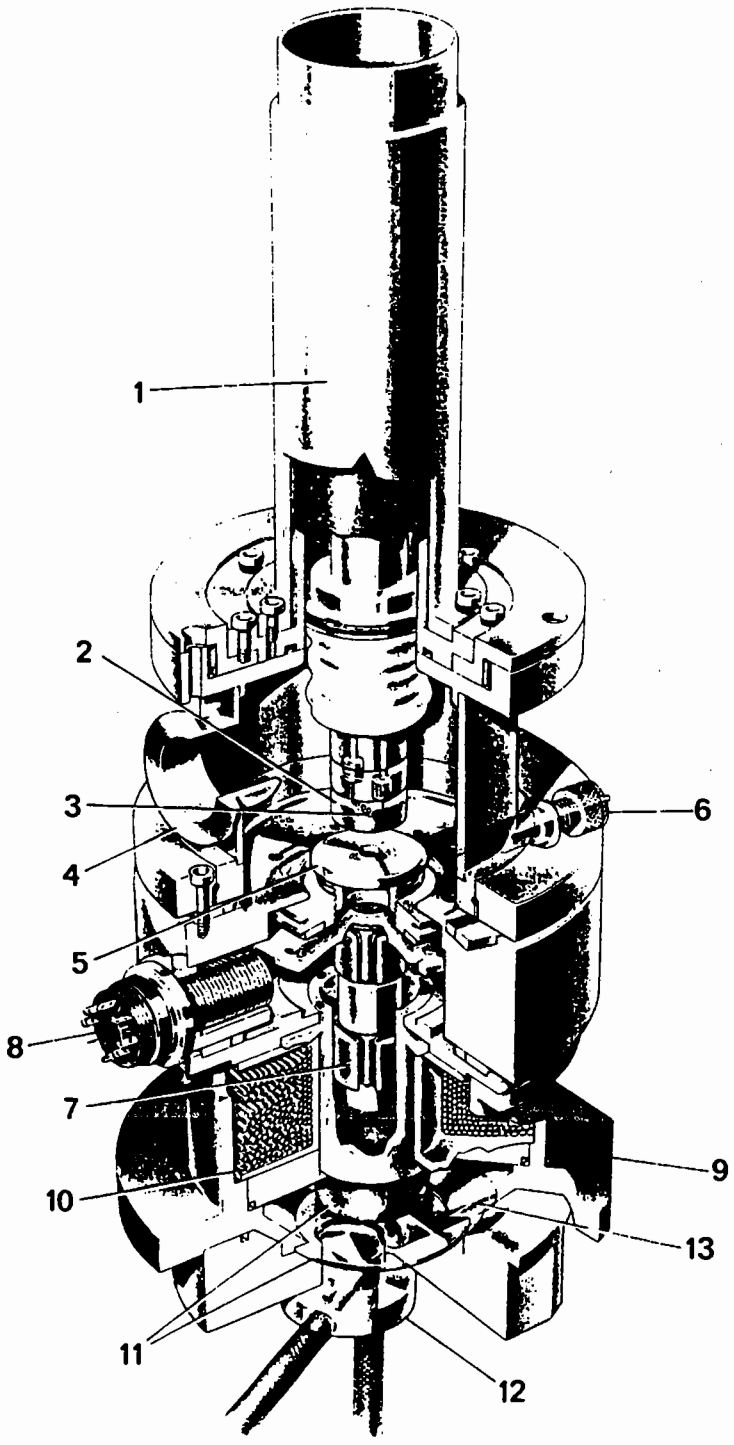


Figure 2. Stationary-anode X-ray Tube

1. Electron gun with centering adjustments. 2. Wehnelt cathode. 3. Tungsten filament. 4. Vacuum port. 5. Anode aperture. 6. Anode aperture translation (one of two at right angles). 7. Deflection coils (two pairs at right angles). 8. Electrical feed-through for coils. 9. Magnetic condensing lens. 10. Winding. 11. Soft iron pole-pieces. 12. Water-cooled stationary copper target. 13. Ellipsoidal mirror (x y z translations not shown).

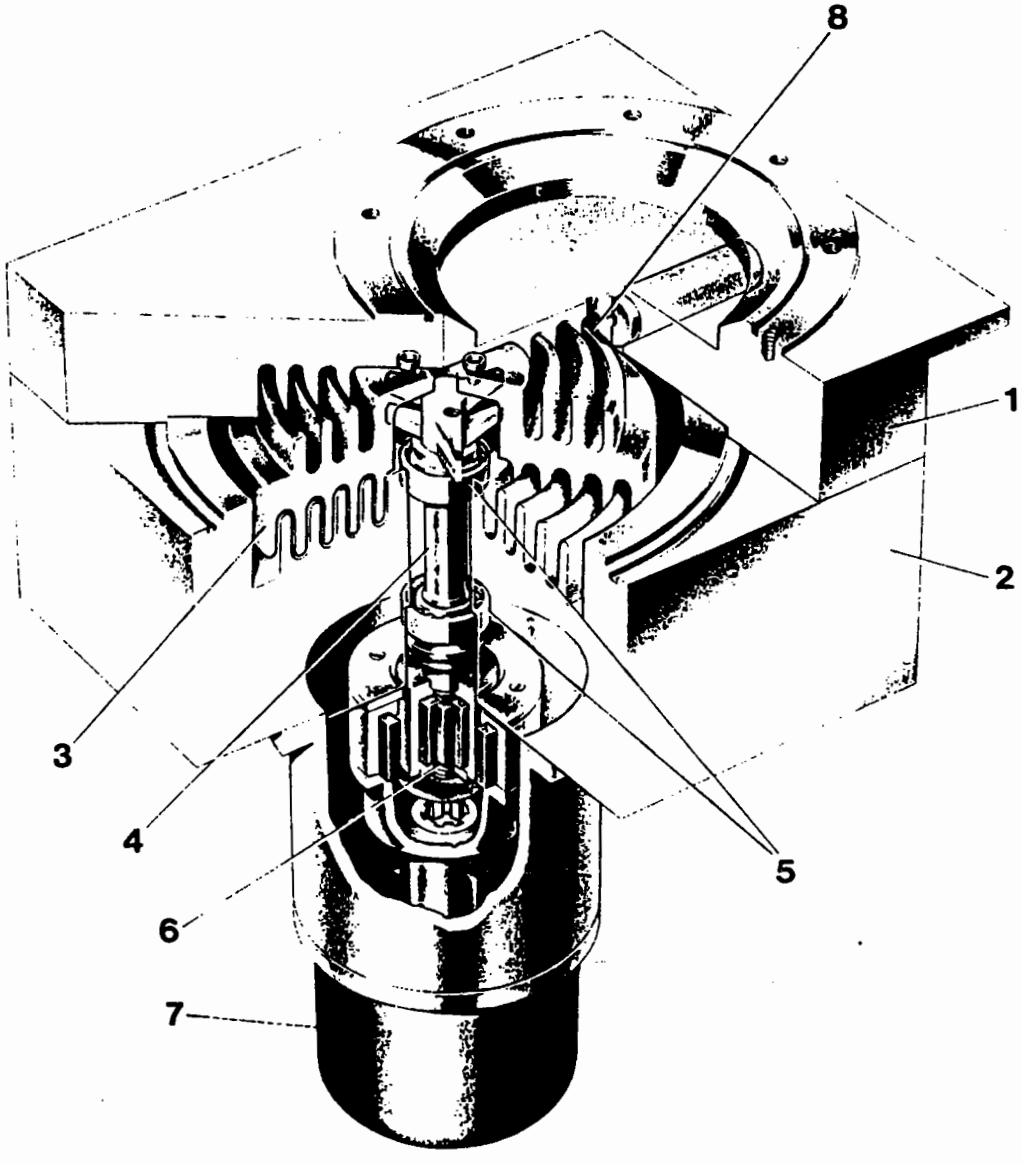


Figure 3. Rotating-anode assembly

1. Upper water-cooled baffle block. 2. Lower water-cooled baffle block.
3. Finned rotating target. 4. Rotor shaft. 5. High-temperature ball-races.
6. Magnetic coupling. 7. 10,000 RPM motor. 8. Electron focus.

Experimental Project

We are engaged in a collaborative project to construct a microfocus X-ray tube with a distance of 10mm between the electron focus and the (re-entrant) tube window and of various focusing mirror optics for use with this tube. This project is funded by the Medical Research Council and by the Royal Society Paul Instrument Fund, and the participants are the author and P. Duncumb and J.V.P. Long in Cambridge who have made the design studies and produced the design of the X-ray tube, Keith Bowen and Harry Hingle at Warwick University who are manufacturing planar-elliptical mirrors and also the masters from which the ellipsoidal mirrors are replicated. R. Hudec, A. Inneman and L. Pina in Prague are undertaking the replication of the ellipsoidal mirrors.

The X-ray tube, the construction of which is approaching completion, is shown in figure 2. The electron beam is focused and positioned magnetically. The first version of the tube is fitted with a stationary target: at a later stage we intend to fit the rotating target shown in figure 3. In view of the relatively low maximum dissipation of 400 watts the finned target is radiation-cooled and driven through a magnetic coupling. There are, therefore, no vacuum or water shaft seals.

Conclusions

We have described the criteria behind the design of a high-brilliance low-power X-ray tube with focusing collimators, intended initially for protein crystallography. The principle of a small source size of which a magnified image is formed on the sample could be applied to other X-ray sources such as plasma sources. The X-ray tube should be usable with other focusing elements, e.g. zone plates or micro-channel plates (Wilkens, *et al.* 1988), or with a combination of focusing elements which concentrating monochromators.

A well-designed laboratory system should be capable of achieving most of the advantages of synchrotron radiation, except tuneability, at least until detectors become available which can utilise the highest X-ray intensities.

References

Arndt, U.W. (1990), J. Appl. Cryst. 23, 161-168.

Barbee, T.W. (1986), Opt. Eng. 25, 898-915.

Elliott, A. (1965), J. Sci. Instrum. 42, 312-316.

Franks, A. (1955), Proc. Phys. Soc. London B68, 1054-1069.

Grider, D.E., Wright, A. & Ausbum, P.K. (1986), J. Phys. D. 19, 2281-2292.

Helliwell, J.R. (1992), Macromolecular Crystallography with Synchrotron Radiation. Cambridge: CUP p. 254.

Hudec, R., Arndt, U.W., Inneman, A. & Pina, L. (1992), XII Internat. Congress on X-ray Optics and Microanalysis, Manchester. Conference Proc.

Kirkpatrick, P. & Baez, A.V. (1948), J. Opt. Soc. Amer. 38, 766-774.

Knapp, G.S. & Smither, R.K. (1986), Nucl. Instrum. & Meth. A 246, 365-367.

Kumakhov, M.A. & Komarov, F.K. (1990), Physics Reports 191, 289-350.

Müller, A. (1927), Proc. Roy. Soc. London A 117, 30-42.

Philip, R., Rivoira, R., Lepêtre, Y. & Rasigni, G. (1988), Appl. Optics 27, 1918-1919.

Smither, R.K. (1982), Rev. Sci. Instrum. 53, 131-142.

Wilkens, S.W., Stevenson, A.W., Nugent, K.A., Chapman, H. & Steenstrup, S. (1989), Rev. Sci. Instrum. 60, 1026-1035.

Auto-indexing of Rotation Diffraction Images and Parameter Refinement

Andrew G. W. Leslie

Medical Research Council,
Laboratory of Molecular Biology,
Hills Road,
Cambridge CB2 2QH

1. Introduction

The primary objective of auto-indexing and refinement is to permit the accurate prediction of reflection positions in terms of their detector co-ordinates (Xd Yd) and their positions and widths in the rotation angle, phi. This is essential in order to minimise the errors in estimating the integrated intensities particularly if profile fitting techniques are employed.

Auto-indexing also greatly simplifies the task of data processing; prior to the advent of auto-indexing procedures the correct determination of crystal orientation was often the rate-limiting step in processing data collected using the rotation method. The incentive for developing auto-indexing algorithms for rotation data was provided by the availability of 2-D area detectors which offered considerable improvements over film methods in both the ease of data collection and processing and in data accuracy. Program suites such as XENGEN (Howard et al., 1987), XDS (Kabsch, 1988b) and MADNES (Messerschmidt and Pflugrath, 1987) were developed to make processing data from area detectors as simple and automatic as possible, and a robust auto-indexing algorithm is an essential component of any such package. There are a number of procedures described in the literature which fall into two classes: those which were designed to work with single oscillation images (Vriend and Rossmann 1987, Kabsch 1988a, Kim 1989, Higashi 1990) and those designed to be used with an area detector where data from a larger segment of reciprocal space is available (Howard, 1986, Tucker 1986, Prange 1986, Tanaka *et al* 1990). However, the basic concept employed in all these procedures is identical, and this will be described.

Once an initial orientation has been determined using auto-indexing procedures, the subsequent refinement of the orientation and in addition crystal, beam and detector parameters using conventional least squares procedures is relatively straightforward.

2. Definition of Crystal Orientation

Following Busing and Levy (1967) the orientation of the crystal can be described as

$$\mathbf{X} = \Phi\phi_z\phi_y\phi_x\mathbf{U}\,\mathbf{B}\,\mathbf{h} \qquad \ldots \qquad (1)$$

where

X is a vector in the laboratory frame giving the position of the reciprocal lattice vector with indices **h**

B is an orthogonalisation matrix, which defines a set of orthogonal axes based on the crystal axes. This matrix depends only on the crystal cell parameters.

U is a pure rotation matrix describing the orientation of the crystal in the laboratory frame in a "standard" setting (for example, with one cell axis parallel to the spindle axis and another along the X-ray beam direction).

ϕ_x,ϕ_y,ϕ_z are (small) rotations around the laboratory frame X,Y,Z axes (missetting angles)

Φ is the rotation around the spindle axis for a single axis device, or more generally the goniostat matrix.

For convenience, the product of the **U** and **B** matrices is often denoted as the "setting matrix" **A**

$$\mathbf{A} = \mathbf{U}\,\mathbf{B} \qquad \ldots \qquad (2)$$

The orthogonalisation matrix B is given by (Busing and Levy, 1967)

$$\begin{pmatrix} a^* & b^*\cos\alpha^* & c^*\cos\beta^* \\ o & b^*\sin\alpha^* & -c^*\sin\beta^*\cos\alpha \\ o & o & c^*\sin\beta^*\sin\alpha \end{pmatrix} \qquad \ldots \qquad (3)$$

which as mentioned earlier depends only on the crystal cell parameters.

It should be pointed out that there is some redundancy in equation (1), which arises for historical reasons. With film methods, it was usual to attempt to orient the crystal by eye in some convenient "standard" orientation, using the crystal morphology and setting photographs as a guide. The crystal orientation was then refined using data from two or more "still" photographs and the true orientation described in terms of missetting angles ϕ_x,ϕ_y,ϕ_z from this standard orientation. In these circumstances the missetting angles will generally be small. In auto-indexing procedures the initial orientation of the crystal is arbitrary, so there is no "standard" setting and ϕ_x,ϕ_y,ϕ_z alone define the crystal orientation. In this case ϕ_x,ϕ_y,ϕ_z are perhaps better described as *setting* angles rather than *missetting* angles and we can re-write equation (1) as

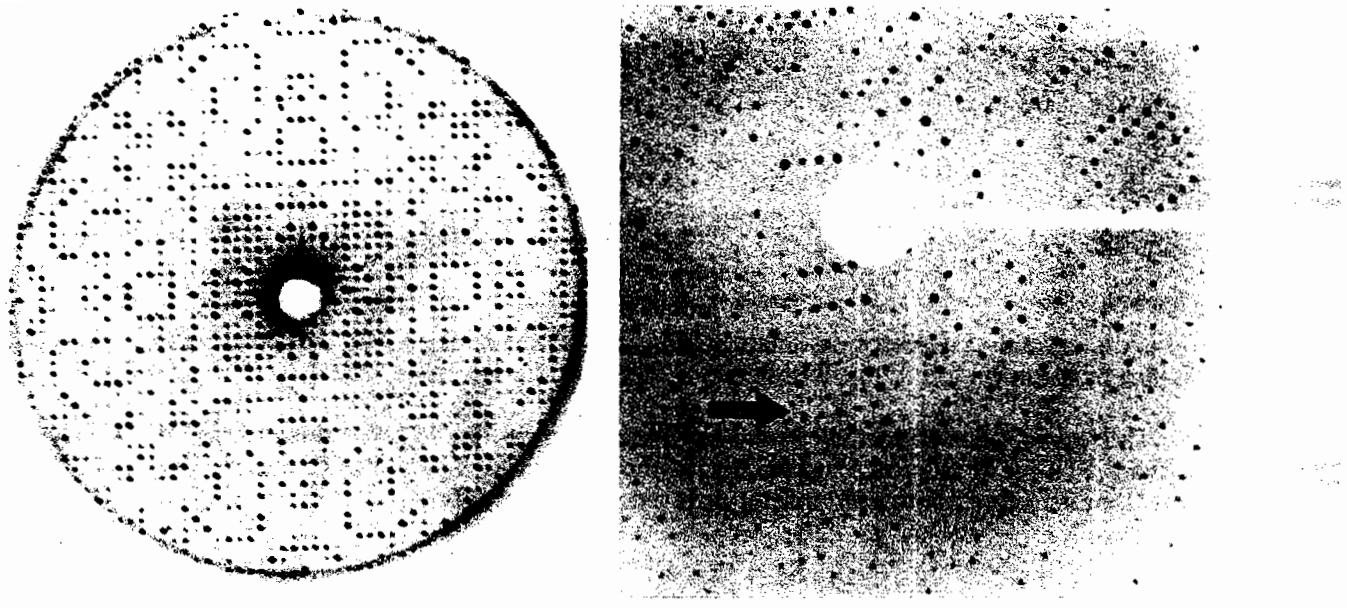
$$\mathbf{X} = \Phi\ \mathbf{U}\ \mathbf{B}\ \mathbf{h} \qquad \ldots \qquad (4)$$

where the rotations ϕ_x,ϕ_y,ϕ_z are incorporated in the **U** matrix. This is the approach adopted in the MADNES package.

If the cell parameters of the crystal are known with reasonable accuracy, then auto-indexing requires a determination of the matrix **U**. When the cell parameters are unknown, both **U** and **B** have to be determined. In general, methods which use data from a single rotation image (typically 0.5°-2°) require prior knowledge of the cell parameters, but those working with data collected in fine phi slices (e.g. 0.1 - 0.5°) can determine both the cell parameters and orientation provided that data corresponding to a rotation of a few degrees is used. (These methods will also work on coarse phi slices, albeit less reliably).

3. Auto-indexing when cell parameters are known

To illustrate the principle underlying auto-indexing it is useful to consider the relatively straightforward task of indexing a zero level precession photograph (Figure 1a). This could be done as follows:



(a) (b)

Figure 1. (a) A zero level precession photograph, which shows an undistorted image of a section of the reciprocal lattice. The choice of two seed vectors which define this lattice is straightforward. (b) The central region of a 1° rotation image recorded on an image plate detector. A pseudo-hexagonal lattice can be identified easily in the principle zone (arrowed).

a) Choose two vectors which together define the lattice of spots in the precession photograph. In Figure 1 the obvious choice is a horizontal and a vertical vector.

b) Measure the lengths of these seed vectors (by measuring the distance between many pairs of spots and taking an average) and the angle between them.

c) Generate, using the known cell parameters, a list of calculated reciprocal lattice vectors with low indices (001, 010, 100, 110, 101, 011, etc.).

d) Identify the seed vectors chosen in (a) by comparing their lengths and relative orientation with the list of calculated lattice vectors from (c). Usually it will be possible to uniquely identify the two seed vectors (ignoring solutions related by crystallographic symmetry). The identification of these two vectors then determines the orientation of the crystal.

The case of indexing a precession photograph is made simpler because the photograph represents an undistorted representation of the reciprocal lattice. In the case of a rotation image, it is also possible to identify sections of the reciprocal lattice, particularly when a crystal is almost aligned on a principle zone (Figure 1b). In this case the lattice is distorted, but using the Ewald sphere construction the observed spot positions $\left(X_d^1, Y_d^1, \Phi^1\right)$ can be mapped back into reciprocal space to give a set of scattering vectors S_i^{Φ}. In vector notation

$$S_i^{\Phi} = \begin{pmatrix} D/r - 1 \\ X_d^i / r \\ Y_d^i / r \end{pmatrix} \quad \ldots \quad (5)$$

where

$$r = \sqrt{X_d^{i2} + Y_d^{i2} + D^2} \quad \ldots \quad (6)$$

D is the crystal to detector distance and X_d, Y_d are the spot coordinates relative to the direct beam position on a flat, untilted detector (see for example Kim, 1989).

This gives the scattering vector when the reciprocal lattice point lies *exactly* on the Ewald sphere. To place all scattering vectors in the same co-ordinate frame, we must correct for the fact that different spots will arise at different Φ values, giving

$$S_i = \Phi_i^{-1} S_i^{\Phi} \quad \ldots \quad (7)$$

This creates a problem when a large rotation angle per image has been used (e.g. 0.5° - 2°) because there is normally no information on the true Φ values for different spots. In practice all spots are assigned a Φ equal to the mid point of the rotation range, but this is of course only an approximation and will inevitably give rise to errors in the vectors S_i. When fine phi slicing is used (e.g. 0.1 - 0.25° rotation per image) the spots will usually extend over several images and an accurate Φ centroid can be determined empirically, and consequently the scattering vectors S_i will be far better determined.

A list of difference vectors:

$$S_{ij} = S_i - S_j \quad \ldots \quad (8)$$

is constructed from the set of scattering vectors S_i. The resulting difference vectors are then sorted on increasing length $|S_{ij}|$. This list will contain "clusters" of vectors corresponding to short reciprocal lattice vectors (clusters rather than points because of errors in the vectors S_i which will be proportionately greater in the difference vectors S_{ij}). The vectors within each cluster are averaged to improve accuracy. The shortest vectors (which will also be those with the highest multiplicity) should correspond to reciprocal lattice vectors with small Miller indices (e.g. 100, 010, 001, etc.). The orientation of the crystal can be determined by identifying two (or more) of these short difference vectors (the seed vectors) and this can be done in exactly the same way as for the precession photograph providing the cell dimensions are known and the seed vectors are non-collinear.

If the scattering vectors arise from a limited Φ rotation (e.g. a single rotation image) there will not, in general, be any short difference vectors in a direction that is significantly out of the plane of the detector. In these circumstances it is only feasible to attempt to identify two seed vectors (although different pairs of vectors can be used). This is the method used by Kabsch's program REFIX (Kabsch, 1988a), and is sometimes known as the 2D approach to auto-indexing. If a larger rotation range has been used (e.g. 5-20 degrees, or more than one Φ segment at different Φ values) then it is an advantage to use *three* seed vectors as this will give enhanced discrimination against incorrect solutions. This approach (3D auto-indexing) is usually used in MADNES, XENGEN, XDS and the Rigaku Image Plate software.

Once an initial orientation has been determined (using either two or three seed vectors) the crystal orientation and cell parameters can be refined by minimising:

$$\Omega = \sum_i \left(\mathbf{h}_i - [\mathbf{h}_i]\right)^2 \quad \ldots \quad (9)$$

where

$$\mathbf{h}_i = \mathbf{A}^{-1}\,\mathbf{S}_i \quad \ldots \quad (10)$$

$[\mathbf{h}_i]$ is the closest integer to $\mathbf{h}_i$

Because of the relatively large errors in the scattering vectors Si, the initial estimate of the orientation will not be very precise and attempts to index the actual scattering vectors Si or even the longer difference vectors Sij using equation (10) will fail. To prevent this, only relatively short difference vectors Sij are used in the first stages of refinement, and the number of difference vectors included is gradually increased and finally the scattering vectors themselves are used. Clearly the inclusion of the longer scattering vectors will improve the accuracy of the refined parameters, but conversely it is the use of very short difference vectors that is crucial to the success of the initial indexing.

4. Auto-indexing when cell parameters are unknown

In general, auto-indexing in the case of an unknown cell is only feasible when data from a wide Φ rotation or several narrow Φ rotations is available (i.e. the 3D rather than the 2D case). The first stages of the procedure are identical up to the calculation of the list of difference scattering vectors Sij. At this point three short, non-coplanar seed vectors are selected with a minimum angle between them (typically 40°). Since the unit cell is unknown, arbitrary indices can be assigned to the three seed vectors (typically (100), (010) and (001)). This assignment is used to generate an initial estimate of the setting matrix **A**, which can be used to index other scattering vectors (initially using difference vectors and then going on to use scattering vectors). The initial orientation and cell parameters can be refined in the same way as described for the case of a known cell (equation 9 and 10).

An additional step is required in this case to produce a final cell, as the cell chosen by the initial assignment of indices to the seed vectors will not necessarily correspond to the optimum choice of unit cell. In particular it may not reflect the symmetry of the reciprocal lattice (as there is no information available on reflection intensities at this stage). One of the auto-indexing procedures in MADNES (due to Paul Tucker) performs a search for 2-fold axes in the resulting lattice to facilitate the correct identification of the lattice symmetry. The user is finally presented with a list of alternative solutions (corresponding to different Bravais lattices) from which one is selected. The true symmetry can, of course, only be determined when integrated intensities are available.

5. Some Practical Considerations

The success of auto-indexing and refinement depends ultimately on the accuracy of the derived scattering vectors Si (and in the case of a known cell, on the accuracy of the initial cell parameters). These in turn depend on the accuracy of the spot co-ordinates (X_d, Y_d,Φ). The algorithm used to locate spots must allow the detection of weak spots without introducing "false" spots arising from noise in the image. It should be able to cope with slightly split spots or unusual spot shapes. For strongly diffracting crystals, the presence of white radiation streaks can cause problems - a low resolution cut-off can help in these cases. Any spatial distortion introduced by the detector must be corrected. As already mentioned, the uncertainty of the individual spot Φ values introduces significant error in coarse Φ slice images, which demands some sophistication in the algorithms used to identify clusters of reciprocal lattice points and in the indexing step following the initial determination of the orientation (e.g. see Kabsch, 1988a).

In order to convert the spot co-ordinates into scattering vectors, the crystal to detector distance and the direct beam position must be known Since the initial orientation is determined from difference scattering vectors, this step is less sensitive to errors in the direct beam position and may be successful, but the auto-indexing procedure may then fail on the subsequent indexing of the scattering vectors themselves. It is therefore good practice to record the direct beam on the image or ensure that its position can be accurately determined (for example by recording powder rings from a sample at the crystal position).

In the event that the auto-indexing is not successful there are a number of possible approaches to take:

a) Examine the image(s) themselves. If the crystal is badly split, twinned or multiple then auto-indexing is very unlikely to succeed. In these cases it may be possible to manually edit the list of spot co-ordinates to select those arising from a single crystal. If the images are very weak there may not be a sufficient number of *short* differences vectors Sij to allow identification of the seed vectors. In this case, the threshold for selecting spots should be reduced or the images re-collected with a longer exposure time. As mentioned earlier, white radiation streaks from strongly diffracting crystals can cause problems. These can usually be eliminated with a low resolution cut-off.

b) Check the direct beam position and the detector distance (and the wavelength for synchrotron data).

c) If possible use data from a larger rotation range or a different Φ segment. However it should be realised that if the crystal has slipped during data collection this may make matters worse rather than better. Thus if a fairly large Φ range was used initially, (e.g. 20°) it may be worth trying a *smaller* range.

d) Change the selection of reflections used in auto-indexing by increasing or decreasing the threshold.

e) Check the lengths of the short difference vectors (these are listed by many programs) against those predicted from the unit cell parameters. If the difference is more than a few per cent, try using the cell parameters predicted by the auto-indexing as input parameters, or adjust the crystal to detector distance (or wavelength) to get a better match.

f) If possible, look at the difference vectors with a graphics display program. This is very effective in showing up twinned or multiple crystals.

g) Recollect the data!

6. Parameter Refinement

Once an orientation matrix and cell parameters have been derived from the auto-indexing, these parameters (and others) are refined further using a different algorithm. In many software packages this refinement is based on the program IDXREF (Nyborg et al., 1975) which was originally developed for processing film data collected by the rotation method. The parameters to be refined can be conveniently grouped into three classes:

a) Crystal parameters: cell parameters, crystal orientation and mosaic spread (isotropic or anisotropic).

b) Detector parameters: the detector position and orientation and (if appropriate) distortion parameters (e.g. the radial and tangential offsets for the Mar image plate scanner).

c) Beam parameters: the orientation of the primary beam and beam divergence (isotropic or anisotropic).

The refinement of these parameters is achieved by least-squares minimisation of two residuals; a positional residual:

$$\Omega_1 = \sum_i \omega_{ix}\left(X_i^{calc} - X_i^{obs}\right)^2 + \omega_{iy}\left(Y_i^{calc} - Y_i^{obs}\right)^2 \quad \ldots \quad (11)$$

where X and Y are the spot co-ordinates on the detector, and an angular residual:

$$\Omega_2 = \sum_i \omega_i \left[\left(R_i^{calc} - R_i^{obs}\right)/d_i^*\right]^2 \quad \ldots \quad (12)$$

where R_i^{calc}, R_i^{obs} are the calculated and observed distances of the reciprocal lattice point d_i^* from the surface of the Ewald sphere. R_i^{obs} is obtained from the Φ centroid if fine Φ slices have been used. For coarse Φ slices, the reciprocal lattice point is either assumed to lie exactly on the Ewald sphere at the midpoint of the rotation, or for partially recorded reflections its position is estimated from the degree of partiality of the reflection (i.e. the way in which the total intensity is distributed between the two abutting images). This latter approach, known as post-refinement because it requires a knowledge of the integrated intensities, requires a model for the rocking curve, and permits refinement of either crystal mosaicity or beam divergence. For fine phi slices the mosaic spread or beam divergence is estimated from the observed reflection width in Φ.

The refinement strategy depends on how the data has been collected. If fine Φ slices have been used, accurate Φ centroids and co-ordinates (X,Y) are available for most strong reflections (excluding those very close to the rotation axis) and both residuals (Ω_1, Ω_2) can be minimised simultaneously using a suitable selection of reflections (strong and evenly distributed over the detector and in Φ). Problems arising due to correlations of different parameters can be avoided either by fixing some parameters or by the use of eigen-value filtering. These problems can be particularly serious for low resolution data, where there is a strong correlation between crystal to detector distance and the cell parameters, or for an offset detector where there is a high correlation between the detector swing angle and the (horizontal) primary beam co-ordinate. If only a narrow Φ range of reflections is used in the refinement then some unit cell parameters will be poorly defined and may be correlated with the crystal setting angles, and there will also be a strong correlation between the detector orientation around the X-ray beam (CCOMEGA in MOSFLM/RIGAKU software, Tau1 in MADNES) and the crystal setting angle around the beam. In such circumstances the refined parameters may assume physically unrealistic values, but this will not necessarily impair the accuracy of the prediction of reflection positions and widths.

If the data is collected with coarse Φ slices, only fully recorded reflections will give accurate spot positions (X,Y), and accurate Φ centroids can only be determined for partially recorded reflections. In the IDXREF program (MOSFLM package) the two residuals are therefore minimised independently. Only the detector parameters are refined when minimising the positional residual, and only cell orientation and optionally beam parameters are refined against the angular residual. The same is true when the refinement is carried out in MOSFLM itself, which uses exactly the same algorithms as IDXREF. This approach does have the advantage that the accuracy of the refined cell parameters does not depend on the accuracy of the crystal to detector distance or direct beam position, providing these are known sufficiently well to allow correct indexing of the reflections. It is the accuracy of the reflection Φ centroids which is important, and this in turn depends on the mosaic spread, beam divergence and the model of the rocking curve.

7. Accuracy of the A matrix

In order to reduce systematic errors in reflection integration using profile fitting methods to an acceptable level, reflection positions need to be predicted to an accuracy of about 5% of

the spot size on the detector. For example if the spot width is 10 pixels, then the desired accuracy is 0.5 pixels (i.e. 50-75μ). Thus the rms positional error (at least for strong reflections) should be in this range.

The angular residual will typically be a few hundredths of a degree, but this will depend critically on the combined beam divergence and mosaic spread, the larger the reflection width, the larger this angular residual can be without significantly affecting the quality of the reflection integration. For the IDXREF program, the angular residual is typically 0.2*(mosaic spread + beam divergence) when processing "still" images, and about half this value when using post-refinement (in MOSFLM or POSTCHK).

Acknowledgements

I am very grateful to Phil Evans for many useful discussions during the preparation of this manuscript.

References

Busing, W.R. and Levy, H.A. Acta Cryst. *22* (1967) 457-464

Higashi, T.J. Appl. Cryst. *23* (1990) 253-257

Howard, A. Proc. EEC Cooperative Workshop on Position-Sensitive Detector Software (Phase I & II), LURE, Paris, 16 May - 7 June 1986, pp.89-94

Howard, A.J., Gilliland, G.L., Finzel, B.C., Poulos, T.L., Ohlendorf, D.H. and Salemme, F.R. J. Appl. Cryst. *20* (1987) 383-387

Kabsch, W.J. Appl. Cryst. *21* (1988a) 67-71

Kabsch, W. J. Appl. Cryst. *21* (1988b) 916-924

Kim, S. J. Appl. Cryst. *22* (1989) 53-60

Messerschmidt, A. and Pflugrath, J.W. J. Appl. Cryst. *20* (1987) 306-315

Nyborg, J., Wonacott, A.J., Thierry, J.C. and Champness, J.N. (1975) unpublished notes

Prangé, T. Proc. EEC Cooperative Workshop on Position-Sensitive Detector Software (Phase III), LURE, Paris, 12-19 November 1986, pp. 12-22

Tanaka, I., Yao, M., Suzuki. M., Hikichi, K, Matsumoto, T., Kozasa, M. and Katayama, C. J. Appl. Cryst. *23* (1990) 334-339

Tucker, P. Proc. EEC Cooperative Workshop on Position-Sensitive Detector Software (Phase III), LURE, Paris, 12-19 November 1986, pp. 9-11, 23-27

Vriend, G and Rossmann M.G. J. Appl. Cryst. *20* (1987) 338-343

AUTOINDEXING in MADNES

by
J. W. Pflugrath
Cold Spring Harbor Laboratory
P. O. Box 100
Cold Spring Harbor, NY 11724
24 September 1987
Revised: 1 May 1989

This article describes how autoindexing has been implemented in MADNES -- the device-independent software for area detector systems in macromolecular crystallography. Most of the ideas presented here were published previously (Busing & Levy, 1967; Sparks, 1976; Sparks, 1982; Howard, 1986). The purpose of autoindexing is to provide a rough estimate of the unit cell parameters and the orientation matrix from the centroids of a few dozen reflections. These rough estimates can then be refined in another part of the program with an eigenvalue-filtered least squares algorithm. From the refined orientation matrix, the program will be able to predict the scattering vector of any reflection of index **h**:

$$\mathbf{s} = \mathbf{UBh} \tag{1}$$

where **s** is the scattering vector,

$$\mathbf{s} = \begin{matrix} s_x \\ s_y \\ s_z \end{matrix} \tag{2}$$

and **h** is the reflection index,

$$\mathbf{h} = \begin{matrix} h \\ k \\ l \end{matrix} \tag{3}$$

B defines the crystal cartesian axes, and **U** the orthogonal matrix relating the crystal cartesian axes to the diffractometer coordinate system (Busing & Levy, 1967). **U** is often designated the orientation matrix, but sometimes so is the matrix **UB**. In MADNES, the matrix **U** is decomposed into rotations or missettings around the three diffractometer axes:

$$\mathbf{U} = \phi_z\phi_y\phi_x \tag{4}$$

where ϕ_x, ϕ_y, ϕ_z designate the right-handed rotation matrices:

$$\phi_x = \begin{matrix} 1 & 0 & 0 \\ 0 & \cos X & -\sin X \\ 0 & \sin X & \cos X \end{matrix} \tag{5}$$

$$\phi_y = \begin{matrix} \cos Y & 0 & \sin Y \\ 0 & 1 & 0 \\ -\sin Y & 0 & \cos Y \end{matrix} \tag{6}$$

$$\phi_z = \begin{matrix} \cos Z & -\sin Z & 0 \\ \sin Z & \cos Z & 0 \\ 0 & 0 & 1 \end{matrix} \tag{7}$$

From Busing & Levy (1967) equation (3):

$$\mathbf{B} = \begin{matrix} a^* & b^*\cos\gamma^* & c^*\cos\beta^* \\ 0 & b^*\sin\gamma^* & -c^*\sin\beta^*\cos\alpha \\ 0 & 0 & c^*\sin\beta^*\sin\alpha \end{matrix} \tag{8}$$

Thus, the job of autoindexing is to find the matrix **UB** and thereby derive the unit cell parameters and the three missetting angles.

In general, the three-dimensional centroids of a few dozen reflections are determined by the FIND routine. Two dimensions arise from the detector coordinates XD and YD, while the third

dimension is the rotation angle φD around an arbitrary rotation axis which intersects the crystal. The reflection centroids (XD, YD, φD) must be converted into scattering vectors in the diffractometer coordinate system. If the *i*th scattering vector is designated $\mathbf{s}_i$, then

$$\mathbf{s}_i = f(XD_i,\ YD_i,\ \varphi D_i,\ \text{rotation\_axis},\ \text{detector\_position},\ \text{beam\_direction},\ \lambda) \tag{9}$$

From the list of n scattering vectors ($\mathbf{s}_i$, $i = 1, n$) a set of difference vectors is calculated. If there are n scattering vectors, there will be $n(n-1)/2$ difference vectors. The differences of the scattering vectors are used in order to provide shorter vectors for indexing, to introduce redundancy in vector coordinates, and to reduce the influence of incorrect detector positional parameters. As an example, if a reflection 10 1 0 gives rise to the scattering vector $\mathbf{s}_j$ and the reflection 11 1 0 gives scattering vector $\mathbf{s}_k$, then the difference vector $\mathbf{d}_{jk}$ would have an index -1 0 0. The difference vector $\mathbf{d}_{lm}$ derived from scattering vectors $\mathbf{s}_l$ and $\mathbf{s}_m$ with indices 8 1 5 and and 9 1 5, respectively, also has index -1 0 0, thus we have two estimates for -1 0 0 which may be combined to produce a better estimate of the vector. The difference vectors are sorted on increasing length. Since $\mathbf{d}_{jk}$ is the same as $-\mathbf{d}_{kj}$, we can require all the difference vectors to lie in a single hemisphere (for example all z coordinates > 0). Difference vectors which are sufficiently close spatially are grouped together and averaged into a single vector.

Three non-coplanar seed vectors are chosen from this new set of averaged vectors which presumably have the smallest error in their measurement. The algorithm in MADNES selects the three shortest averaged vectors above a user-specified minimum vector length and separated by at least a user-specified angle. Another criteria is also included: the averaged difference vector must arise from at least n difference vectors, where n is a user-specified number which defaults to 3. If the length is relatively short, then the indices will be small and easily indexed, but the estimates of the cell parameters and orientation matrix will be worse than if a longer minimum vector length were chosen. If the user-specified angle is large (>75°) then a better estimate of the orientation matrix will be found, but one may not find three vectors which meet this condition, especially if all the original reflection centroids were derived from a limited rotation range. In practice, an angle of at least 20° between pairs of seed vectors has been sufficient for a rotation range of about 2°.

If we designate the three seed vectors as $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$ with corresponding indices $\mathbf{h}_1$, $\mathbf{h}_2$ and $\mathbf{h}_3$ then we have from Eq. 1:

$$\mathbf{s}_1 = \mathbf{U}\mathbf{B}\mathbf{h}_1 \tag{10}$$
$$\mathbf{s}_2 = \mathbf{U}\mathbf{B}\mathbf{h}_2 \tag{11}$$
$$\mathbf{s}_3 = \mathbf{U}\mathbf{B}\mathbf{h}_3 \tag{12}$$

These can be combined by forming 3 x 3 matrices from the column vectors $\mathbf{s}_i$ and $\mathbf{h}_i$

$$\mathbf{S} = (\mathbf{s}_1\ \mathbf{s}_2\ \mathbf{s}_3) \tag{13}$$
$$\mathbf{H} = (\mathbf{h}_1\ \mathbf{h}_2\ \mathbf{h}_3) \tag{14}$$

to yield the following equation:

$$\mathbf{S} = \mathbf{U}\mathbf{B}\mathbf{H} \tag{15}$$

We make one restriction; ($\mathbf{s}_1\ \mathbf{s}_2\ \mathbf{s}_3$) must form a right-handed system. In other words *det*S must be greater than 0 (it cannot be 0 since $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$ are non-coplanar). This presents no problems since if *det*S < 0, we set $\mathbf{s}_1 = -\mathbf{s}_1$.

Now let

$$\mathbf{U}\mathbf{B} = \mathbf{A} \tag{16}$$

then Eq. 15 becomes

$$\mathbf{S} = \mathbf{A}\mathbf{H} \tag{17}$$

and pre-multiplying both sides by A^{-1} gives

$$A^{-1}S = H \tag{18}$$

Post-multiplying both sides of this equation by S^{-1} yields

$$A^{-1} = HS^{-1} \tag{19}$$

If we now decompose A^{-1} and H into row vectors (Sparks, 1982):

$$A^{-1} = \begin{matrix} r_1 \\ r_2 \\ r_3 \end{matrix} \tag{20}$$

$$H = \begin{matrix} g_1 \\ g_2 \\ g_3 \end{matrix} \tag{21}$$

we can re-write Eq. 19 as three separate equations:

$$r_1 = g_1S^{-1} \tag{22}$$
$$r_2 = g_2S^{-1} \tag{23}$$
$$r_3 = g_3S^{-1} \tag{24}$$

We can determine S^{-1} and we can loop over small integers to form the integer triplets g_1, g_2, g_3 in order to calculate r_1, r_2, r_3. Recalling Eq. 18, a solution is found when

$$r_i \cdot s_j = k \tag{25}$$

where s_j is a column in S or for that matter any difference or scattering vector and k is an integer.

Equations 32 through 36 of Busing & Levy (1967) tell us that

$$|r_1| = a \tag{26}$$
$$|r_2| = b \tag{27}$$
$$|r_3| = c \tag{28}$$

$$r_2.r_3 / |r_2||r_3| = \cos\alpha \tag{29}$$
$$r_1.r_3 / |r_1||r_3| = \cos\beta \tag{30}$$
$$r_1.r_2 / |r_1||r_2| = \cos\gamma \tag{31}$$

so that if the unit cell parameters are known roughly we can apply these criteria to a solution in addition to that provided by Eq. 25.

In the autoindexing algorithm used, the inverse of matrix S (formed from the seed vectors, Eq. 13) is calculated, then integer triplets are generated by using 0's and +-1's first, then combinations of increasingly larger integers from which the row vector r_1 is calculated. If $|r_1|$ is within preset limits a_{min} and a_{max}, then an 'integerness residual' based on Eq. 25 is calculated:

$$R_{Ij} = \left(\sum_{i=1}^{n} (P_{ij} - [P_{ij}])^2 / n \right)^{1/2} \tag{32}$$

where $P_{ij} = r_j.s_i$ (Eq. 25) for all the scattering or difference vectors s_i, $[P_{ij}]$ denotes the integer nearest P_{ij} and n is the total number of scattering or difference vectors. If R_{I1} is below a user-specified minimum (default 0.15), then g_1 gives the h indices of the 3 non-coplanar vectors in S.

Guesses for g_2 are made in the same way to determine a valid solution for r_2 which results in b. With g_1 and g_2 a further check is made to see if γ falls within the user-specified limits γ_{min} and γ_{max}. When valid answers for g_1 and g_2 are made, then guesses for the integer triplet g_3 are made with [c_{min},

c_{max}], [α_{min}, α_{max}] and [β_{min}, β_{max}] used as criteria along with the integerness residual test. A final test is that *det*H must be positive.

Limits on the guesses used for g_1, g_2 and g_3 are based on a_{max}, b_{max} and c_{max}. All valid answers are saved and sorted on increasing overall R_{Io}, where

$$R_{Io} = ((R_{I1}^2 + R_{I2}^2 + R_{I3}^2) / 3)^{1/2} \qquad (33)$$

For each solution, the missetting angles around the X, Y and Z axes are calculated as follows. We know from Eqs. 16 and 20 that

$$\begin{matrix} r_1 \\ r_2 \\ r_3 \end{matrix} = \quad A^{-1} = \quad (UB)^{-1} \qquad (34)$$

so A^{-1} is inverted to get UB. With the unit cell parameters derived from Eqs. 26 - 31, the normal direct to reciprocal transformations and Eq. 8 we form the B matrix. We invert B to give B^{-1} which is then post-multiplied with UB to give U. The U matrix is then decomposed into the missetting angles around the diffractometer axes X, Y, Z by comparison of elements of the matrix resulting from the right-hand side of Eq. 4:

$$U = \begin{matrix} \cos Y\cos Z & -\cos X\sin Z + \sin X\sin Y\cos Z & \sin X\sin Z + \cos X\sin Y\cos Z \\ \cos Y\sin Z & \cos X\cos Z + \sin X\sin Y\sin Z & -\sin X\cos Z + \cos X\sin Y\sin Z \\ -\sin Y & \sin X\cos Y & \cos X\cos Y \end{matrix} \qquad (35)$$

The 20 solutions with the lowest integerness residual are listed in order of increasing R_{Io} and the user selects one. Often there is only a single solution and its equivalents. The chosen solution is used as a starting point in the refinement algorithm which ultimately gives the accurate orientation matrix and unit cell parameters.

References

Busing, W. R. & Levy, H. A. (1967) *Acta Cryst.* **22**, 457-464.

Howard, A. (1986) *Proc. EEC Cooperative Workshop on PSD Software* **2**, 89-94.

Sparks, R. A. (1976) in *Crystallographic Computing Techniques* (F. R. Ahmed, ed.) Copenhagen, Munskgaard, 452-467.

Sparks, R. A. (1982) in *Computational Crystallography* (D. Sayre, ed.) Oxford, Clarendon Press, 1-18.

Not cited but useful

Kabsch, W. (1988) *J. Appl. Cryst.* **21**, 67-71.

Kim, S. (1989) *J. Appl. Cryst.* **22**, 53-60.

Oscillation data reduction program

Zbyszek Otwinowski

Department of Molecular Biophysics and Biochemistry, Yale University

Program Denzo allows data reduction of single crystal oscillation images. Diffraction data are indexed, cell and detector parameters are refined and reflections are integrated by a weighted profile fitting algorithm. The dependence of the precision of the integrated data on the assumptions made in data reduction programs is described.

The program Denzo integrates reflections from single crystal diffraction data measured on film or phosphofluorescence Image Plate (IP) detector. Data can be collected by oscillation, Weissenberg or precession method. Detector can be either flat or cylindrical. Detector readout can be either in rectilinear or spiral, converted to rectilinear, coordinate system. Program allows for random changes in position and sensitivity of the detector between consecutive exposures.

Analysis and reduction of the single crystal diffraction data consists of six major steps:

1) Visualization of the original (unprocessed) detector data.
2) Indexing of the diffraction pattern.
3) Refinement of the crystal and detector parameters.
4) Integration of the diffraction maxima.
5) Finding relative scale factors between measurements.
6) Merging and statistical analysis of the measurements related by space group symmetry.

Visualization of the diffraction space

Program Denzo can reduce diffraction data that form an image of the 3-dimensional reciprocal space. Such image is collected as a series of 2-d images, each of them representing different (curved) slice of the reciprocal space. For the program to integrate diffraction maxima they have to be separated in the individual images. The distortion of the image is function of the data collection method, diffraction geometry and characteristics of the detector. For the data reduction to be successful, the distortion of the reciprocal space as view by the detector has to be correctly accounted by the program. The distortion of the image of the reciprocal space can vary even with the same detector being used. The position of the detector, X-ray wavelength, oscillation range, pixel size, gain and exposure level can all significantly affect the diffraction image and in some cases make the data difficult or impossible to process.

One should start data reduction with careful inspection of the data in their original form. One should check if the diffraction maxima are resolved, detector placed correctly for the right resolution range and if the exposure level is appropriate. Many problems with the detector can be discovered and corrected before the whole data set is collected. If there are many diffraction maxima in the image that form a pattern characteristics for a diffraction from a single crystal the next step is finding a crystal lattice that accounts for such pattern.

Autoindexing

Program Denzo offers two methods to index the diffraction pattern, automatic and interactive. In most cases the automatic method is used due to its simplicity. The automatic method starts with peak search. One needs to find large number of peaks in the automatic method. For a small molecule crystal 20 diffraction maxima are often sufficient, whereas for a macromolecule crystal 100 is often enough. Current version of the program accepts peaks for autoindexing only from a single oscillation image. It is important that the oscillation range be small enough so the lunes formed by diffraction peaks are resolved. If the previous condition is not satisfied reflections can have more than one index consistent with a particular position on the detector. The next step in autoindexing is mapping diffraction maxima to the reciprocal space. Because the precise angles at which reflections diffract are unknown the center of the oscillation range value is used instead.

The autoindexing is based on a novel algorithm that does complete search off all possible indexing of all reflections simultaneously, one index at a time. Finding one index (for example **h**) of all reflections is equivalent to finding one real space direction (in this case **a**) of the crystal, for this reason such indexing can be called "real space indexing".

The real space indexing method is quite robust because it assignees indexes to all reflection simultaneously. Small percentage of incorrectly identified diffraction maxima usually does not affect the method. Unlike reciprocal space indexing methods the real space indexing is insensitive to how many short difference vectors can be created from the peak search list. The autoindexing in Denzo is based on complete search of all possible real space vectors within some reasonable length range. Advantage of a complete search is that it is not dependent on previous knowledge of the crystal unit cell.

After the search for real space vectors is completed, program finds three best linearly independent vectors that would index all peaks and generate crystal unit cell with minimal volume. Such three vectors are unlikely to form a standard basis for description of the unit cell. The process of finding a standard basis is called "cell reduction". Program follows the definitions in the International Tables and finds best cells for all 14 Braveis lattices. The best triclinic lattice that fits the peak search list has to be distorted to fit into any higher symmetry lattice. The distortion index is a guide to which lattice type is consistent with observed positions of the reflections. Due to experimental errors the fit is never perfect for correct crystal lattice. Sometimes the observed reflections can be fitted into higher symmetry lattice than one defined by space group symmetry. Such condition is called lattice (or metric tensor) pseudo symmetry and can make lattice determination and spot indexing quite complicated. Program Denzo calculates the distortion index for all 14 Braveis lattices. It is up to the user to define lattice and space group symmetry, as the program at this stage of the calculation cannot distinguish lattice symmetry from pseudo symmetry.

Interactive indexing

There are two alternatives to autoindexing in Denzo. One is to input an orientation matrix from another autoindexing programs like REFIX (Kabsch. 1988). Alternatively, the user can determine the approximate orientation by an iterative process resembling that used for aligning precession photographs, except that the predicted pattern is being changed in every iteration,

rather than the diffraction image. The manual indexing is helped by graphics feedback. The interactive indexing process is also helped by the ability of the program to define the current orientation relative to any (principal or higher order) zone being in the center of the image. This is particularly useful in centered space groups and in characterizing unknown lattices. Manipulation of the predicted patters can be used to simulate diffraction experiments. Simulation can identify potential data reduction problem before data collection even starts. Simulation of the diffraction patter is also invaluable tool in teaching crystallography.

Failure of indexing

Autoindexing is based of assumption that spots are correctly mapped from detector coordinates to diffraction (reciprocal) space. Due to carelessness or lack of knowledge of detector coordinates wrong parameters can be input to the program, nevertheless program can sometimes find a indexing consistent with the incorrectly specified detector geometry. If the error is in the position of the origin of the diffraction space (same as the position of the direct beam on the detector) the program can shift the indexing of the diffraction pattern by an integer vector. Such misindexing can be totally self consistent till scaling of the diffraction data. In fact, the program will shift the origin of the diffraction space to the nearest grid point of the best primitive lattice it finds. Initial error in the direct beam position by 0.48 reflection separation will lead to correct indexing, error by 0.52 reflection separation will misindex the diffraction pattern by one index. Misindexing by one index will never get corrected by subsequent refinement of crystal and detector parameters, Misindexing may or may not, produce poor agreement between predicted and observed positions of reflections.

Refinement of the crystal and detector parameters.

The initial values of crystal and detector orientation parameters may require refinement. The refinement process typically is quite simple for a series of images collected on a well characterized on-line detector but can be more complicated if both the detector and the crystal orientation are only crudely known. The refinement strategy is fully under user control and can consist of several steps. In each step the user defines the resolution limits and parameters to be fitted. Both detector and crystal parameters are fit together using data only from one image. This avoids problems presented by crystal slippage. Program Denzo refines all parameters by a rapidly converging least squares method. Refinement can be unstable due to high correlation among parameters. In such cases eigenvalue filtering removes the most correlated components from refinement. If eigenvalue filtering becomes active, users are encouraged to set some of the correlated parameters to known values even if these are only approximate. Fortunately, if a parameter cannot be precisely determined from a diffraction image, use of a somewhat incorrect value of such parameter will not significantly affect the prediction of the diffraction pattern. In such case integration can proceed without hindrance.

One might think that integration of spot intensity requires only its approximate coordinates, as the summation of the peak area is not affected by its precise placement. However, accurate prediction of spot positions is often necessary for spot integration. In many cases the detector is placed as close as possible to the crystal consistent with complete separation of diffraction spots. In such case small errors in spot prediction would make one diffraction peak to

intrude upon the predicted background of the next peak. The most important reason for accurate position prediction results from the application of profile fitting. Profile prediction calculates an average of profiles shifted by the predicted separation between spots. If the predicted positions have errors, the average profile will be broadened and/or displaced from the actual profile of the reflection. The consequences of this effect on calculated intensity will be quantified later.

Errors in the prediction of spot position affect the accuracy of the summed intensity by a different mechanism. If predictions do not match the peak position exactly, one has to enlarge the expected spot area to sum the intensity of the whole spot. Enlargement of the predicted spot area increases the total background to be subtracted. Larger background has larger variance that adds to the measurement variance. Auto centering of the spot area can compensate for errors in the prediction, but it only works for strong spots and if applied individually to every spot would seriously bias the calculated intensity. Some programs do auto centering by averaging the local deviations between observed and predicted position. This is not done explicitly in Denzo, however, the profile prediction algorithm used in Denzo has a similar effect.

The detector and crystal parameters are refined by a least squares method that minimizes deviation of the reflection centroids from their predicted positions. Such refinement is seriously deficient when applied to a single oscillation image, since one crystal rotation parameter is undefined and others are highly correlated and/or poorly defined. To overcome this problem Denzo adds another term, in which the intensity of partial reflections is compared to the predicted partiality times an average intensity in the same resolution range. This residual is very similar to the one used in 'postrefinement', except that the error in the predicted fully recorded intensity is very large, equal to the expected intensity. Concomitant position and partiality refinement used in Denzo is stable and very accurate. A major benefit is uniform treatment of detector and crystal variables in the whole refinement process. The most important effect of combined positional and partiality refinement is in reduced correlation between detector and crystal parameters.

Description of real detectors

Correct understanding of detector geometry is essential to accurate positional refinement. Unfortunately most detectors deviate from perfectly flat or cylindrical geometry. These deviations are detector specific. Primary sources of error include misalignment of the detector position sensors (MAR, R-AXIS), non-planarity of film or IP during exposure or in the scanner, inaccuracy of the wire placement and distortions of the position readout in MWPC, optical distortion (can also be due to magnetic field acting upon image intensifier) in the TV or CCD based detectors. If the detector distortion can be parametrized, then these parameters should be added to the refinement. For example, in the case of the spiral scanners there are two parameters describing the end position of the scanning head. In the perfectly adjusted scanner these parameters would be zero. In practice, they may deviate from zero by as much as 0.5 mm. Such misalignment parameters can correlate very strongly with other detector and crystal parameters. If the program does not have ability to describe detector distortions, other parameters such as unit cell and detector to crystal distance will be systematically wrong. With film and IP handled manually in cassettes, the biggest problem is in keeping the detector flat during exposure and subsequent scanning. In manual systems it is much harder to model the possible departures from ideal flat or cylindrical geometry, and most programs make limited attempts to correct for such

distortions. Non-ideal film/IP geometry is one of the main factors behind variable quality of data collected with manual systems.

Denzo accepts the diffraction data formatted in any rectilinear coordinate system. Denzo has flexible definition of mapping reciprocal space on the detector coordinates so any detector geometry, flat or cylindrical can be handled. The detector can be in any position in space and in any angular orientation. In particular, for cylindrical geometry, there is no requirement for the crystal to be on the cylinder axis, nor that the scanning coordinate system be parallel and perpendicular to the cylinder axis. All detector parameters can be refined. They are described as series of rotations, translations and scale factors rather than as one transformation matrix. This allows for each of the detector parameters to be individually fixed or refined according to user specification.

Correction for the non-linear response function of a detector to the photon flux is applied internally in the program so that it can read original data without the need for any transformations (this does not yet apply to the data from spiral scanners). Pixel values can represent two special cases: no measurement or detector overload. Overloaded pixels are assumed to be close to the center of gravity of the diffraction spots and as such they are used in determining spot centroids. Pixels that are either overloaded or had no measurement are ignored in calculating the spot intensity by profile fitting method, but the existence of such pixels in the spot area is flagged by negation of the sigma estimate. Profile-fitted intensities seem to be reliable even if such pixels exist in the spot area. Denzo has no automatic facility for definition of an active area of the detector. For some detector types there is separate program that automatically finds cassette and the beam stop shadows. This separate program uses the position of the cassette shadow as a substitute for the fiducial marks on the image plate. This facility has been very useful with manually loaded cassettes at synchrotron beamlines where there was no option to take a direct beam exposure. On-line Image Plate scanners do not require fiducials, as they have fixed relation between the scanned and exposed positions.

Profile fitting

To calculate the diffraction intensity, the detector background must be estimated and then subtracted from the reflection profile. The standard method to estimate the background value is to calculate an average detector signal in the neighborhood of a specific reflection. Some programs assume that background is a linear function of detector coordinates. Denzo makes the simpler assumption that the background is constant around every spot. This usually does not affect intensity measurements because the background is measured symmetrically around each reflection. With symmetrical definitions of the areas where background and peak values are estimated, the effect of linear background variation is almost zero. However, profile fitting is slightly affected by linear background variations. A small effect can also be due to background editing that removes pixels asymmetrically from background measurement area. Denzo removes pixels from the background area in two cases: when they have been flagged as no measurement by an auxiliary program or when they are in the spot area of another nearby reflection. Removal of pixels is based on predicted, rather than measured, spots positions.

Profile fitting is a two-step process: First, the profile is predicted based on the profiles of other reflections. Second, the observed profile M_i is describes as sum of the background B_i and

the predicted profile p_i times a constant (index i represents all pixels in 1,2 or 3-dimensional profile). If the predicted profile is normalized $\sum_i p_i = 1$, then the constant is the fitted intensity I. The calculation of the fitted intensity has an equivalent but simpler explanation: each pixel provides an estimate of spot intensity $\frac{M_i - B_i}{p_i}$ with variance $\frac{V_i}{p_i^2}$. Variance is a function of expected signal in a pixel, in the case of a counting detector $V_i = \langle M_i \rangle = Ip_i + B_i$. A profile fitted intensity is then simply a weighted average of all observations:

$$I = \frac{\sum \frac{p_i^2}{V_i} \frac{(M_i - B_i)}{p_i}}{\sum \frac{p_i^2}{V_i}} = \frac{\sum \frac{p_i (M_i - B_i)}{V_i}}{\sum \frac{p_i^2}{V_i}} \qquad \text{(Eq. 1)}$$

Equivalent, but expressed as a solution to best fit problem, approach was first implemented by Diamond in 1969 for the 1 dimensional case. However, in 1974 Ford proposed a simplified formula where V_i is constant. This was based on mistaken idea that variance of the optical density value of the X-ray exposed film is independent of the degree of X-ray exposure. Equation 1 become simpler:

$$I = \frac{\sum p_i (M_i - B_i)}{\sum p_i^2} \qquad \text{(Eq. 2)}$$

Most of the subsequent programs followed the formulation of Ford rather than of Diamond, even when applied to data collected with counters or IP. The unweighted formula proposed by Ford works quite well where peak spot intensity is not much higher than background intensity. This happens more often with data collected on film, which has a high intrinsic background, or when crystals have low scattering power due to a very large unit cell, high solvent content or disorder. The unweighted profile fitting improves accuracy (compared to straight summation) of weak reflections but at the cost of reducing the accuracy of the strong ones. This observation did lead to a partial solution based on taking a weighted average between profile fitted and summed intensity, where the weight is function of reflection intensity.

Denzo uses the weighted formula (eq. 1). Years of experience with that formula show that it does not deteriorate the accuracy of strong, low resolution reflections. Thus, the previously observed problem with the unweighted formula is in weighting, rather than in the accuracy of the predicted profile.

Errors of the predicted profiles

The prediction of the spot profile has been based on three approaches: modeling of the spots in detector coordinates by an analytical function, averaging of spot profiles in detector coordinates, and averaging of the spot profiles in reciprocal space coordinates. None of these approaches has an intrinsic advantage over others, what is more important it is how well the detailed assumptions about profile shape and its variability match what happens during data collection. Denzo is based on the averaging of profiles in detector coordinates. It is different from other programs in that it averages profiles separately for each spot. This approach has two main

advantages: first, it chooses only nearby spots, ones with most similar profiles; second, additional shifts by a single pixel are introduced in Denzo to make the average profiles center on the position of the predicted reflection to eliminate the need for interpolation.

The profile prediction is never exact. Predicted profiles can be of different shape and displaced from the measured spot profile. Diamond analyzed the case of one dimensional Gaussian profiles and unweighted profile fitting formula. The important parameters are: w - root mean square (rms) width of the actual profile, f - root mean square (rms) width of the predicted profile. We can define relative change in reflection width square $\Delta^2 = (f^2 - w^2)/w^2$. For no displacement the fitted intensity will be wrong by a factor (Diamond, 1969):

$$\sqrt{1+\frac{\Delta^2}{2+\Delta^2}}$$

The averaging of profiles adds $r^2/3$ (r - raster size) to the f^2. Averaging will increase profile fitted intensity of most reflection by a constant multiplicative factor. The interpolation broadens profile by a factor dependent upon position of the predicted reflection relative to the pixel boundaries. Interpolation will also increase f^2, by number between zero and $r^2/2$. Profile fitted intensity will be affected differently by interpolation on different reflection. The intensities will increased on average, but also interpolation will add random noise to the reduced data.

In almost all cases errors due to inaccurate prediction of profiles are less significant than increased random error in summed estimate of integrated intensity. Profile fitted intensities have been used in MAD (Multiply Anomalous Dispersion) method that is most sensitive of all crystallographic techniques to the errors in the estimates of the diffracted intensities.

The data reduction of diffraction from macromolecular crystal is now a laboratory routine. The quality of the data is frequently short of what one would like to use in solving the crystal structure. To improve the data quality one has to plan experiment understanding data reduction process. The program Denzo allows for optimal treatment of large class of experimental data.

Diamond, R. (1969) *Acta Cryst.* A**25**, 43-55.
Ford, G. (1974) *J. Appl. Cryst.* **7**, 555-564.
Kabsch, W. (1988) *J. Appl. Cryst.* **21**, 67-71.

RECENT EXTENSIONS OF THE DATA PROCESSING PROGRAM XDS

by

WOLFGANG KABSCH

Max-Planck-Institut für medizinische Forschung, Abteilung Biophysik,
Jahnstraße 29, 6900 Heidelberg, Germany

1. INTRODUCTION

The availability of electronic area detectors and imaging plate systems has greatly increased the speed of data acquisition from macromolecular crystals. As compared with film, the new x-ray recording devices require much less exposure time which results in a drastic increase of the useful data that can be obtained from each crystal.

The program package XDS [1,2] has been specifically designed for automatic reduction of rotation data thereby exploiting the potential of the new hardware to increase the quality of the data. Recently, the capabilities of the program have been extended to handle area detector (Siemens/Nicolet) as well as imaging plate (Marresearch) data from crystals of unknown orientation, symmetry and cell constants. The ideas underlying these new program features are decribed below.

2. SPACE-GROUP DETERMINATION

Obtaining good crystals can be a rare event and in this case their symmetry and unit cell constants are usually unknown. As described below it is now possible to collect and process all x-ray data in the absence of this knowledge thereby extracting as much useful information as possible from the few available crystals. Space-group symmetry and cell constants are derived directly from the x-ray data: Strong diffraction spots occuring in the rotation pictures are automatically located and used to determine a reduced cell. The spots are then indexed with respect to the triclinic reduced cell and used for a refinement of the parameters controlling the x-ray diffraction geometry. Processing of all data frames is carried out as in the older version of XDS. Finally, a small number of possible space-groups is tested against the observed diffraction geometry and the integrated intensities after reindexing the reflections with respect to the appropriate conventional cell. A more detailed report of the underlying concept will appear elsewhere [3].

2.1 DETERMINATION OF A LATTICE BASIS

Reciprocal lattice vectors $\mathbf{p}_i^*$ $(i = 1, \ldots, n)$ corresponding to each automatically located spot X_i, Y_i, ϕ_i are formed as described earlier [1]. The determination of a basis underlying this lattice is carried out in three steps.

In the first step the list of given reciprocal lattice points is reduced to a small number m of low-resolution difference vector clusters $\mathbf{v}_\mu^*$ $(\mu = 1, \ldots, m)$[1]. f_μ is the population of a difference vector cluster $\mathbf{v}_\mu^*$, that is the number of times the difference between any two reciprocal lattice vectors $\mathbf{p}_i^* - \mathbf{p}_j^*$ is approximately equal to $\mathbf{v}_\mu^*$.

In the second step a best set of three linear independent vectors $\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3^*$ is selected from the list $\mathbf{v}_\mu^*$ $(\mu = 1, \ldots, m)$ that maximizes the function Q.

$$Q(\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3^*) = \max_{\substack{1 \le \mu_1 < \mu_2 < \mu_3 \le m \\ \mathbf{v}_{\mu_1}^* \cdot (\mathbf{v}_{\mu_2}^* \times \mathbf{v}_{\mu_3}^*) \neq 0}} Q(\mathbf{v}_{\mu_1}^*, \mathbf{v}_{\mu_2}^*, \mathbf{v}_{\mu_3}^*)$$

$$Q(\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3^*) = \sum_{\mu=1}^{m} f_\mu q(\xi_1^\mu, \xi_2^\mu, \xi_3^\mu)$$

$$q(\xi_1^\mu, \xi_2^\mu, \xi_3^\mu) = exp\Big(-2 \sum_{k=1}^{3} \{[\max(|\xi_k^\mu - h_k^\mu| - \epsilon, 0)/\epsilon]^2 + [\max(|h_k^\mu| - \delta, 0)]^2\}\Big)$$

$$\mathbf{b}_k \cdot \mathbf{b}_l^* = \begin{cases} 1, & \text{if } k = l; \\ 0, & \text{otherwise} \end{cases} \qquad \xi_k^\mu = \mathbf{v}_\mu^* \cdot \mathbf{b}_k \qquad \mathbf{v}_\mu^* = \sum_{k=1}^{3} \xi_k^\mu \mathbf{b}_k^*$$

$$h_k^\mu = \text{ nearest integer to } \xi_k^\mu$$

The absolute maximum of Q is assumed if all difference vectors can be expressed as small integral multiples of the best triplet. Deviations from this ideal situation are quantified by the quality measure q. The value of q sharply declines if the expansion coefficients ξ_k^μ deviate by more than ϵ from their nearest integers h_k^μ or if the indices are absolutely larger than δ. Excellent results have been obtained using $\epsilon = 0.05$ and $\delta = 5$.

In the third step the best vector triplet found above is refined against the difference vector clusters. From the refined vector triplet a reduced cell is derived as defined by Buerger [4]. In most cases the expansion coefficients for all difference vector clusters with respect to this reduced vector triplet assume nearly integral values. Occasionally it happens that some indices turn out to be very close to half-integers. In this case one of the vectors of the reduced triplet is replaced by the difference vector cluster with half-integers. The new triplet is Buerger-reduced. All difference vector clusters with integral indices with respect to the old base and the vector with half-integral indices now all have integral indices with respect to the new reduced base.

2.2 INDEXING OF LATTICE POINTS

Using the nearest integers $\mathbf{h}^i$ of $\mathbf{p}_i^* \cdot \mathbf{b}_k$ $(k = 1, 2, 3)$ as indices of the reciprocal lattice vectors $\mathbf{p}_i^*$ $(i = 1, \ldots, n)$ could easily lead to a misindexing because of inaccuracies in the basis vectors $\mathbf{b}_k$. A solution of this problem is provided by the *local indexing method* described in an earlier paper [1]. The idea is to use only small index differences $\mathbf{h}^{ij}$ between pairs of neighbouring lattice vectors. For large n the original implementation was found impractical and has been replaced by an alternative one which is highly efficient, both with respect to storage utilization and computing time.

The reciprocal lattice points can be considered as nodes of a tree. The tree connects the n points to each other with the connections as its branches. The length ℓ_{ij} of a possible branch between nodes i and j is defined here as

$$\ell_{ij} = 1 - exp\Big(-2\sum_{k=1}^{3}\{[\max(|\xi_k^{ij} - h_k^{ij}| - \epsilon, 0)/\epsilon]^2 + [\max(|h_k^{ij}| - \delta, 0)]^2\}\Big)$$

$$\xi_k^{ij} = (\mathbf{p}_i^* - \mathbf{p}_j^*) \cdot \mathbf{b}_k \qquad h_k^{ij} = \text{ nearest integer of } \xi_k^{ij} \qquad k = 1,2,3$$

ℓ_{ij} is 0 if none of the indices h_k^{ij} is absolutely larger than δ and the ξ_k^{ij} are integer values to within ϵ. Typical values of ϵ and δ are $\epsilon = 0.05$ and $\delta = 5$. Defining the length of a tree as the sum of the lengths of its branches, a shortest among all possible n^{n-2} trees is determined by the elegant algorithm described by Dijkstra [5]. Starting with arbitrary indices $0,0,0$ for the root node the local indexing method consists then of traversing the shortest tree thereby assigning each node visited the indices of its predecessor plus the small index differences between the two nodes. Bad points are recognized by large values of the lengths of their connecting branches and removed from the list. Finally, a constant offset is determined and added to the indices such that the centroids of the good observed reciprocal lattice points $\mathbf{p}_i^*$ and their corresponding grid vectors $\sum_{k=1}^{3} h_k^i \mathbf{b}_k^*$ are as close as possible.

2.3 DETERMINATION OF THE BRAVAIS LATTICE

The determination of the possible Bravais lattices is based upon the concept of the reduced cell whose metrical parameters characterize 44 lattice types as described in the *International Tables for Crystallography* [6]. The reduced cell is defined there by a number of conditions (inequalities) which must be satisfied by the components of its metric tensor. Each of the 44 lattice types is characterized by additional equality relations among the six components of the reduced cell metric tensor. Any primitive triclinic cell describing a given lattice can be converted into such a reduced cell (see for instance Andrews & Bernstein [7]). It is well known, however, that the derived reduced cell is sensitive to experimental error. Hence, the direct approach of first deriving the correct reduced cell and then finding the lattice type is unstable and may in certain cases even prevent the identification of the correct Bravais lattice.

Despite these difficulties a stable procedure has been developed that identifies all lattice types compatible with an observed lattice basis. Stability is obtained by avoiding any decision what the "true" reduced cell is. The essential ingredients of this procedure are: (a) a data base of possible reduced cells (b) a backward search strategy that finds the best fitting cell in the data base for each lattice type.

The data base is derived from a seed cell consisting of the three shortest linear independent observed lattice vectors sorted in increasing length. All cells of the same volume as the seed cell are included in the data base whose basis vectors can be linearly expressed in terms of the seed vectors by indices $-1, 0$, or $+1$. In addition, each of the three basis vectors of a cell must be of the same length – within 10%– as the corresponding seed vector. Now each unit cell in the data base is considered as a potential reduced cell although some of the defining conditions as given in chapter 9 in the *International Tables for Crystallography* may be violated. These violations are treated as being due to experimental error.

The backward search strategy starts with the hypothesis that the lattice type is already known and identifies the best fitting unit cell in the data base of possible reduced cells. Contrary to a forward directed search it is now always decidable which conditions on the components of the metric tensor of the reduced cell have to be satisfied. The total amount by which all these equality and inequality conditions are violated is used as a quality index. This measure is defined below for lattice type 20 mC testing a potential reduced cell $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ from the data base for agreement. If none of the conditions are violated a quality index $p_{20} = 0$ will result. Positive values indicate that some conditions are not satisfied.

$$A = \mathbf{b}_1 \cdot \mathbf{b}_1 \qquad B = \mathbf{b}_2 \cdot \mathbf{b}_2 \qquad C = \mathbf{b}_3 \cdot \mathbf{b}_3$$
$$D = \mathbf{b}_2 \cdot \mathbf{b}_3 \qquad E = \mathbf{b}_1 \cdot \mathbf{b}_3 \qquad F = \mathbf{b}_1 \cdot \mathbf{b}_2$$

$$\begin{aligned} p_{20}(\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3) = {} & max(0, A - B) + max(0, B - C) + max(0, 2|D| - B) \\ & + max(0, 2|E| - A) + max(0, 2|F| - A) + max(0, -D) \\ & + max(0, -E) + max(0, -F) + |B - C| + |E - F| \end{aligned}$$

All potential reduced cells in the data base are tested and the smallest value for p_{20} is assigned to lattice type 20. This test is carried out for all 44 possible lattice types using quality indices derived in a similar way from the defining conditions as listed in chapter 9 in the *International Tables for Crystallography* [6]. For each of the 44 lattice types thus tested the procedure described here returns the quality index, the conventional cell parameters, and a transformation matrix relating the original indices of the observed lattice points to the new indices with respect to the conventional cell. These index transformation matrices are derived from those given in Table 9.3.1 in the *International Tables for Crystallography* [6]. (Note that the matrix for lattice type # 17 mC given there is wrong: instead of $1\bar{1}0/110/\bar{1}0\bar{1}$ it should be $1\bar{1}0/\bar{1}\bar{1}0/\bar{1}0\bar{1}$.)

The quality index as defined for lattice type 20 is not particularly good since its contributing terms are not independent of each other. Despite these shortcomings it serves its purpose to clearly indicate lattice types which are in agreement with the observed lattice points. At this stage there is no automatic decision making; rather decisions are made by the crystallographer after all rotation pictures have been processed and integrated intensities are available. These decisions are then based on the rms-deviations between the observed and refined spot positions as well as on the R-factor statistics of symmetry-related reflection intensities.

3. INTENSITY ESTIMATION BY PROFILE FITTING

Originally, XDS has been developed for the Siemens/Nicolet electronic area detector to process sequences of adjacent rotation pictures with narrow oscillation ranges (typically 0.16°). This device has a relatively short data read-out time to allow this mode of operation. It is then possible to extract three-dimensional reflection profiles from the data frames which leads to accurate estimated intensities based on a fine-grained distinction between signal and background regions. In contrast to the electronic area detectors, imaging plate systems require a much longer read-out time which leads one to increase the oscillation range of each rotation picture to minimize their number. As a consequence, the representation of each reflection by its three-dimensional profile has been modified as described below.

3.1 REFLECTION PROFILE COORDINATES

The shape of its three-dimensional profile strongly depends on the specific path of the reflection through the Ewald sphere. As described in an earlier paper [2] for the case of rotation diffraction data, this effect can be eliminated by representing the profile with respect to a suitable coordinate system specific for each reflection. Let $\mathbf{n}$ be a unit vector perpendicular to the incident-beam wave vector $\mathbf{S}_0$, and $\mathbf{S} = \mathbf{S}_0 + \mathbf{x}$ the diffracted-beam wave vector when the Laue equations are satisfied. The basis vectors of the local coordinate system are $\mathbf{a}, \mathbf{b}, \mathbf{c}$; the origin is on the surface of the Ewald sphere at the endpoint of $\mathbf{S}$.

$$\mathbf{a} = \mathbf{n} \times \mathbf{S}/|\mathbf{n} \times \mathbf{S}| \qquad \mathbf{b} = \mathbf{S} \times \mathbf{a}/|\mathbf{S} \times \mathbf{a}| \qquad \mathbf{c} = (\mathbf{S} + \mathbf{S}_0)/|\mathbf{S} + \mathbf{S}_0|$$

The unit vectors $\mathbf{a}$ and $\mathbf{b}$ are tangential to the Ewald sphere while $\mathbf{c}$ is perpendicular to $\mathbf{x}$ and lies in the diffraction plane. What are the profile coordinates α, β, γ of a point in the neighbourhood of the reflection ? Assuming this point is located at X, Y on the detector and the crystal is rotated by $\phi' - \phi$ about the spindle axis $\mathbf{u}$ from the ideal angular position at ϕ when the Laue equations are satisfied, its profile coordinates are

$$\alpha = \mathbf{a} \cdot (\mathbf{S}' - \mathbf{S})/|\mathbf{S}| \qquad \beta = \mathbf{b} \cdot (\mathbf{S}' - \mathbf{S})/|\mathbf{S}| \qquad \gamma = \mathbf{c} \cdot (\mathbf{x}' - \mathbf{x})/|\mathbf{x}|$$

Here, the diffracted-beam wave vector $\mathbf{S}'$ is calculated from the pixel coordinates X, Y and the position of the detector and

$$\mathbf{x}' = D(\mathbf{u}, \phi' - \phi)\mathbf{x} \simeq \mathbf{x} + \mathbf{u} \times \mathbf{x} \sin(\phi' - \phi)$$

It can be shown that

$$\gamma \simeq \zeta \cdot \sin(\phi' - \phi) \qquad \zeta = \mathbf{c} \cdot (\mathbf{u} \times \mathbf{x})/|\mathbf{x}| = \mathbf{u} \cdot (\mathbf{S} \times \mathbf{S}_0)/|\mathbf{S} \times \mathbf{S}_0|$$

where ζ is closely related to the Lorentz factor.

3.2 REFLECTION PROFILE REPRESENTATION BY SLICES

In the original version of XDS designed for processing electronic area detector data the small size of the oscillation range in each data frame was neglected. The scattered intensity recorded in a pixel was treated as if the crystal was fixed at the angular position in the center of the oscillation range. The corresponding profile coordinates α, β, γ were computed and the intensity was distributed among the eight nearest grid points in the profile box. For larger oscillation ranges this simple method is invalid and has been replaced by a slice-representation of the reflection profiles as follows.

The profile axis of length $2\gamma_0$ along γ is divided into $2n_\gamma + 1$ equal intervals Γ_i of width Δ_γ (see fig.1).

$$\Gamma_i = \{\gamma \mid (i-1/2)\Delta_\gamma \leq \gamma \leq (i+1/2)\Delta_\gamma\} \qquad i = -n_\gamma, \ldots, n_\gamma$$
$$\Delta_\gamma = \gamma_0/(n_\gamma + 1/2) \qquad n_\gamma \geq 0$$

If $c(\alpha, \beta, \gamma)$ is the number of counts after background subtraction at α, β, γ, a reflection profile is then represented by $2n_\gamma + 1$ slices

$$C_i(\alpha, \beta) = \int_{\Gamma_i} c(\alpha, \beta, \gamma) d\gamma$$

Contributions to the reflection profile arrive from a number m of adjacent data frames each covering an interval Γ'_j of the same width starting at γ_1 along the γ-axis. This width Δ'_γ is a function of the oscillation range Δ_ϕ of the data frames and changes for each reflection.

$$\Gamma'_j = \{\gamma \mid \gamma_1 + (j-1)\Delta'_\gamma \leq \gamma \leq \gamma_1 + j\Delta'_\gamma\} \qquad \Delta'_\gamma = \zeta \sin \Delta_\phi \qquad j = 1, \ldots, m$$

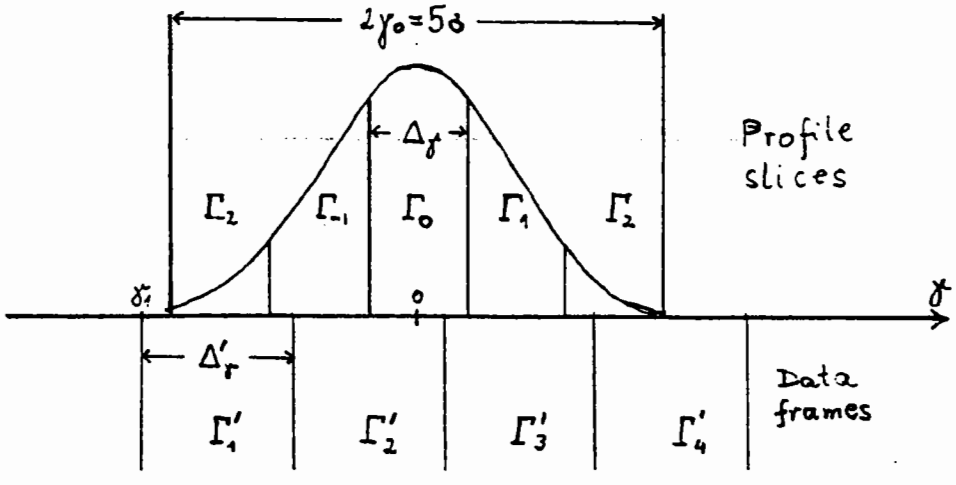


Figure 1: Profile representation by slices

The number of counts coming from data frame j is

$$C'_j(\alpha,\beta) = \int_{\Gamma'_j} c(\alpha,\beta,\gamma)d\gamma$$

and these counts must be distributed among the $2n_\gamma + 1$ slices of the reflection profile. The fraction f_{ij} of counts belonging to slice i which are contributed by data frame j is

$$f_{ij} \cdot C'_j(\alpha,\beta) = \int_{\Gamma'_j \cap \Gamma_i} c(\alpha,\beta,\gamma)d\gamma$$

Although the function $c(\alpha,\beta,\gamma)$ is unknown, all reflections roughly assume Gaussian profiles along γ and this approximation is used to estimate f_{ij}.

$$f_{ij} \approx \int_{\Gamma'_j \cap \Gamma_i} \exp(-\gamma^2/2\sigma^2)d\gamma \Big/ \int_{\Gamma'_j} \exp(-\gamma^2/2\sigma^2)d\gamma \qquad \sigma = \gamma_0/2.5$$

The integrals which are equivalent with the error function are evaluated numerically.

The essential parameters of the profile representation by slices are the oscillation range Δ_ϕ, the reflecting range of the crystal γ_0, and the number of subdivisions (slices) along γ. A proper choice of the parameter values is important for the quality of the integrated intensity data as can be seen from table 1.

Δ_ϕ	n_γ	Reflections	6Å	4Å	3Å	2.6Å
1/6°	4	unique,	1121	2293	3624	2356
		observed	8017	13436	21032	10148
		R-factor	6.6	7.7	12.7	31.6
1/2°	4		7.0	8.2	14.6	40.9
1°	4		7.3	9.8	19.5	62.6
1°	3		7.4	9.9	19.5	62.1
1°	2		7.5	10.1	19.8	61.5
1°	1		7.8	10.5	20.2	61.1
1°	0		8.5	12.2	24.1	70.1

Table 1: Effect of profile rastering n_γ and oscillation range Δ_ϕ on data quality as function of resolution. Data quality is defined by a symmetry R-factor, $100 \cdot \sum_{h,l} |I_{h,l} - I_h| / \sum_{h,l} I_{h,l}$ where h are unique reflection indices and $I_{h,l}$ are the intensities of symmetry equivalent reflections giving a mean value of I_h.

588 data frames were collected on a Siemens/Nicolet area detector from a hexagonal crystal ($P6_222, a = 87.7$Å,$c = 170$Å) at a distance of 13cm. The oscillation range was $\Delta_\phi = 1/6°$. Two additional data sets were derived by merging three or six consecutive original frames resulting in 196 ($\Delta_\phi = 1/2°$) and 98 ($\Delta_\phi = 1°$) new frames, respectively. All data sets were processed by the program XDS using $\gamma_0 = 0.5°$ for the original and $\gamma_0 = 0.6°$ for the derived data sets. The resulting R-factors are shown in table 1 as a function of resolution. As expected, the best results are obtained for the original data with the smallest oscillation range. Using small oscillation ranges for imaging plate detectors does not seem to be very attractive because of the long read out time. This disadvantage might be overcome by developing systems with several imaging plates which could be processed while data acquisition goes on, perhaps similar to the endless track design described by Sakabe [8] for data collection with a Weissenberg camera.

References

1. Kabsch, W. *J.Appl.Crystallogr.* **21** (1988a) 67–71.
2. Kabsch, W. *J.Appl.Crystallogr.* **21** (1988b) 916–924.
3. Kabsch, W. *J.Appl.Crystallogr.* (1993) submitted.
4. Buerger, M. J. *Z.Kristallogr.* **109**, (1957) 42–60.
5. Dijkstra, E. W. (1976). *A discipline of programming*, pp. 154–167.
New Jersey: Prentice-Hall, Inc.
6. *International Tables for Crystallography* (1989). Vol. A. Dordrecht:Kluwer Academic Publishers, pp. 738–749.
7. Andrews, L. C. & Bernstein, H J. *Acta Cryst. A44* (1988) 1009–1018.
8. Sakabe, N. *Nucl. Instr. and Meth. A303* (1991) 448–463.

LEAP, Laue Evaluation Analysis Package, for Time-Resolved Protein Crystallography

Soichi Wakatsuki
LMB, Rex Richards Building
South Parks Road, Oxford, OX1 3QU

A program package, LEAP (Laue Evaluation Analysis Package) is developed for data analysis of Laue diffraction patterns from macromolecules. The package is modular and thus easily extendible for inclusion of any new routines. Main capabilities are visualisation of data, various diagnostic routines for systematic errors, integration of Laue diffraction patterns, and post-integration data reduction. Extensive visualisation routines are available for the various stages of the data analysis. These include three-dimensional views of diffraction patterns for fast evaluation of the quality of diffraction patterns and precision of spot prediction; diagnosis of systematic errors; display of the completeness of data both before and after data reduction; inspection of systematic incomplete sampling which is unique to Laue diffraction, analysis of the integration results; and scaling and merging Laue data into unique data set. Routines related to integration are determination of granularity for appropriate evaluation of variances of observation, refinement of film parameters, preparation of profiles, integration of dense Laue patterns based on profile fitting to both spatial and non-spatial overlaps, analysis and trial integration of the profiles generated during the integration. The post-integration analysis consists of film-response, wavelength normalisation, transmission-absorption correction, and deconvolution of energy overlaps.

1. Introduction

One of the most important advantages of Laue diffraction for macromolecules is its ability for simultaneous recording of a large number of reflections. Very often ten to hundreds of thousand of spots on a film are observed. The precise measurement of the intensities of these reflections has presented a challenge. Each of the reflections arises from different wavelength, thus wavelength normalisation is required. A significant number of reflections may be spatially overlapping, especially for crystals of large unit cell. For instance, if the crystal-to-plate distance is chosen to cover the maximum number of reciprocal lattice points using a film, up to 30% (for a pea lectin crystal, $P2_12_12_1$, a=50.37Å, b=60.58Å, c=135.5Å) and 70% (for doubled cell dimensions) of the reflections could be spatially overlapped assuming a spatial resolution limit of 200 μm (Cruickshank *et al.*, 1991, Tables 3 and 4). For a broad band incident radiation a certain proportion of the reflections are generated as overlapping energy harmonics at λ, $\lambda/2$, $\lambda/3$ and so on, which occur at the same scattering angle, thus harmonics deconvolution is required. Cruickshank *et al.*, (1987) have shown that up to 17% of them could be overlaps of harmonics requiring deconvolution under certain conditions. These are some of the reasons for the difficulty of obtaining structure amplitudes with quality and completeness comparable to those of monochromatic data. There have been three major programs/program packages available for processing macromolecular Laue diffraction, Daresbury package (Helliwell *et al.* 1989, Shrive *et al.*, 1990) which now includes J. Campbell's new autoindexing and prediction program LAUEGEN, LAUE by Chess group (Smith Temple and Moffat 1987), and DESY package, LAUEMAD, based on MADNESS (Bartunik and Borchert, 1989). In these packages, each stage of the analysis is treated by a separate program with exception of Chess package, LAUE, which analyses film-response and wavelength normalisation in one program. These problems are highly convoluted in nature and best treated as a composite constraint. For instance, many of energy-overlaps cannot be deconvoluted with

data from only 6 films of a film-pack, but the integrated intensities on the series of films could still be used as constraints in structure refinement.

LEAP aims to provide a combined approach to the problems, (1) evaluation of data, (2) integration, and (3) post-integration analysis to obtain structure amplitudes. It does not include auto-indexing and prediction since reliable programs are already available in CCP4 package (NEWLAUE and GENLAUE written by I. J. Clifton, see also Helliwell *et al.*, 1987). There is now a new program, LAUEGEN, written by J. Campbell as a replacement of NEWLAUE and GENGLAUE which could be used as a complimentary package for LEAP.

2. Description of LEAP

The package is a collection of routines, mostly written in C language, which can be called from the top level menu tablet which has two sections: data bins and command lines. It can contain unlimited number of buffers each of which can contain any number of data bins, as many as the memory allows. Each buffer can be stored into a file on a disk and later restored from a disk file back into the program. Data bins can be X-ray diffraction images (Optronics-scanned film images, MAR-system and Raxis-II, Molecular Dynamics, and Fuji BAS100, so far), generate files (GE1 and GE2), LCF files or profile bins which contain calculated profiles with other information of integration (see below). The conversion of the data format to new MTZ is planned. Tablet software was first written by E. Merritt and has been extensively expanded by A. D. Cox, and adapted for LEAP. Each bin is divided into sections: header, data, and history. History of a bin is updated every time the bin is processed or any text can be added to it at any stage. There is a routine which allows editing/checking contents of header of a GE1 bin. Another routine is available for editing/listing reflection data, which can be used to analyse the integration results. All the information for each reflection included in generate file can be listed and/or edited. For the editing, part of the reflections can be selected depending on their attributes such as multiplicity, wavelength, nodal index, intensity, standard deviation, etc..

LEAP is run on VAX-VMS environment using X-window for interactive graphics operation, UNIRAS graphics software package and NAG routines.

2-1. Viewing images

Images can be viewed as 2D or 3D plots on screen or printed to a hard-copy device. For successful data reduction, all the parameters for integration and postintegration analysis must be set right. The 3D plots are especially useful for fast evaluation of diffraction images, detection of artefacts introduced by a scanner, and estimation of spot radius and size of background area for the integration. Predicted positions of reflections superimposed with the 3D image (Fig. 1) show immediately if there are any problems in the prediction of spots, for instance, film-parameters and unpredicted strong diffraction spots.

2-2. Diagnosis of systematic errors

There are five routines to evaluate the distribution and quality of generate (GE1/GE2) and LCF files. Two routines are available for viewing the coloured reflections of GE1 generate bins in 3D reciprocal space, one routine for many different kinds of hardware using UNIRAS (Fig. 2) another for on-line rotation, shift, zooming on X-window. Other two routines can show similar 3D distribution of reflections from LCF bins. The first is to plot slices of reflections on, for instance, (h, k) planes at each l level in which reflections are coloured according to their attributes (presence (Fig.3), intensity, standard deviation, wavelength, etc.). This immediately shows systematic absence of reflections, systematic errors, and incomplete sampling.

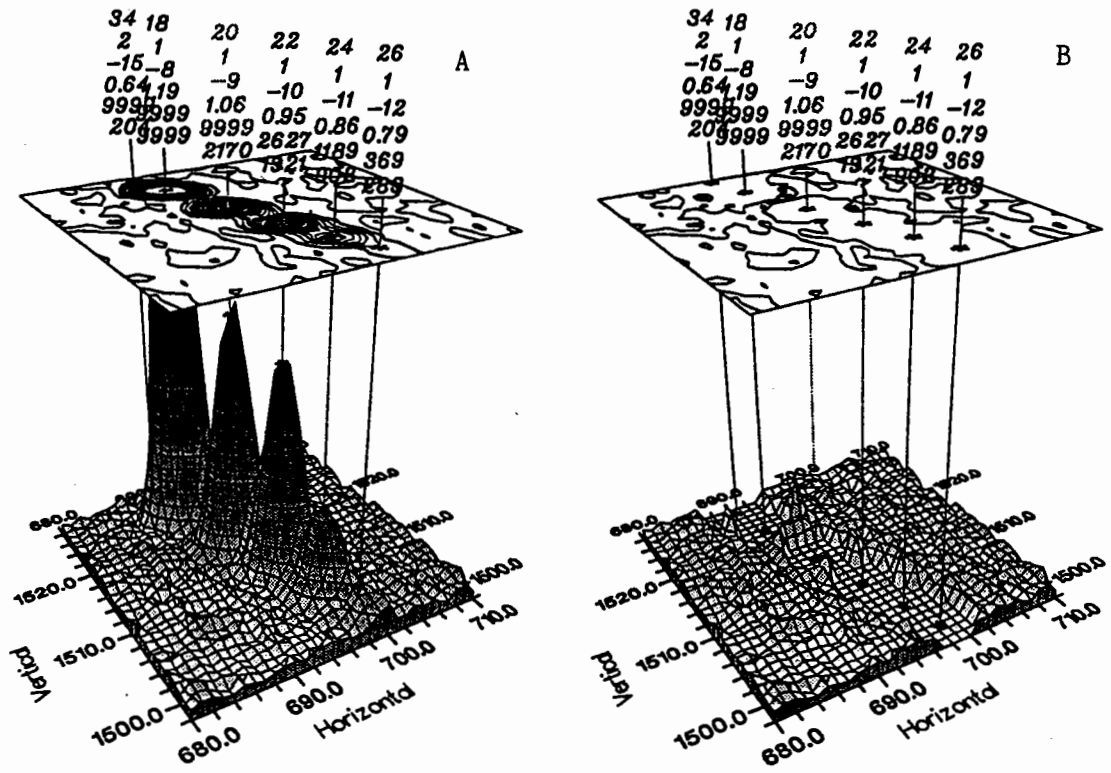


Fig. 1 (a) A 3D representation of a Laue photograph from cytochrome *c* peroxidase with a 2D contour map above. Predicted peak positions are indicated by *'s and lines with their Miller indices, (*h*, *k*, *l*), wavelength λ(Å) and box- and profile-intensities. X and Y coordinates are given in 50 μm raster units. (b) The same view but the pixels within the radius of 175 μm from the predicted positions are removed and set to the lowest level showing the pixels to be used for background determination.

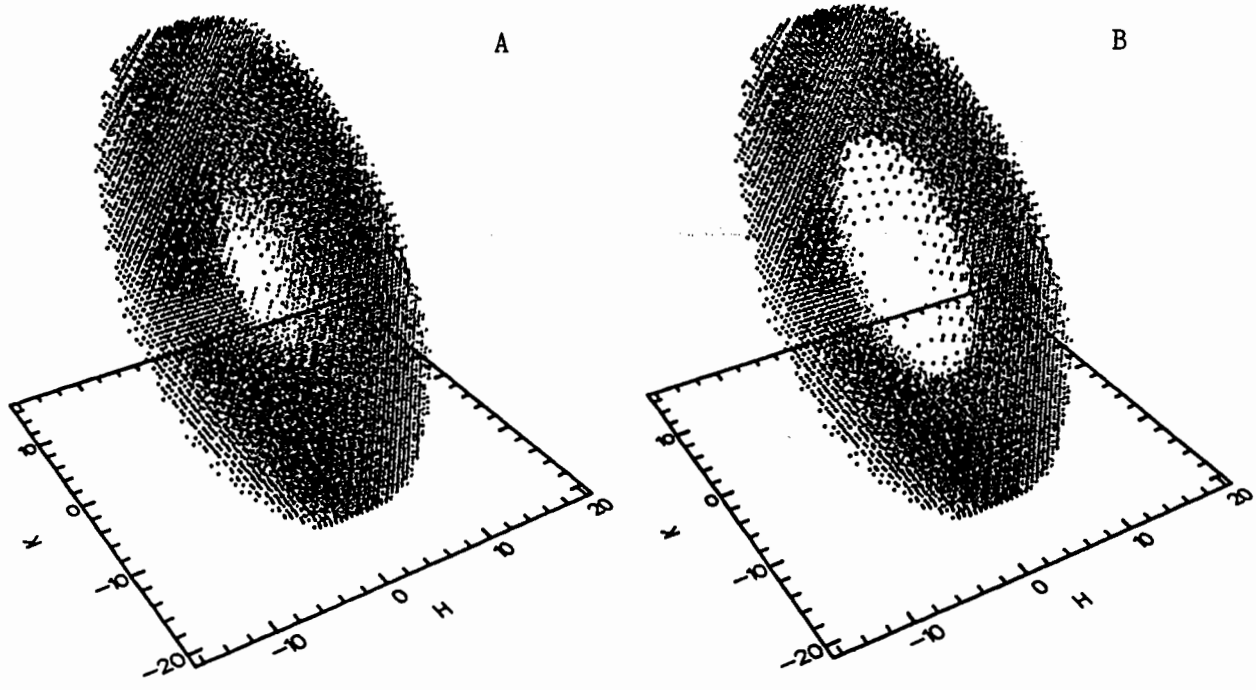


Fig. 2 Distribution of reflections in reciprocal space coloured according to wavelength (blue: short - red: long), when a colour device is used. A: all the reflections, B: singlets only.

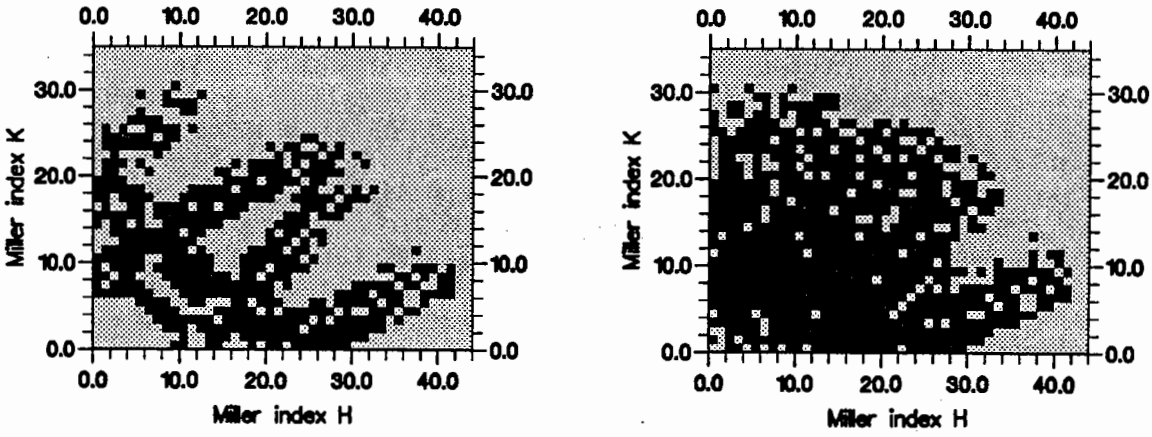


Fig. 3 Distributions of unique reflections in (h,k,11) plane, one film pack (left) and combination of 4 film packs (right).

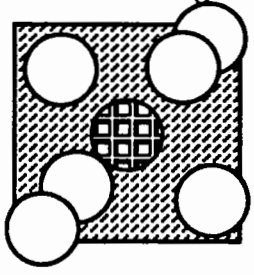
3. Integration

Granularity of the film or gain of the image plate scannter is needed for estimating variances of the measurements in the integration routine. They should be known for each kind of film or image plate scanner. In practice, however, it varies with the conditions of film development or the read-out apparatus. A routine is available to determine Selwyn granularity or the gain for each image using pixels in background areas around reflections.

Integration of Laue photographs can be performed interactively if the number of predicted reflections is fewer than 50000. For data sets larger than that and also for convenience, the package can generate an input parameter file for a separate stand-alone integration program which can be submitted as a batch job. This stand-alone version works in the exactly the same way as the interactive one except for the input and output. The integration program, either interactive or batch, copies the input GE1 generate file to an output GE1 file at the beginning for saving the integration results at the end, rather than modifying the contents of the input file. It also creates a secondary file containing profiles generated during the integration, listing of members of groups of spatially overlapped spots, which can be analysed separately by the following routines.

3-1. Background determination

In Laue photographs where many reflections are very close or even overlapped, use of a narrow background region often prevents reliable calculation of the background level. In order to circumvent this problem, this package defines a square region (dashed area in the figure below) around a reflection (central circle) excluding any pixels belonging to other reflections within the square (open circles). Typically the size of the background area is 2 to 3 times as large as the spot size.



This allows a substantially wider area around a reflection to be used for the determination of the background even if reflections are next to each other. The program then fits a plane to the pixels in the square but not in any of the spots (see Fig. 1 b). After this initial step, the program excludes pixels which are farther than user-specified distance, for instance, $3\sigma_{bg}$, from the background plane and recalculates the background. Goodness of the fit is tested using the incomplete Gamma function which gives a probability of finding a solution as bad as χ^2.

3-2. Refinement of film parameters

For successful integration of Laue photographs, it is essential to predict spots within at least one pixel. Root-mean-square (RMS) deviation between peak and predicted positions is often misleading; even if RMS deviation is, say, 0.8 pixel, many spots might be found to be off by 1 or 2 pixels if one browses through the film using the 3D view described above. A poor mismatch is the result of insufficient refinement in the prediction stage (GENLAUE) since it uses rough estimate of background. In order to overcome this problem, the package has a routine to refine film parameters using the very accurate background determination described below. Currently, it is an interactive process which requires a few cycles but will be implemented in the integration routine. After the refinement using this routine, the spot predictions are found much closer to the actual peak positions.

3-3. Profile fitting and deconvolution of spatial overlaps

Here we use n good spots to form a profile which best fits all the n spots. Optical densities are first corrected for background using the method described above. The pixels in each spot are numbered 1 through m in each spot. Let $\rho_{i,k}$ be a background-subtracted optical density measurement of the i-th spot at k-th pixel $(x_{i,k}, y_{i,k})$:

$$\rho_{i,k} = \rho_{i,k(measured)} - (ax_{i,k} + by_{i,k} + c)$$

We are now ready to integrate spots using the profiles generated.

$$\rho_k = \sum_{i=1}^{n} J_i \tilde{p}_{i,k} \quad (k = 1,2,,,m)$$

where ρ_k is background-corrected optical density at the k-th pixel J_i intensity of the i-th spot, and $\tilde{p}_{i,k}$ value of the profile for the i-th spot at the k-th pixel. The design matrix for finding intensities which best fit the background-corrected optical density is prepared by partial-differentiation of the sum of the squared differences between the two sides of the above equation. Weighted normal equation is

$$\begin{pmatrix} \sum_{k=1}^{m} \tilde{p}_{1,k}\tilde{p}_{1,k}/\sigma_k^2 & \sum_{k=1}^{m} \tilde{p}_{1,k}\tilde{p}_{2,k}/\sigma_k^2 & \cdots & \sum_{k=1}^{m} \tilde{p}_{1,k}\tilde{p}_{n,k}/\sigma_k^2 \\ \sum_{k=1}^{m} \tilde{p}_{2,k}\tilde{p}_{1,k}/\sigma_k^2 & \sum_{k=1}^{m} \tilde{p}_{2,k}\tilde{p}_{2,k}/\sigma_k^2 & \cdots & \sum_{k=1}^{m} \tilde{p}_{2,k}\tilde{p}_{n,k}/\sigma_k^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{m} \tilde{p}_{n,k}\tilde{p}_{1,k}/\sigma_k^2 & \sum_{k=1}^{m} \tilde{p}_{n,k}\tilde{p}_{2,k}/\sigma_k^2 & \cdots & \sum_{k=1}^{m} \tilde{p}_{n,k}\tilde{p}_{n,k}/\sigma_k^2 \end{pmatrix} \begin{pmatrix} J_1 \\ J_2 \\ \vdots \\ J_n \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{m} \rho_k \tilde{p}_{1,k}/\sigma_k^2 \\ \sum_{k=1}^{m} \rho_k \tilde{p}_{2,k}/\sigma_k^2 \\ \vdots \\ \sum_{k=1}^{m} \rho_k \tilde{p}_{n,k}/\sigma_k^2 \end{pmatrix}$$

3-4. Analysis of profiles and test integration

Number of spots used in generating profiles of bins arranged in concentric grid can be displayed in order to see if there are enough spots evenly distributed on the films. Profiles generated during the integration can be viewed in 2D/3D for diagnosis of the spot radius and accuracy of spot prediction. For testing effects of various parameters used for integration, there is a routine to perform test integration. For spatially overlapped spots, the same routine can also list members of each group with the integration results.

3-5. Resolving power of deconvolution of spatial overlaps

The above method can deconvolute spatial overlaps very reliably with better than a few percent accuracy as long as they are separated by half the spot size. If they are closer than that, accuracy of the weaker of the two is more seriously affected. Spatially overlapped spots whose distances are shorter than the half of the spot size are excluded from the deconvolution of the group of the spatial overlaps. Pixels which belong to the too-heavily overlapped spots are removed from the calculation of the evaluation matrix. It is therefore important to predict all the spots regardless of their degree of overlap.

3-6. Results of integration

So far, the integration procedure has been applied to a few cases from relatively less crowded (cytochrome *c* peroxidase, CCP), intermediate (T-state phosphorylase *b*), and very dense (Rubisco and viruses). In all cases, significant reductions in R-merge ($=\Sigma_j\Sigma_h|I_{hj} - \langle I_h \rangle|/\Sigma_j\Sigma_h\langle I_h \rangle$) were observed using the package.

Laue photographs from CCP crystals were relatively well-separated and there was little streaking of spots, thus an easy problem to tackle. Yet, the integration using LEAP has resulted in a clean difference map between the semistable, doubly oxidized intermediate, Compound I, and the parent protein. In addition to the shift of the haem iron, which was previously observed, a pair of positive and negative peaks shows a movement of Arg48 toward the haem iron at the active site (Fülöp *et al.*, in preparation). The shift of ARG48 was not apparent in difference maps based on intensities obtained with INTLAUE.

In case of T-state phosphorylase, the difference maps calculated using LEAP integration has improved the noise level significantly and now show phosphate at the attacking position in the complex of the enzyme and heptenitol (Duke *et al.*, in preparation). The refinement of the film parameters was the major step forward in improving the quality of the difference map.

Another problem in the analysis of Laue diffraction data is crystals which diffract to high resolution in which case many of spots are closely arranged along lunes. Rubisco is a good example of this type of problem where almost half of the spots are spatially overlapped due to the high resolution. While the use of a much larger detector will be clearly helpful, the problem could be partially solved if one uses the differential absorption of the films/image plates since high and low energy spots often alternate along a lune (see below).

4. Post-integration analysis

Once integration is complete, the GE1 generate files are further processed using a series of methods for post-integration analysis of Laue diffraction, the full detail will be described elsewhere.. It consists of internal determination of film-response (Victoreen curves), wavelength normalisation, transmission-absorption correction due to crystal and non-diffracting materials such as buffer and capillary, and deconvolution of energy overlaps. Propagation of variances are carefully monitored, which are used for χ^2-tests in the linear least-squares minimisation throughout the process using the

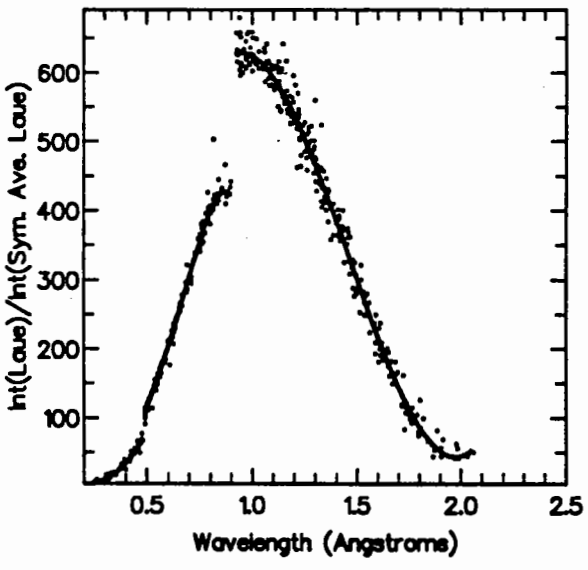


Fig. 4 An example of wavelength normalisation curves.

inverse of the normal matrix for the covariance matrix rather than using the goodness of the fit for estimating the variances.

An iterative non-linear procedure for wavelength normalisation is proposed based on the comparison of all the measurements from symmetry equivalent spots. A typical wavelength normalisation curve is shown in Fig. 4. Using the discrepancy between the symmetry related singlet reflections after applying the wavelength normalisation, an empirical, wavelength dependent algorithm is developed for positional correction of transmission and absorption due to the crystal and other non-diffracting material. As the final stage of the post-integration analysis, energy harmonics are deconvoluted using all the measurements from symmetry equivalent multiplets as a combined least-squares problem with appropriate χ^2-tests which eliminates outlying measurements. The quality of deconvoluted multiplets is similar or only slightly inferior to that of the singlet reflections. The method, combined with the previously described integration procedure, can produce difference maps of significantly higher quality.

5. Low resolution hole and completeness of data

Completeness of data sets as a function of resolution, either generate (GE1/GE2) or LCF files, can be calculated. If necessary, various screening conditions, such as spatial overlaps, nodal spots, and Miller indices, can be applied before calculating coverage. Although there is a small hole in the distribution of reflections in reciprocal space from one film pack of CCP data (Fig. 3a), the predicted completeness (Fig. 5a) does not show a significant loss of data between 12Å and 4.4Å (d_{min} = 2.2Å). This is due to the symmetry operation (space group of the CCP crystal is $P2_12_12_1$). When four film packs are combined, the distribution of completeness is more uniform (Fig. 5b). Thus, the geometry of the Laue methods is NOT the source of the poor completeness of the data nor the low-resolution hole. Yet, the final results are often not so encouraging (Fig. 5c and d). Why is the completeness of final data set in the Laue methods so low? First, we discuss the low resolution hole and then the implication of experimental set-up on the poor completeness in general.

When multiplets are not deconvoluted, there are two consequences. First, reflections whose resolution is below $2xd_{min}$ are not recovered. Except for a minor contribution from the low resolution singlets which occur at short wavelength, majority of the low resolution data are lost in this way, thus called low-resolution hole (Fig. 3b). The loss of low resolution data causes fragmentation of difference maps (Hajdu *et al.*, 1991, Duke *et al.*, 1992). Second, the higher harmonic components of multiplets, which have Miller indices that are multiple of 2, 3 and so on, are not sampled. This systematic incomplete sampling gives rise to artificial peaks at the half length of the unit cell in the point spread function (Duke *et al.*, 1992), and ghost images in difference maps (Wakatsuki, in preparation).

Even though the above method for deconvoluting energy harmonics using symmetry equivalents has improved the statistics of the structure amplitudes, the completeness of the low resolution data is still far too low. This is largely due to the poor energy resolution of X-ray films which are hardly sensitive to higher energy X-ray photons. The typical set-up of 6 X-ray films in a film-pack is not sufficient to give differential absorption wide enough for deconvoluting energy-harmonic components whose absorption coefficients are different only by a factor of 2 to 5. Image plate is not only sensitive at higher energy but also its absorption is nearly 100% up to about 20 keV but 20-30% at 50keV. Thus, by stacking image plates in one pack (Wakatsuki, Soltis, Phizackereley, in preparation) low energy component of a doublet is completely absorbed on the top image plate while the higher energy component can be resurrected from the subsequent image plates.

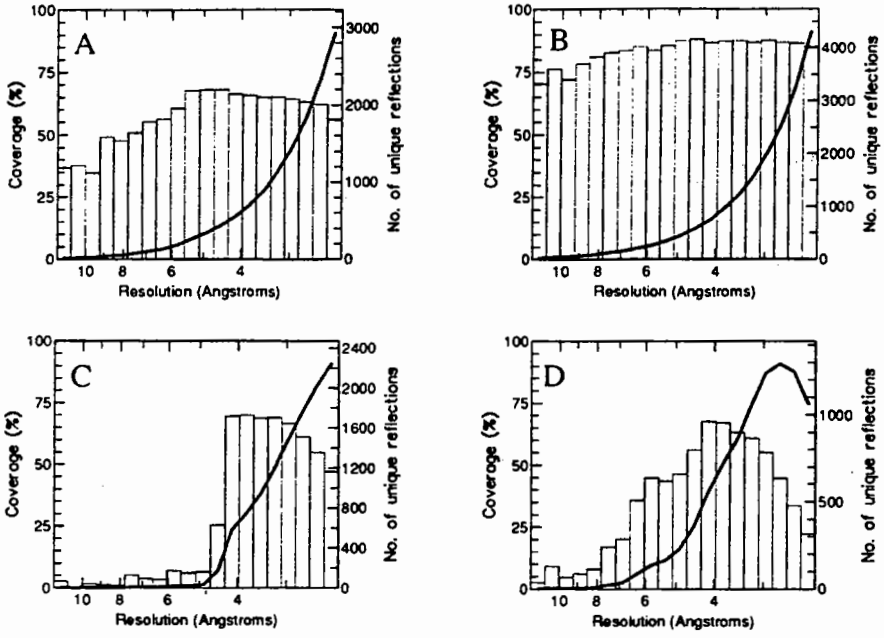


Fig. 5 Completeness of the data. Histograms show percent completeness while the lines show number of reflections in each resolution bin. (a) prediction from one film pack, (b) combination of predicted spots from 4 film packs, (b) reflections used in the difference method (Hajdu, et al., 1987), and (d) those used in LEAP.

The choice and set-up of a detector is one of the most important experimental factors in collecting high completeness Laue data set, apart from the need for high quality crystals. The use of 5-inch X-ray films does not compare favourably with much larger image plate detector systems, with active size ranging from 9cm radius circle (Mar System) to 40cm by 80cm rectangular Weissenberg. Given the general rule of the spot-to-spot resolution of half the spot size and the scanning raster size of 100 or 150μm of the image plate scanners, an image plate detector to be used for Laue diffraction from protein crystals must be as large as possible.

The other problem of experimental set-up is the optimal exposure time. As long as the crystal can survive the radiation damage, the exposure time has to be long enough to catch weak, high resolution data. Use of image plate with wider dynamic range helps in this respect, but the recent experiment at SSRL (Wakatsuki, Soltis, Phizackereley, in preparation) showed that the 10^4 dynamic range is not enough for efficient data collection for both strong and very weak spots. In order to expand the dynamic range even wider without sacrificing the sensitivity, we propose to stack image plates to recover very strong spots from the lower image plates while collecting weak spots on the top image plate by long exposure time. This has also an implication of deconvolution of energy overlaps as discussed above.

Of course, there is the problem of the streaking of spots due to disorder, an inherent problem of the Laue method. But, we have to wait for accumulation of image plate Laue data before making a fair comparison between the Laue method and the fast monochromatic, oscillation or Weissenberg, data collection using large image plate.

6. Extension

The viewing routine using X-window has been extended for analysis of monochromatic oscillation data. So far an interactive indexing routine has been implemented, which has been used successfully to determine cell parameters and index reflections from *E. coli.* phosphorylase. Deconvolution of spatial overlaps is being incorporated for crystals with large unit cells and large-angle oscillation Weissenberg data (Hajdu, 1993). UNIX version is also under consideration.

Distribution of the package

The package is available from S. Wakatsuki, Laboratory of Molecular Biophysics, Rex Richards Building, South Parks Road, Oxford OX1 3QU, UK, +44-865-275379 (Office), +44-865-510454 (Fax), Wakatsuki@vax.molecular-biophysics.oxford.ac.uk.

Acknowledgements

LEAP has been developed in the groups of Prof. L. N. Johnson and Dr. J. Hajdu with the invaluable help from the members of the groups, I. Andersson, I. J. Clifton, E. M. H. Duke, V. Fülöp, A. Hadfield, M. K. Jaynes, P. Nordlund, S. Walters and P. Williams. Menu tablet used in the package was developed by E. Merritt and A. D. Cox. The first template of the analysis package was provided by A. D. Cox.

References

Bartunik, H. D. and Borchert, T. (1989) *Acta Cryst.* **A45**, 718-726.

Cruickshank, D. W. J., Helliwell, J. R. and Moffat, K. (1987). *Acta Cryst.* **A43**, 656-674.

Cruickshank, D. W. J., Helliwell, J. R. and Moffat, K. (1991). *Acta Cryst.* **A47**, 352-373.

Duke, E. M. H., Hadfield, A., Walters, S., Wakatsuki, S., Bryan, R. K., and Johnson, L. N. (1992) *Phil. Trans R. Soc. Lond.* A (1992) **340**, 245-261.

Hajdu, J. et al. (1987) *Nature* 329, 178-181.

Hajdu, J., Almo, S. C., Farber, G. K., Prater, J. K., Petsko, G. A., Wakatsuki, S., Clifton, I. J., and Fülöp, V. (1991). in *Crystallographic Computing* 5: *From Chemistry to Biology*, Oxford University Press, 27-49.

Hajdu, J. (1993) Fast Weissenberg data collection as an alternative to the Laue method in kinetic crystallography, in *Synchrotron Radiation in Biosciences*, ed. N. Sakabe, Oxford University Press.

Helliwell, J. R., Habash, J., Cruickshank, Harding, M. M., Greenhough, T. J., Campbell, J. W., Clifton, I. J., Elder, M., Machin, P. A., Papiz, M., Zurek, S. (1989) *J. Appl. Cryst.* 22, 483-497.

Shrive. A. K. et al, (1990), *J. Appl. Cryst.* 23, 169-174.

Smith Temple, B. R. and Moffat, K. (1987). In computational Aspects of Protein Crystal Data Analysis, Rep. DL/SCI/R25, ed. J. R. Helliwell, P.A. Machin, M. Z. Papiz, p.84. Daresbury, UK; SERC Daresbury Lab.

Szebenyi, D. M. E., Bilderback, D. H., LeGrand, A., Moffat, K., and Schildkamp, W. (1992) *J. Appl. Cryst.* 25, 414-423.

"THE CHOICE OF X-RAY WAVELENGTH

IN

MACROMOLECULAR CRYSTALLOGRAPHY"

J.R. Helliwell

Department of Chemistry, University of Manchester, M13 9PL and
SERC, Daresbury Laboratory, Warrington, WA4 4AD

ABSTRACT

The freedom to select the wavelength in data collection is one of the most important benefits of synchrotron X-radiation. At Daresbury the use of the wiggler radiation at a monochromatic wavelength of ≈0.9Å enhanced the absorption efficiency of film as the detector due to the AgBr grains and the Br K edge, well matched to the wiggler critical wavelength (0.93Å). Film has now been largely superceded at Daresbury. The use of 0.9Å is still very beneficial because sample absorption is reduced and lifetime, in general, increased (e.g. for virus crystals which have so far not responded to freezing strategies for radiation protection). Careful tuning of the wavelengths in this range also is useful for SIROAS of heavy metal derivatives (e.g. Pt, Au, Hg) and, of course, other anomalous scattering techniques. Higher energy synchrotron sources such as CHESS, ESRF, APS, SPRING-8 and the proposed DIAMOND machine for Daresbury will offer copious fluxes at even shorter X-ray wavelengths e.g. 0.5 or even 0.3Å, where detector absorption efficiency (e.g. for IP or CsI coupled CCD) can still be high. First tests have been made and reported from use of CHESS. Experience on a range of samples at λ's $\leqslant$ 0.5Å still has to be gained to see the general effect on sample lifetime and the potential benefit of greatly reduced absorption, as a systematic error, particularly for high resolution data. Given the progress made in the field with protein crystal freezing techniques, a proposal is made here for the use of a longer wavelength (e.g. 2.5Å) for increasing the scattering efficiency of very small crystals (e.g. (20 μm)^3) of proteins whereby the lifetime is now controlled by the freezing; sample absorption errors will in any case be small, even at the long wavelength, due to the small thickness of the crystal.

1. Introduction

Synchrotron radiation is intense, tunable and well collimated and is extensively used now in macromolecular crystallography to overcome a variety of technical hurdles encountered with data collection in the home laboratory. The freedom to select wavelength is one of the most important benefits of synchrotron X-radiation (SR). In conjunction with the other beneficial properties of SR, referred to above, new strategies for data collection have opened up. This paper describes in essence why we use the wavelengths that we do at current SR sources and outlines future prospects. An appendix outlines a new idea to use a longer wavelength for protein microcrystal data collection.

2. Terminology

It is useful to define some terms. Short wavelengths generally refer to use of monochromatic beams at ≈0.9Å. Very-short wavelengths are at ≈0.5Å and ultra-short wavelengths are referred to for ≈0.3Å beams. These are terms as used by Helliwell (1992). In this paper a new proposal for longer wavelengths for data collection from very small protein crystals is made, which is meant to cover ≈2.5Å wavelength.

3. Short Wavelengths

The use of a monochromatic beam of 0.9Å has various benefits. Sample absorption is reduced significantly and variations in the correction factor which needs to be applied are also reduced. Hence, systematic errors in the intensity data are significantly reduced. This improves the quality of protein model refinement and of the isomorphous and anomalous differences between heavy atom derivative and native data.

The lifetime of a crystal as a function of wavelength is predicted to be improved at shorter wavelengths i.e. that more data can be measured per crystal sample. This theoretical prediction is based on consideration of the **number** of protein molecules destroyed as a result of the **number** of photons absorbed. Improved crystal lifetime has indeed been noted by the virus crystallographers (Acharya et al. (1989), Liddington et al. (1991)). Virus crystal samples have so far resisted attempts at freezing to prolong their lifetime.

The actual choice of 0.9Å is a technical one due to the machine spectral output and the detector available. At Daresbury the SRS wiggler has a λ_c of 0.93Å for the machine running at 2 GeV and the wiggler at 5T. Photographic film has an increase in its absorption efficiency at 0.92Å due to the AgBr grains in the emulsion. Hence, a choice of 0.9Å at Daresbury is a good combination of spectral output of the machine and of the optimisation of film as a detector. The use of 0.9Å instead of CuKα also allows a longer crystal to detector distance and so gives

improved signal to noise via a reduced background under a spot. Another aspect of interest in the short wavelength range is the optimisation of the anomalous scattering of the heavy atoms (e.g. Pt, Au, Hg) in the method of single or multiple isomorphous replacement. At CuKα f" is ≈7e⁻ whereas on the short wavelength side of the L_I absorption edge f" jumps to ≈12e⁻ (e.g. L_I edge of Hg is at 0.83Å). Simultaneously the protein absorption is considerably reduced. The shortest wavelength used for data collection on the SRS wiggler protein crystallography station 9.6 was 0.6Å for a data set recorded from nitrogenase crystals by J. Bolin in 1984. This was to optimise the molybdenum f" at the K absorption edge of 0.62Å.

4. Very–short and ultra–short wavelengths

High machine energies exist at CHESS and at ESRF producing a copious output of very–short and ultra–short wavelengths. The arguments presented in the previous section can be taken to a logical conclusion, for sample lifetime and reduced absorption, at a wavelength of OÅ! According to the theory the benefit would be infinite sample lifetime and no absorption correction at all. The detector would also be infinitely far from the crystal and so the background under a spot would be zero. To exploit this of course needs infinite flux (because the scattering efficiency decreases at λ^2), zero divergence and a perfect crystal (to keep the spot size finite).

In practice how short a wavelength can one go whilst preserving a reasonable exposure time and crystal to detector distances as well as detector absorption efficiency?

A high brilliance, high energy machine like ESRF (or APS or SPRING–8) should allow use of 0.3Å wavelength, a value set by the sudden increase in detector absorption at Cs or I or Ba (depending on CsI coupled CCD or barium in an image plate; the K edges of Cs, I and Ba are respectively 0.3445Å, 0.3738Å and 0.3310Å).

Perhaps the disadvantage of this wavelength range is in terms of the anomalous scattering of the heavy atom derivative L edges (referred to in Section 3). However, the K edges of these elements become accessible at ≈0.15Å! In summary, at a wavelength of 0.5 or 0.3Å sample lifetime could hopefully be long enough to consistently allow a complete data set from one crystal. The background on the detector would be especially low. These are conditions which would allow the highest possible resolution of data to be measured from a crystal, useful for direct methods and model refinement (Helliwell et al. (1993). Novel data collection schemes also suggest themselves. With a large enough detector (e.g. 1m × 1m image plates) a large–angle oscillation technique (LOT) has been suggested whereby all the data from a protein crystal is put onto a single image (Weisgerber and Helliwell (1993)).

5. Longer wavelengths and data collection from frozen protein "microcrystals"

As sample volume decreases the scattering efficiency also decreases linearly with volume. A sample of size $20 \times 20 \times 20\ \mu m^3$ will scatter $10^{-3} \times$ that of a crystal of size $200 \times 200 \times 200\ \mu m^3$. However, sample absorption is reduced and use of a longer wavelength would be allowed to increase the scattering efficiency (by $\approx \lambda^2$).

Working with photographic film as detector and use of station 7.2 at Daresbury, Mahendrasingam et al. (1986) collected data from a lysozyme crystal at a wavelength of 2.6Å. The mylar capillary was needed instead of a glass capillary to avoid the capillary absorption (a factor of 20) at 2.6Å wavelength. A new version of this mode of data collection would be to freeze the crystal on the end of a glass fibre or in a wire loop (as reported elsewhere in these proceedings) and so avoid the glass capillary directly.

Freezing a crystal could sufficiently preserve its lifetime in a long wavelength beam. The use of 2.5Å instead of 0.9Å wavelength would increase the scattering efficiency by 8.8 × (see Appendix). The linear absorption coefficient of protein at this wavelength would be $\approx 5 mm^{-1}$. Hence, the transmitted beam intensity for a 20μm thick crystal would be $\exp^{-}(0.02 \times 5) = e^{-0.1} = 0.91$ (i.e. 9% absorbed). At a wavelength of 2.5Å, a 2.5Å d_{min} would require a Bragg angle of 30° (i.e. 60° 2θ). Tests with an image plate, for example, at such a wavelength needs to be made. These ideas parallel arguments made in favour of use of CuKα over MoKα in chemical crystallography of small crystals of smaller molecules, which are radiation stable, with conventional X-ray sources (M. Helliwell et al. (1993)).

Concluding remarks

As the field of macromolecular crystallography at the synchrotron has developed a great variety of options for different data collections has opened up. The freedom of choice of wavelength is an essential parameter in experimental design. Just how short a wavelength one can use is partly determined by the synchrotron machine energy and the magnet field strength, which determines the output flux at a given wavelength, and by the perfection of the sample, the collimation of the beam and, in the case of anomalous scattering or detector absorption efficiency, where the absorption edges are!

Acknowledgements

The SERC, Daresbury Laboratory is thanked for provision of synchrotron radiation facilities allowing the use of the wavelengths mentioned of 0.6, 0.9 and 2.6Å. For the wavelengths of 0.5Å and 0.3Å the use of CHESS at Cornell University, U.S.A. is gratefully acknowledged.

REFERENCES

Acharya, K.R., Fry, E., Stuart, D., Fox, G., Rowlands, D. and Brown, F. Nature, *337* (1989) 709–16.

Helliwell, J.R. "Macromolecular Crystallography with Synchrotron Radiation" (1992) Published by Cambridge University Press.

Helliwell, J.R., Ealick, S., Doing, P., Irving, T. and Szebenyi, D. Acta Cryst. D, *49* (1993) 120–128.

Helliwell, M., Gallois, B., Kariuki, B., Kaucic, V. and Helliwell, J.R. Acta Cryst., *B49* (1993) In press.

Liddington, R., C. Yan Y., Moulai, J., Sahli, R., Benjamin, T.I. and Harrison, S.C., Nature *354* (1991) 278–284.

Mahendrasingam, A., Sowerby, A., Helliwell, J.R. and Thompson, A.W. (1986) Daresbury Laboratory Annual Report Appendix (1985/86) on Synchrotron Radiation page 183.

Weisgerber, S. and Helliwell, J.R. Faraday Transactions (1993) submitted.

APPENDIX

Monochromatic data collection and wavelength dependent factors

The total energy in the diffracted beam, for a reflection hkℓ, recorded by the detector is given by

$$E(hk\ell) \sim \frac{I_o}{\omega} \lambda^3 \, {}^{\lambda}L \, {}^{\lambda}P \, {}^{\lambda}A \, \frac{V_x}{V_o^2} \, |{}^{\lambda}\underline{F}(\underline{h})|^2 \, {}^{\lambda}D \qquad (1)$$

λ is the wavelength ($= 2d_{hk\ell} \sin\theta$)

I_o is the intensity of the incident beam and depends on λ (i.e. SR spectral curve)

ω is the angular velocity of the crystal

${}^{\lambda}L$ is the Lorentz factor and depends on λ

${}^{\lambda}P$ is the polarisation factor and depends on λ

${}^{\lambda}A$ is the absorption of the sample and depends on λ

V_x is the crystal volume illuminated by the X-ray beam

V_o is the unit cell volume

$|{}^{\lambda}\underline{F}(\underline{h})|$ is the structure factor amplitude for reflection $\underline{h}(=hk\ell)$ and depends on λ in the case of anomalous scatterers being present in the structure.

${}^{\lambda}D$ is the detector absorption efficiency and depends on λ.

The Lorentz factor, L, takes account of the relative time spent by each reflection in the reflecting position. It depends on the precise diffraction geometry used. In the rotation method

$$L = \frac{1}{(\sin^2 2\theta - \zeta^2)^{1/2}} \qquad (2)$$

where ζ is the coordinate of a point P is reciprocal space parallel to the rotation axis, as the axis of a cylindrical coordinate system, relative to the origin of reciprocal space and θ is the Bragg angle of diffraction.

For $\zeta = 0$ i.e. for a reciprocal lattice point (RLP) lying in the equatorial plane (which is perpendicular to the rotation axis and for which $\Upsilon = 2\theta$, where Υ is the angle between the projections of the incident and diffracted beam vectors $\hat{\underline{S}}_o$ and $\hat{\underline{S}}$ respectively onto the equatorial plane; see figure 1) equation 2 becomes

$$L = \frac{1}{\sin 2\theta} \tag{3}$$

The polarisation factor, P, for the case of synchrotron radiation reflected from a singly bent triangular monochromator is (after Kahn et al. (1982)).

$$P = P_o - P' \tag{4}$$

where $$P_o = \frac{(1 + \cos^2 2\theta)}{2} \tag{4a}$$

and $$P' = \frac{\tau'}{2} \cos 2\psi \sin^2 2\theta \tag{4b}$$

where $$\tau' = \frac{\alpha(1+\tau) - (1-\tau)}{\alpha(1+\tau) + (1-\tau)} \tag{4c}$$

and $$\alpha = \cos 2\theta_M \tag{4d}$$

for a perfect crystal monochromator with its Bragg angle = θ_M

and $$\tau = \frac{I_{//} - I_{\underline{1}}}{I_{//} + I_{\underline{1}}} \tag{4e}$$

where $I_{//}$ = the flux delivered by the source with a parallel component of polarisation and $I_{\underline{1}}$ is the perpendicular component. ψ is the azimuthal angle in the detector plane of the diffraction spot (figure 1).

In the limit of very-short wavelengths both θ and θ_M are small, $P_o \rightarrow 1$ and $P' \rightarrow O$. Also, since $L = {}^1/_{\sin 2\theta}$, for the equatorial case, then $L \rightarrow {}^1/_{2\sin\theta} = d_{hk\ell}/\lambda$. Hence, equation 1 becomes, in such a limit,

$$E'(hk\ell) \sim \frac{I_o}{\omega} \lambda^3 . \frac{d_{hk\ell}}{\lambda} . 1 . 1 . \frac{V_x}{V_o^2} \; |\lambda \underline{F}(\underline{h})|^2 \, \lambda D \tag{5}$$

where the sample absorption λA is determined by the transmitted intensity $I_T = I_o \exp(-\mu t)$ and μ is the linear absorption coefficient of the sample and varies as λ^3, t is the sample thickness so that the sample becomes transparent at very-short wavelengths and no absorption correction is required (i.e. $\lambda A = 1$). λD, the

detector absorption also depends on λ but by judicious choice can essentially be designed to absorb all the intensity in the diffracted beam even for very–short wavelengths.

Hence,
$$E'(hk\ell) \sim \lambda^2 \qquad (5a)$$
$$\lambda \to O$$

i.e. in the short wavelength limit.

In the case of longer wavelengths (e.g. >1.5Å) explicit account must be taken of the θ dependence in the $^{\lambda}L$ and $^{\lambda}P$, particularly because cos 2θ can no longer be approximated to unity. In particular, as regards $^{\lambda}P$, it is not ideal to use a horizontally dispersing monochromator with a horizontally polarised SR beam because the polarisation attenuatoin of this component is significant, varying as cos $2\theta_M$; e.g. for Ge(111) at λ's = 1.5Å and 2.5Å, cos $2\theta_M$ is respectively 0.895 and 0.71. A vertically dispersing monochromator (double crystal type) will attenuate only the vertical component of the polarisation (twice, once at each reflection) but this is a very small intensity component anyway. Hence, to within a few percent τ' (eqn, 4c) can be set to 1 in such a case. Hence, in terms of the λ and θ dependent factors we get for E(hkℓ),

$$E''(hk\ell) \sim \lambda^3 . \frac{1}{(\sin^2 2\theta - \zeta^2)^{1/2}} \left\{ \frac{(1 + \cos^2 2\theta)}{2} - \frac{\cos 2\psi \sin^2 2\theta}{2} \right\} \qquad (6)$$
$$\lambda > 1.5\text{Å}$$

and again it is assumed that the detector absorption is optimised to obtain essentially 100% absorption of the diffraction pattern. In order to explore the λ dependence of E(hkℓ), in the limit of longer wavelengths, it becomes necessary to explicitly consider the θ dependence. Consider the situation at the highest Bragg angle e.g. for λ = 2.5Å, d_{min} = 2.5Å, θ = 30°. In the equatorial plane, ζ = 0 and ψ = 90°.

Hence, $E''(hk\ell) \sim \lambda^3 \frac{1}{2\sin\theta \cos\theta} . \frac{1}{2} \{(1 + \cos^2 2\theta) + \sin^2 2\theta\}$ (7)

$$\sim \lambda^3 \frac{1}{2\sin\theta \cos\theta}$$

At θ = 30°

$$E''(hk\ell) \sim \lambda^3 \frac{1}{2 \times 0.5 \times 0.866}$$

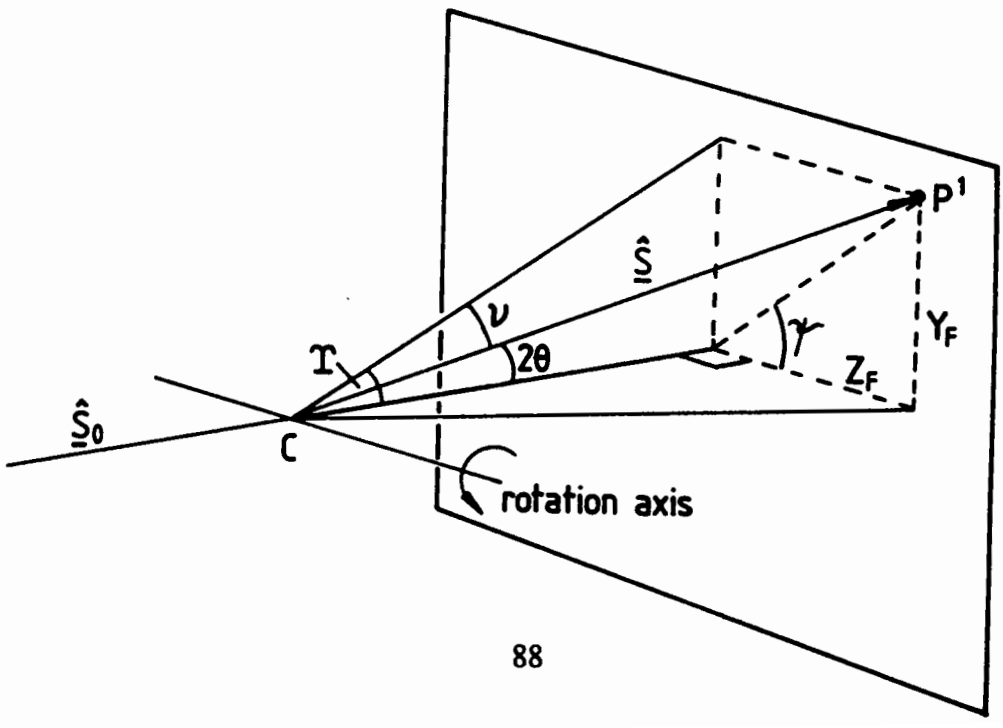
$$\sim \lambda^3 \frac{1}{0.866} \qquad = 18.04 \text{ (arbitrary units)}$$

Table 1 Variation with wavelength, λ, of the energy in a diffracted beam, recorded in the equatorial plane.

λ	0.9Å	1.5Å	2.0Å	2.5Å
E"(hkℓ) (arb. units)	2.05	5.9	10.9	18.04
E"(hkℓ) Relative to 0.9Å wavelength	1.0	2.88	5.32	8.8
Ratio of λ^2 (relative to 0.9Å λ)	1.0	2.78	4.93	7.7

Table 1 compares values of E"(hkℓ) calculated as a function of various wavelengths for a diffracted beam in the equatorial plane. Essentially the trend varies as λ^2 although deviations from this dependence occur at longer wavelengths.

Figure 1: Diffraction geometry: horizontal rotation axis of the crystal sample; X-ray beam from a vertically dispersing double crystal monochromator is essentially horizontally plane polarised. At ψ = 90° a diffraction spot is recorded in the equatorial plane and suffers no polarisation losses, at all θ values, in fact. At ψ = 0° polarisation losses would be serious at high 2θ values but these spots would occur in the blind region anyway. C = crystal position.



ALTERNATIVES IN MAD DATA COLLECTION AND ANALYSIS

Jeffrey T. Bolin and Janet L. Smith
Dept. of Biological Sciences, Purdue University
West Lafayette, Indiana, 47907-1392 USA

INTRODUCTION

Because of a variety of technical advances, particularly the increased availability of appropriate synchrotron data collection facilities, the determination of macromolecular crystal structures by methods that exploit multiwavelength anomalous diffraction (MAD) data is no longer a rare occurrence. Although earlier examples exploring the use of MAD techniques exist, it seems reasonable to date the establishment of the technique in 1988, when the structure of the cucumber blue Cu protein was solved by the interpretation of electron density maps phased exclusively by MAD methods augmented by solvent flattening [1]. Since then, most MAD structure determinations have followed the principles of experimental design and data analysis developed and described in the late 1980s by Hendrickson and coworkers [2,3]; we will treat the methods developed by Hendrickson and coworkers as the 'standard' techniques for MAD data collection and analysis. Recent reviews, which summarize current and historically significant applications as well as the methods, have been published by Hendrickson [4] and Smith [5]. Following a brief summary of the important experimental and computational principles, we will describe a few recent alternatives to the standard techniques.

THEORETICAL BACKGROUND

The MAD phasing method developed by Hendrickson and coworkers makes use of a description of the total structure factor amplitude (due to J. Karle [7]) that separates all wavelength dependence into terms that depend only on the scattering factor components of the anomalous scatterers. Because this description, represented by the form for a single type of anomalous scatterer in equation (1), allows an algebraically exact analysis, it is sometimes referred to as the "algebraic formalism". The total atomic scattering factor, f, is defined by (2) in the usual way as the complex sum of the wavelength-independent, 'normal' scattering factor °f and two wavelength-dependent anomalous scattering factors, f′ and if''. The quantities a(λ), b(λ), and c(λ) are ratios of these components, as defined in (3); the sign ambiguity in the equation for c(λ) accounts for differences between Bijvoet mates.

$$(1)\quad |{}^{\lambda}F_{obs}|^2 = |{}^{\circ}F_T|^2 + a(\lambda)|{}^{\circ}F_A|^2 + b(\lambda)\,|{}^{\circ}F_T|\,|{}^{\circ}F_A|\cos({}^{\circ}\phi_T - {}^{\circ}\phi_A) + c(\lambda)\,|{}^{\circ}F_T|\,|{}^{\circ}F_A|\sin({}^{\circ}\phi_T - {}^{\circ}\phi_A)$$

$$(2)\quad f = {}^{\circ}f + f' + if''$$

$$(3)\quad a(\lambda) = (f_\lambda'^2 + f_\lambda''^2)\,/\,f^{\circ 2};\quad b(\lambda) = 2(f_\lambda'/f^{\circ});\quad c(\lambda) = \pm 2(f_\lambda''/f^{\circ})$$

In (1), quantities marked by the degree symbol '°' are structure dependent but wavelength-independent, and quantities subscripted by 'T' and 'A' refer to the total structure and the structure of the anomalous scatterers, respectively. Thus $|^{\circ}F_T|$ and $^{\circ}\phi_T$ refer to the amplitude and phase of the structure factor that results from the 'normal' scattering of the total structure. In contrast, the terms $a(\lambda)$, $b(\lambda)$, and $c(\lambda)$ account for all wavelength-dependence and are nominally structure independent. It is this separation of wavelength dependence that lies at the heart of the method. A vector diagram showing the relationships in the complex plane between the various structure factor components may be found in [5].

EXPERIMENTAL DESIGN

The basic design of a typical MAD data collection, including the importance and characteristics of appropriate synchrotron beamlines and instruments, have been ably described in a number of recent reviews [4-6], and will only be briefly considered here. To appreciate the major aspects, one needs only recognize that any MAD phase determination depends on the measurement of *small* differences in structure factor amplitudes caused by the wavelength-dependence of the anomalous scattering of a few atoms. Consideration of plots illustrating the behavior of f' and f'' as a function of wavelength (see Fig. 1 in [4]) should establish that these factors change gradually except in a very narrow region associated with an absorption edge. To maximize the difference signals, one must use optics suitable for the measurement of these very sharp absorption spectra; the ability to provide energy resolution on the order of a few eV is one factor that distinguishes appropriate beamlines. A second criterion is the ability to change rapidly and reproducibly between wavelengths (see below).

The choice of wavelengths is driven by the need to maximize the differences, so that it is common to measure at the inflection point of the absorption edge, at the wavelength of maximum absorption, and at one or more wavelengths remote from the edge. To understand these choices, it is useful to consider separately how they maximize the dispersive difference signals, which arise from measurements made at different wavelengths, and the Bijvoet difference signals, which arise from F^+ and F^- measurements made at the same wavelength. The extreme value of f' occurs at the inflection point, so that large dispersive differences are found between amplitudes derived from measurements made at the inflection point, and amplitudes derived from measurements made at remote wavelengths where $|f'|$ is small. Bijvoet differences are largest at the point of maximum absorption, and are significant both at the inflection point and at any remote point of shorter wavelength where f'' is large. Thus the remote point of first choice is one of shorter wavelength that will provide significant Bijvoet differences as well as a strong dispersive signal.

The expected magnitudes of the dispersive and Bijvoet diffraction signals are described by diffraction ratios [5] that express the magnitude of the signal as a fraction of the average amplitude. Typically, the maximum values of these ratios are on the order of 3-5%. Thus a major consideration in experimental design is to seek to minimize all sources of systematic error by making all measurements that contribute to the determination of the phase for a given reflection from the same crystal as contemporaneously as possible. Pairs of reflections used for dispersive measurements should have identical Miller indices; in most circumstances, pairs

used for the measurement of Bijvoet differences are most conveniently observed as reflections related by a mirror plane, particularly if the detector has sufficient size to measure both reflections simultaneously. It is sometimes necessary to measure Friedel pairs by 'inverse beam geometry' in order to acquire the Bijvoet differences. In any case, the typical experimental design calls for the near-contemporaneous measurement of Bijvoet related reflections at three or more wavelengths from the same crystal, a total at least six measurements. Thus the optics and crystal orienter must provide rapid and reproducible changes in wavelength and crystal orientation, and a means for the periodic verification of the wavelength calibration.

Apart from the diffraction measurements, MAD data collection also requires the measurement of X-ray fluorescence or X-ray absorption spectra. Such spectra are a necessity to define the position and near-edge structure for the crystals under study in order to establish the optimum wavelengths for the diffraction experiments. In most cases to date, high precision spectra have also been used to determine experimental values for f' and f'' as a function of wavelength in the vicinity of the edge. The experimental spectrum is typically merged with a theoretical spectrum, and the result is used for the determination of the scattering factors by techniques similar to those described in [2]. Given well-determined scattering factors, the ratios $a(\lambda)$, $b(\lambda)$, and $c(\lambda)$ can be treated as knowns in the subsequent application of equation (1) for phase determination. As discussed below, recent developments suggest that use of of the fluorescence/absorption spectra to determine scattering factors may be bypassed.

PHASE ESTIMATION BY THE ALGEBRAIC APPROACH

A system of programs usually referred to as MADSYS has been developed for the purposes of grouping and scaling the MAD data as well as phase determination (see footnote 15 in [8]). It is beyond the scope of this article to describe each of the evolving procedures in detail, but a few general features should be understood.

An important consideration throughout the suite of programs is to maintain data groupings that are consistent with the measurement scheme. Thus multiple observations of a particular reflection at a particular wavelength, say $F_{\lambda 1}^{+}$, are not collected and scaled together. Rather, reflections are grouped according to the design principle governing contemporaneous measurements, as described above. Typically, local scaling of Bijvoet-related observations is the first step (program ANOSCL), followed by local scaling of all observations distinguished by wavelength (program WVLSCL).

On the assumption that reliable estimates are available for $a(\lambda)$, $b(\lambda)$, and $c(\lambda)$, data scaling is followed by analysis of the set of observational equations defined by equation (1) (program MADLSQ) to determine the unknown quantities $|^{\circ}F_T|$, $|^{\circ}F_A|$, and $(^{\circ}\phi_T - {}^{\circ}\phi_A)$ (see [2]). What remains is to recover the phase for the total structure factor, $^{\circ}\phi_T$. This is accomplished by defining the structure of the anomalous scatterers, so that $^{\circ}\phi_A$ may be calculated, by analysis of $|^{\circ}F_A|$, Bijvoet difference, or dispersive difference Patterson maps, or by application of direct methods to the amplitudes $|^{\circ}F_A|$. A 'hand' problem remains, in that these methods do not distinguish between enantiomorphic structures for the anomalous

scatterers. This problem is usually solved by evaluation of ($|^{\circ}F_T|$, $^{\circ}\phi_T$) maps derived from the two enantiomorphic structures. It is important to be aware that the maps so produced are not mirror images of one another [5], so that the proper 'hand' should be readily determined. Procedures also exist for the calculation of Hendrickson-Lattman phase coefficients (program MADABCD, [9]) and for the combination of phase information derived from multiple determinations of the phase for a given reflection (program MERGIT, [10]) should multiple contemporaneous observation sets exist.

ALTERNATIVES AND VARIATIONS

Most successful MAD structure determinations have involved proteins of modest size ([4], [5]). The extension of MAD phase analysis to larger proteins leaves room for further development of the method because of difficulties related to the problem of introducing a sufficient signal, the problem of assigning the structure of the anomalous scatterers, and the problem of measuring sufficiently precise data from weakly diffracting crystals [5]. It is not our intention to address these issues here. Rather, we wish to consider alternative MAD phasing procedures that represent responses to problems associated with recent or current applications: the problem of determining scattering factors and the problem of how to exploit incomplete MAD data to obtain phases for as many reflections as possible.

Alternatives in the Determination of Scattering Factors

The experimental determination of accurate anomalous scattering factors has been an important consideration in most MAD structure determinations. Nevertheless, the sensitivity of the scattering factors to undefined instability in monochromator settings as well as the potential anisotropy of the scattering [3], can limit the utility of scattering factors determined from a single X-ray fluorescence experiment. Fortunately, both problems have been addressed recently and it seems likely that the requirement for precisely determined scattering factors will be of less significance in future for phase determinations which rely on the MADSYS programs. Moreover, the requirement may also be reduced or eliminated by the use of alternative phasing procedures (see below).

An extreme example of monochromator infidelity and a discussion of the techniques used to diagnose the problem as well as to recover the affected data may be found in a report by W. Weis *et al.* on the structure determination of a Ca-binding lectin domain [11]. In essence, the anomalous scattering factors, which normally are used in the MADLSQ procedure in the form of defined ratios $a(\lambda)$, $b(\lambda)$, and $c(\lambda)$, were iteratively refined within MADLSQ, for each temporal block of data, in alternation with the usual refinement of $|^{\circ}F_T|$, $|^{\circ}F_A|$, and $\Delta\phi$. It should be noted that the anomalous diffraction in this case arose from the L_{III} absorption edge of Ho^{3+}, so that the anomalous scattering factors and diffraction ratios were relatively large. Other uses of the technique in cases where the anomalous diffraction signals are quite small have been reported [8], but we are aware of no data specifically comparing the relative merits of experimentally determined versus refined scattering factors in such a case. If the technique is indeed generally useful, it is reasonable to expect that the requirement for precise, experimental determination of scattering factors may be reduced or eliminated [11], particularly for commonly used scatterers.

Fanchon and Hendrickson [12] have recently examined the potential problem of anisotropic anomalous scattering in detail and have developed techniques for evaluating and using f′ and f″ as tensor quantities within the MADSYS procedures. Their general conclusion was that this effect "does not cripple the MAD method ... phases uncorrupted by these effects can be recovered".

Alternatives in the Calculation of Phases with Incomplete Data

It is an unfortunate fact that even the best planned MAD data collection is subject to the general problems associated with the use of synchrotron radiation facilities as well as the unexpected difficulties that limit the completeness of the data or lead to the acquisition of useless diffraction patterns. Thus it is not at all uncommon for an experimenter to design an elegant four- or five-wavelength MAD data collection experiment, with adequate counting times for every measurement, only to face the necessity of compromising the design because of time pressure. One may also discover after leaving the synchrotron that inadequate data were obtained because of problems that were not recognized while the experiment was in progress. When rational compromise is possible, the first choice is to reduce the number of wavelengths used. Nevertheless, a frequent result is the acquisition of an incomplete MAD data set, where the data needed for phasing by the algebraic approach are available for only a fraction of the reflections.

1. The use of alternative phasing programs such as MLPHARE

Several investigators have attacked the problem of incomplete or inadequate MAD data by calculating phases with procedures conventionally used for isomorphous replacement phasing in the presence of anomalous scatterers. A useful example is found in the recent work of V. Ramakrishnan *et al.* [13], who determined the crystal structure of the globular domain of histone H5 at 2.5 Å resolution based on MAD analysis of the selenomethionyl protein (2 Se per 89 amino acids). Phase calculations were performed both with the algebraic approach and by analyzing the MAD data as if they represented a series of isomorphous derivatives; solvent flattening was used for phase improvement in both cases. In the isomorphous replacement scheme, data from wavelength λ_1 were designated as the reference, or "native", data for a protein with intrinsic anomalous scatterers, whereas the data for two additional wavelengths were treated as if they came from heavy atom derivatives. In essence, dispersive differences, $(f'_{\lambda j} - f'_{\lambda 1})$, were treated as the equivalent of isomorphous replacement signals and used to refine variations in real occupancy factors whereas variations in anomalous occupancy factors for each wavelength were determined from f″ values. Refinement of the real and anomalous occupancy factors of the Se atoms was accomplished with the CCP4 [14] implementation of Otwinowski's maximum likelihood phase refinement program MLPHARE [15].

In this case, the MAD data were both incomplete and noisy. The diffraction measurements were made on beamline X12C at the National Synchrotron Light Source associated with Brookhaven National Laboratory (USA) at three wavelengths using two gain settings of an Enraf-Nonius FAST detector and more than thirty crystals. In terms of data completeness, the telling statistic is the fraction of reflections that could be phased by the algebraic approach, 70%, which compares unfavorably with the assignment of phases for 98% of the reflections by MLPHARE. Problems with data quality, in the sense of signal vs. noise, are

illustrated by a comparison of expected and observed Bijvoet diffraction ratios: the maximum expected ratio was roughly 5%, whereas the observed ratio at the same energy was more than 11% for acentric reflections and nearly 6% for centric reflections, for which no signal is expected. Thus the maximum expected signal is comparable in magnitude to an observed difference that should be indicative of the noise.

Both qualitative and quantitative assessments of map interpretability suggest that maps calculated with MLPHARE-derived phases were superior. For example, the correlation coefficient between a model phased map and the MLPHARE-phase mapped was 0.54 before solvent flattening and 0.67 after; the corresponding statistics for the other phase set are 0.43 and 0.67. The difference in map quality apparently arises in part from from a substantial difference in the completeness of the phase set, but also from a difference in phase accuracy. When maps were calculated with the subset (70%) of reflections for which both procedures assigned phases, the map correlation statistic was 0.48 for the MLPHARE-phased map and 0.43 for the map phased by the algebraic procedure [V. Ramakrishnan, personal communication].

A. Aggarwal and coworkers also used both MLPHARE and MADLSQ in their determination of the structure of selenomethionyl-modified endonuclease BamH I [16]. In this study, the MADLSQ phases were of high quality, as judged by value of the statistic $< \Delta(\Delta\phi) > = 30°$, which defines the mean of the differences between multiply determined values of the MALDSQ resultant $\Delta\phi$; for comparison, a similar statistic from the above case had a value of 38°. Nevertheless, an MLPHARE-phased map based on roughly 95% of the possible reflections was judged to be more continuous and interpretable than the MADLSQ-phased map, which included 85% of the possible data. It should be noted, however, that additional data sets were used in the MLPHARE phase calculations and that this comparison provides no information on phase accuracy.

The above examples illustrate the probable advantage of using a procedure such as MLPHARE, which yields phase estimates for a larger fraction of the reflections in an incomplete MAD data set. They also suggest that the phases calculated by MLPHARE may be more accurate if the MAD data are not of high quality. Are there any other possible advantages? One possible advantage relates to the heretofore standard use of experimentally determined anomalous scattering factors as known quantities in the algebraic approach. In an assessment of the utility of MLPHARE for MAD phase determination by E. Dodson [personal communication], it was observed that refinement of the real and anomalous occupancy factors by MLPHARE was remarkably accurate when performed on an arbitrary scale without assuming starting values for the anomalous scattering factors. The test case involved three-wavelength data for selenobiotinyl-bound core streptavidin, which contains 2 Se per roughly 130 protein residues and is expected to produce maximum diffraction ratios of about 2.5%. Real occupancies were refined against dispersive differences for centric reflections, whereas anomalous occupancies where refined against non-centric data with $F/\sigma \geq 3$; the resultant occupancies were typically within 5-10% of the expected values. Additional developments in the adaptation of maximum likelihood estimation to macromolecular phase determination offer the promise of eliminating the assignment of a reference data set as well as further reducing the heretofore stringent dependence of MAD phase

determination on accurate difference measurements [20].

2. The use of observation-specific phase probabilities

J. Smith has used an alternative response to the problem of incomplete MAD data in the determination of the structure of the glutamine PRPP amidotransferase from *B. subtilis*. Aspects of this work related to the function of the enzyme, data collection, determination of the structure of the anomalous scatterers, and early stages in the phase determination were described at the 1991 study weekend [17].

The asymmetric unit of the monoclinic crystal contains one 200 kDa homotetramer, which binds four Fe_4S_4 clusters, one per monomer. The anomalous scattering of these 16 Fe atoms was exploited in a three-wavelength MAD phase analysis. It is predictable that the application of MAD phasing methods to such a large structure will present difficult and perhaps new problems with respect to data acquisition and analysis. This is particularly true when Fe is the anomalous scatterer, so that anomalous diffraction must be measured at relatively low energy (K-edge energy ca. 7130 eV), and when the expected signal strength is low: in this case, the maximum diffraction ratios are on the order of 5-6% at low resolution (where the atoms of the clusters are not resolved) and 2.5-3% at high resolution. Thus, for a variety of reasons (see [17]), sufficient data to support MAD phase analysis by the MADLSQ program were available for only 40% of the possible reflections to 3 Å resolution, despite the fact that roughly 90% of the reflections were measured when data from all crystals and all wavelengths were merged.

Smith exploited the Hendrickson-Lattman-style probability equation for MAD phasing [9] to obtain separate phase probability coefficients for every observation in all data sets. The equations for the probability coefficients depend on the scattering factor ratios $a(\lambda)$, $b(\lambda)$, and $c(\lambda)$ as well as on estimates for $|^{\circ}F_T|$, $|^{\circ}F_A|$, and $^{\circ}\phi_A$, which are usually determined, as described above, in earlier stages of MAD analysis by the algebraic approach . By using the mean |Fobs| over all measurements as an estimate for $|^{\circ}F_T|$, using values for $|^{\circ}F_A|$ and $^{\circ}\phi_A$ calculated from the structure of the anomalous scatterers, and carefully placing both $<|Fobs|>$ and $|^{\circ}F_A|$ on an absolute scale, phase probability coefficients were calculated for individual observations. The resulting coefficients were combined to provide a joint probability distribution, and a net estimate for the phase, based on all available observations. Thus phases were obtained in a consistent fashion for 94% of the possible reflections to a resolution of 5.5 Å whereas the standard pathway through MADLSQ yielded phases for roughly 60% of the reflections to a comparable resolution (5.2 Å). A readily interpretable 3.0 Å resolution map was subsequently obtained by using these 5.5 Å MAD phases as the starting phase set for phase improvement and extension by solvent flattening and fourfold molecular averaging [18].

3. Other possibilities: changing the overall data collection strategy

One of the fundamental principles of MAD experimental design is to collect all of the observations contributing to the phase determination for a particular reflection *from the same crystal as close together in time as is reasonable.* The experiences of J. Bolin and colleagues related to an ill-fated attempt to determine MAD phases for nitrogenase MoFe protein suggest that, at least in some cases, it may

be possible to violate this principle. Specifically, it may be possible to acquire data at a remote wavelength in a separate experiment either at the synchrotron or in the home laboratory.

By way of background, the MoFe protein is the component of Mo-dependent nitrogenases that catalyzes the reduction of N_2 to NH_3. It is a highly oxygen-sensitive, 220 kDa $\alpha_2\beta_2$ tetramer which binds 2 Mo and 30 Fe atoms in the form of 4 metal-sulfur clusters of two types; each cluster contains eight metal atoms. Crystals of MoFe protein from *C. pasteurianum* have space group $P2_1$, with a= 69.9Å, b= 151.2 Å, c= 121.8 Å, β= 110.2° and one tetramer per asymmetric unit. Given these unit cell dimensions and the fact that the dominant anomalous scatterer is Fe, so that data must be collected with X-rays of wavelengths near 1.75 Å, very few synchrotron beamlines (in 1991) were useful for a MAD experiment because of limitations in the physical size or location of the detector, or limitations in the accessible wavelength range. Thus a three-wavelength data collection strategy was planned for the multiwire area detector station at SSRL (beamline I-5AD). Since the crystals grow in a habit that makes data collection by rotation about the *b* axis impractical, an inverse beam data collection strategy was required to measure Bijvoet differences. Because of the low flux of the beam and the limited strength of the diffraction from the crystals, the original experimental plan was abandoned in favor of measurement of the unique data at the wavelengths that would have been used for the full experiment: 1.742 Å, 1.737 Å, and 1.542 Å. It was hoped that dispersive anomalous differences could be extracted from these measurements and used to augment existing phase information from MIR and single-wavelength (Cu-Kα) anomalous diffraction experiments.

In fact, this hope was justified in the sense that the dispersive differences between the 1.742 Å and 1.542 Å data sets were measured with sufficient precision to give strong, if not dominant, peaks for the metal clusters in a dispersive anomalous difference Patterson map and to contribute to the solution of the phase problem as outlined above (see also [19]). However, it subsequently was discovered that improved Patterson maps, as judged by peak heights, and phases of equivalent or better quality, as judged by figures of merit after phase combination with the other phase sources, could be obtained from differences calculated using the SSRL 1.742 Å and Cu-Kα data measured on a conventional X-ray instrument. The superiority of the Cu-Kα data is likely a consequence of the fact that they were measured for single-wavelength anomalous diffraction phasing and the data set is of very high quality: the R_{merge} is 5% for a data set that is more than 95% complete, with sixfold overall redundancy, to 2.7 Å.

These results suggest that it may be practical to collect only the edge and inflection point data sets under the usual conditions for MAD exeriments, for it is only these measurements that demand high energy resolution and must experience whatever reduction in beam intensity accompanies the associated optical requirements. If data collected at energies remote from the edge can be measured under different conditions and at different times, it is possible to foresee several advantages: more counting time can be devoted to the difficult measurements at the edge resulting in higher precision data; a wider bandpass can be used for the measurements at remote energies, also leading to an increase in precision; the demands are reduced on the monochromator and the optical alignment of the

instrument with respect to rapid and reproducible settings at different energies; a smaller fraction of the total time will be spent changing the wavelength and/or optimizing the instrument alignment.

Acknowledgments

The authors acknowledge support from the US Dept. of Agriculture (NRICRGP # 91-37305-6661 to JTB), the US Public Health Service (DK42303 to JLS), and the Lucille P. Markey Foundation.

References

1. Guss, J.M., Merritt, E.A., Phizackerley, R.P., Hedman,B., Murata, M. Hodgson, K.O., and Freeman, H.C. Science *242* (1988) 806
2. Hendrickson, W.A., Smith, J.L., Phizackerley, R.P. and Merritt, E.A. Proteins: Struc. Funct. Genet. *4* (1988) 77
3. Hendrickson, W.A., Pähler, A., Smith, J.L., Satow, Y., Merritt, E.A. and Phizackerley R.P. Proc. Natl. Acad. Sci. U.S.A. *86* (1989) 2190
4. Hendrickson, W.A. Science *254* (1991) 51
5. Smith, J.L. Curr. Opin. Struct. Biol. *1*, (1991) 1002
6. Fourme, R. and Hendrickson, W.A. in Synchrotron Radiation and Biophysics edited by Hasnain, S.S. (1990) Ellis Horwood Ltd, Chichester, UK, 156
7. Karle, J. Int. J. Quantum Chem.: Quantum Biol Symp *7* (1980) 357
8. Leahy, D.J., Hendrickson, W.A., Aukhil, I. and Erickson, H.P. Science *258* (1992) 987
9. Pähler, A., Smith, J.L. and Hendrickson, W.A. Acta Crystallogr. Sect. A. *46* (1990) 537
10. Yang, W, Hendrickson, W.A., Crouch, R.J., and Satow, Y. Science *249* 1398
11. Weis, W.I., Kahn, R., Fourme, R., Drickhamer, K. and Hendrickson, W.A. Science *254* (1991) 1609
12. Fanchon, E. and Hendrickson, W.A. Acta Crystallogr. Sect. A *46* (1990) 809
13. Ramakrishnan, V., Finch, J.T., Graziano, V., Lee, P.L. and Sweet, R.M. Nature *362* (1993) 219
14. CCP4 (1979) *The SERC (UK) Collaborative Computing Project No. 4, a Suite of Programs for Protein Crystallography*, distributed from Daresbury Laboratory, Warrington WA4 4AD, UK
15. Otwinowksi, Z. in *Isomorphous Replacement and Anomalous Scattering*, edited by Wolf, W., Evans, P.R. and Leslie, A.G.W. (1991) Science & Engineering Research Council Daresbury Laboratory, Daresbury, U.K., 80
16. Newman, M., Strzelecka, T., Dorner, L., Schildkraut, I. and Aggarwal, A. abstract HH03 in *Abstracts of the American Crystallographic Association: Series 2* (Albuquerque, New Mexico, May 23-28, 1993)

17. Smith, J.L., Zaluzec, E., Wery, J.-P. and Satow, Y. in *Isomorphous Replacement and Anomalous Scattering*, edited by Wolf, W., Evans, P.R. and Leslie, A.G.W. (1991) Science & Engineering Research Council Daresbury Laboratory, Daresbury, U.K., 96

18. Smith, J.L., Zaluzec, E., Wery, J.-P, and Satow, Y. abstract MS-02.02.04 in *Abstracts, XVIth Congress and General Assembly of the International Union of Crystallography* (Beijing, 21-29 Aug. 1993)

19. Bolin, J.T., Campobasso, N, Muchmore, S.W., Morgan, T.W. and Mortenson, L.E. in *Molybdenum Enzymes, Cofactors, and Model Systems*, edited by Stiefel, E.I., Couccouvanis, D. and Newton, W.E., American Chemical Society, Washington, DC, (1993) 186

20. de la Fortelle, E., Bricogne, G., Kahn, R. and Fourme, R. abstract MS-02.02.03 in *Abstracts, XVIth Congress and General Assembly of the International Union of Crystallography* (Beijing, 21-29 Aug. 1993)

NORMAN: APPLICATIONS IN DATA ANALYSIS

by

G. DAVID SMITH
Medical Foundation of Buffalo, Inc.
Buffalo, NY 14203 U.S.A.

INTRODUCTION:

NORMAN [1] is a computer program which compares the amplitudes and standard deviations of two sets of measurements by a normal probability plot analysis [2]. The results of such an analysis provide a global view of individual differences in the entire set of measurements and can be used to identify individual outliers, subsets of the data which are in poor agreement, or the contribution of a heavy atom to a derivative data set.

For any large set of replicate measurements, the mean, the standard deviation from the mean, and the standard deviation of the population can be easily calculated. If there is a normal distribution of errors, the population of the differences from the mean divided by the variance is the familiar bell-shaped curve. One property of this curve is that it will have a mean of zero. The normal probability integral, equation (1), provides us

$$P(x) = (1/\sqrt{2\pi}) \int_{-x}^{x} \exp(-\alpha^2/2)\,d\alpha \qquad (1)$$

with an estimate of the percentage of data which should lie within a given number of standard deviations from the mean. Thus, for x = 0.674, P(x) is equal to 0.50, or 50% of the measurements will lie within 0.67 standard deviations.

The same principle can be applied to two sets of data obtained from an X-ray diffraction experiment. It is convenient to define δm(real) as:

$$\delta m(real) = (F_1 - KF_2) \,/\, (\sigma^2(F_1) + K^2\sigma^2(F_2))^{1/2} \qquad (2)$$

where K is a scale factor which minimizes the sum of the δm^2. Again, if the errors are normally distributed, the population of the δm(real), which corresponds to x in the expression for P(x) in equation (1), should produce a bell-shaped curve.

Although the expected value for δm(real) can be calculated for any distribution, its use in crystallography usually assumes a normal distribution of errors. Following the calculation of δm(real) for each of the j pairs of measurements, the collection is sorted in ascending order. For the ith value of δm_i(real), an order statistic is calculated from $|(j-2i+1)/j|$. This order statistic is the percentile ranking of each point within the entire

collection of differences and is equivalent to $P(x_i)$, where x_i is equal to δm(expected) and P(x) is the normal probability function, equation (1). The sign of x_i is positive for $i>j/2$ and negative for $i<j/2$. The calculation of the order statistic for each of the pairs of measurements, followed by a lookup and interpolation from a table of normal probabilities, provides individual values for δm(expected). A plot of δm(real) against δm(expected) should be linear with a slope of unity and an intercept of zero if the errors in the measurements are normally distributed. The calculation of the scale factor for equation 2 is not trivial; as a first approximation, NORMAN calculates the unweighted scale factor and then employs an iterative procedure to minimize the sum of the $\delta m(real)^2$ [1].

The equations of two least-squares straight lines as a function of δm(expected) are also calculated.

$$\delta m(real) = Slope * \delta m(expected) + Intercept \qquad (3)$$

The first includes all data which were used to generate the plot. The second is usually calculated using only those data with δm(expected) between plus and minus 0.674; this subset comprises the central 50% of the data, i.e., those data which are in the best agreement and generally deviate least from linearity. A comparison of the two least-squares straight lines provides an immediate indicator of the linearity of the entire plot. If one is attempting to verify the presence of a heavy atom in derivative data or to isolate the source of a potential error in the measurements which may be a function of resolution, it is very useful to perform an analysis of the deviations of the individual points from the 2nd least-squares straight line. NORMAN performs this analysis in several ways. If all errors were normally distributed, all of the δm(real) would lie on the straight line. Thus, δm(calc) can be obtained from the right side of equation (3). The difference between δm(real) and δm(calc) can then be used to calculate a residual, an average and rms deviation, and an average and rms deviation normalized to a slope of unity for each point of the plot as a function of resolution. These statistics are useful in identifying the extent to which the presence of a heavy atom contributes to the higher resolution data.

It should be emphasized that δm(real) is not only a function of the amplitudes but is also a function of their standard deviations. Thus, inaccurate estimates of the standard deviations will have a considerable effect upon both the slope as well as the linearity of the plots. Therefore, it is very important that the signal to noise ratio is optimal, since overestimation of the standard deviations of either or both sets of measurements will significantly reduce the values of δm(real) and mask real differences in the amplitudes. Likewise, underestimation of the standard deviations will suggest that there are significant differences in the two sets of data. Of course, it is pointless to perform any analysis of differences in the two sets of data if there is a lack of isomorphism.

The choice of using F or F^2 in the analysis seems to have little effect. However, caution must be exercised in using the weak data if the analysis is performed on F since the calculation of F and σ(F) from F^2 and σ(F^2) can result in an underestimate of F and an overestimation of σ(F). This problem is discussed in some detail by Blessing [3] and Rees [4].

There are many potential applications for NORMAN. It can be used to verify the presence of a heavy atom in a derivative data set, to compare independent sets of data, to identify those data which have the largest anomalous signal and to compare observed and calculated amplitudes.

APPLICATIONS:

The use of NORMAN to verify the presence of a heavy atom in a derivative data set has been described in some detail [1]. Reproduced in Figure 1 is a comparison of three putative heavy atom data sets with the native data for macromomycin [5], a relatively small protein (M_r = 11,000). Although Patterson maps were generated for all three derivatives, only the platinum derivative produced a high occupancy, single site. Examination of the probability plots shows that the platinum derivative has a relatively linear plot with a slope in excess of 20.0. A satisfactory solution for the Patterson from the iridium derivative was never obtained and the slope of this plot is just slightly larger than 5.0. For a larger protein, in which the heavy atom signal is a much smaller percentage of the total amplitude, one would expect to see smaller differences in the amplitudes and hence, smaller slopes. Nevertheless, if sufficient care is taken in the measurement of the amplitudes and the standard deviations, NORMAN can provide information regarding the presence or absence of a putative heavy atom as well as the extent to which it contributes.

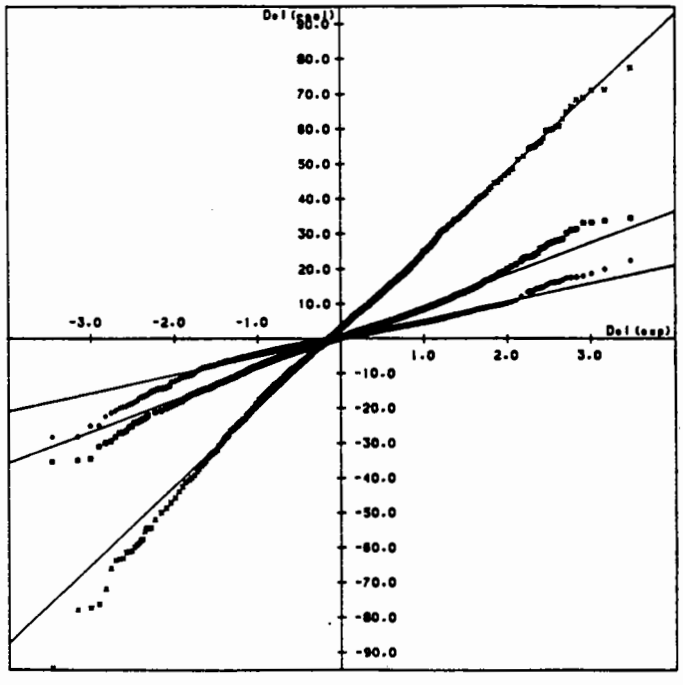


Figure 1. Plot of δm(real) versus δm(expected) for native macromomycin data versus three heavy atom derivatives; × is $K_2Pt(NO_2)_4$; ■ is $HgCl_2$; ◆ is $(NH_4)_3IrCl_6$.

In the original description of NORMAN [1], it was shown that both the resolution and sigma cutoff had little effect upon the slope and intercept of the resulting plot. This is shown in Table 1 by a comparison of data measured on two rhombohedral insulin crystals, obtained from the same crystallization vial. The first set of data were measured on an Enraf-Nonius CAD-4 diffractometer and the second set were measured at a later date on an R-axis image plate system. The normal probability plot for the first entry is illustrated in Figure 2. The plot is reasonably linear and the slope of 1.3

TABLE 1

Amplitude (F or F^2)	Sigma Cut-off	Resol.	R-merge	Slope	Intercept	Number of data
F^2	0	∞-2.5	11.1	1.31	0.36	3019
F^2	1	10-2.5	11.4	1.33	0.43	2644
F	2	10-2.5	9.3	1.39	0.36	2644
F^2	2	10-2.5	10.8	1.37	0.42	2260
F^2	2	8-3	10.7	1.41	0.24	1638
F	4	8-3	6.9	1.43	0.17	1638
F	2	3.5-2.5	14.3	1.24	0.31	1492

would suggest that either the standard deviations are underestimated or that there is a larger error in the amplitudes from that expected on the basis of a normal distribution of errors. The intercept, which is significantly different from zero, is somewhat more difficult to explain. In principle, the

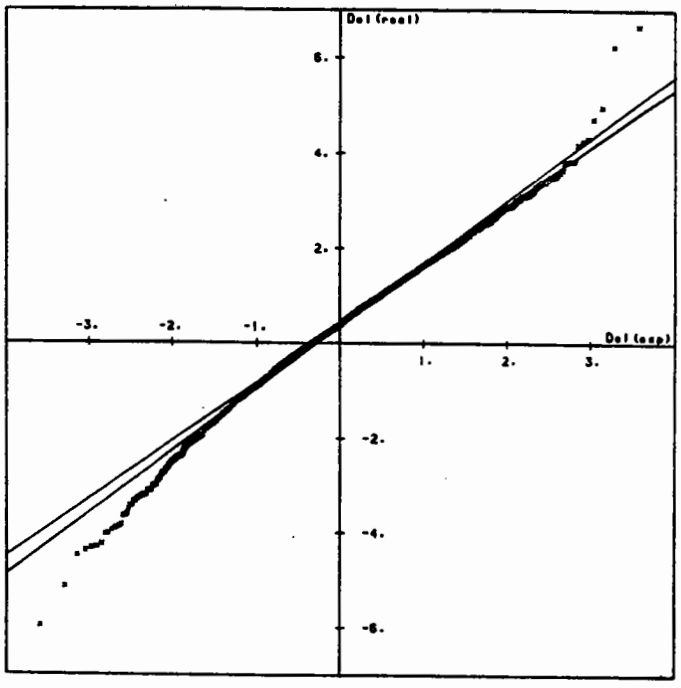


Figure 2. The normal probability plot of entry 1 in Table 1. The analysis was performed using F^2 (all data for which $F^2>0$) in the resolution range ∞-2.5Å.

number of positive and negative δm(real) should be equal. The positive intercept shows that there is a greater number of diffractometer amplitudes which are larger than the image plate amplitudes. This is also reflected by the fact that the average value of F^2 for the diffractometer data is larger than that of the R-axis data. There are several possible explanations for this phenomena: one is the absence of an absorption correction for the

R-axis data; a second might be the inability of the interframe scaling to completely correct for X-ray damage to the crystal; and finally, the quality of the R-axis data, obtained from an "aged" crystal may be less than optimal.

It is also useful to compare symmetry related reflections. A set of R-axis image plate measurements on a rhombohedral insulin crystal were processed as if they were triclinic. Each of the "triclinic" amplitudes was compared to the averaged rhombohedral amplitudes and the resulting plot

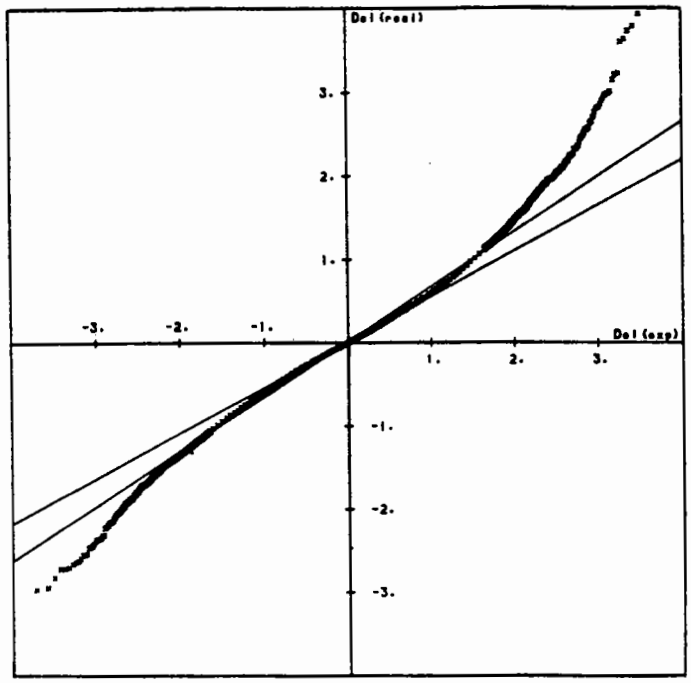


Figure 3. A normal probability plot analysis of data from a rhombohedral insulin crystal processed as triclinic compared to the rhombohedral averaged data. The comparison was performed on F^2 over a resolution range of ∞ to 1.9Å; R-merge was 0.02. The slope and intercept for the least-squares straight line using all data was 0.660 and 0.01, respectively, and 0.55 and 0.01 for the central 50% of the data.

is shown in Figure 3. Although the extremes of the plot are somewhat sigmoidal, particularly in the upper right corner, the plot is linear between plus and minus 2.0 in δm(expected) which comprises 95% of the data. While some deviation from linearity at the extremes is common, the sigmoidal behavior at the upper right suggests the presence of an additional source of error, perhaps differences in absorption. The slope of 0.66 suggests that the standard deviations are probably underestimated. An analysis of the deviations of the points from the least-squares straight line as a function of resolution shows a smooth decrease in the average deviation as the resolution increases (0.15 at 6.4Å and 0.05 at 2.0Å for the average deviation normalized to a slope of unity and resolution, respectively). Thus, from the point of view of a normal distribution of errors, the largest differences in the symmetry equivalent reflections are found in the lower resolution shells of data.

A δR plot is a very useful means to compare the calculated and observed amplitudes at the completion of the refinement. In this case, δ(real) is defined as:

$$\delta m(\mathrm{real}) = (|F_o| - K|F_c|) \,/\, \sigma(F_o) \qquad (4)$$

and the scale factor, K, is known from the results of the refinement. If one has confidence in the amplitudes which are calculated on the basis of the model, then the presence of outliers from the least-squares straight line are indicative of errors in the measurements. Figure 4 shows a δR plot which was calculated for a rhombohedral insulin structure at the completion of the refinement; the standard deviations which were used in the refinement and to calculate the δR plot were based upon an empirical weighting scheme.

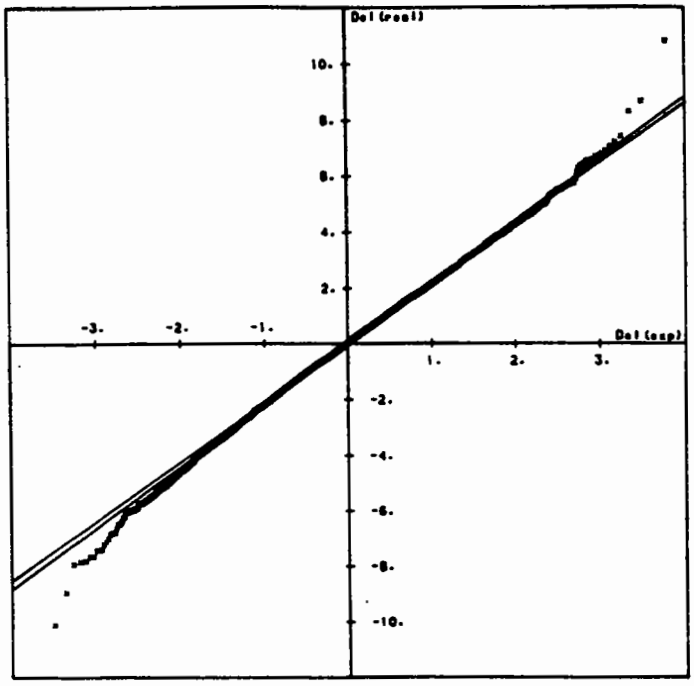


Figure 4. A δR plot, generated at the end of the refinement for a rhombohedral insulin structure. The final residual was 0.169 and all data used in the refinement (6316) were used to construct the plot. The resolution range was 8 to 1.9Å and the slope and intercept of the least-squares straight line for all data were 2.21 and -0.02, respectively.

It was noted earlier [1], that an analysis of two sets of data measured with either monochromatic or with polychromatic radiation appeared to agree reasonably well, but a comparison of monochromatic versus polychromatic data was significantly sigmoidal. At the time, it was suggested that this may be due to wavelength dependent effects. A preliminary set of Laue data on insulin were measured at DESY in Hamburg; the crystals came from the same vial as used previously to measure the R-axis and diffractometer data compared in Figure 2. A comparison of the Laue data with the monochromatic diffractometer data is shown in Figure 5. The data were processed using the Daresbury suite of programs [6] and an unmerged file containing lambda values was output; merging of these data was accomplished with the DREAD data reduction package [7]. Although the plot is reasonably linear, the intercept is distinctly different from zero and negative, showing that there are more differences for which the Laue amplitude is larger than the diffractometer amplitude. One might argue that this is indicative of the presence of wavelength dependent errors in the Laue data. However, this may also be due to the lower intensities of the diffractometer data in the higher resolution ranges as compared to the Laue data. For example, 75% of the Laue data fall between 3.5 and 2.5Å and of these, only 13% have $F^2<2\sigma(F^2)$; by contrast, 66% of the diffractometer data fall into this resolution range, but 40% of the intensities are less than $2\sigma(F^2)$.

Thus, the improvement of the signal to noise ratio in the Laue data may be responsible for the negative intercept.

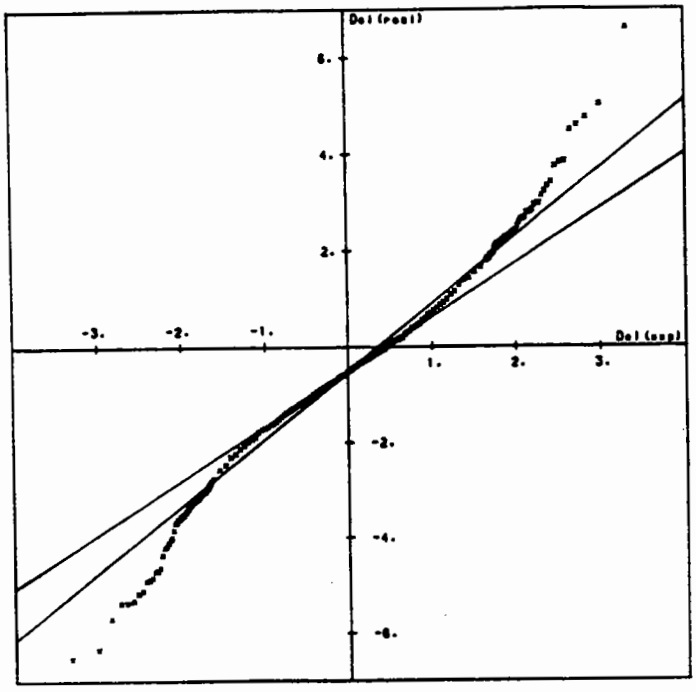


Figure 5. Normal probability plot of Laue versus monochromatic data. The analysis was performed on F^2 using data for which $F^2>2\sigma(F^2)$; the slope and intercept of the straight line for all data is 1.41 and -0.51, respectively.

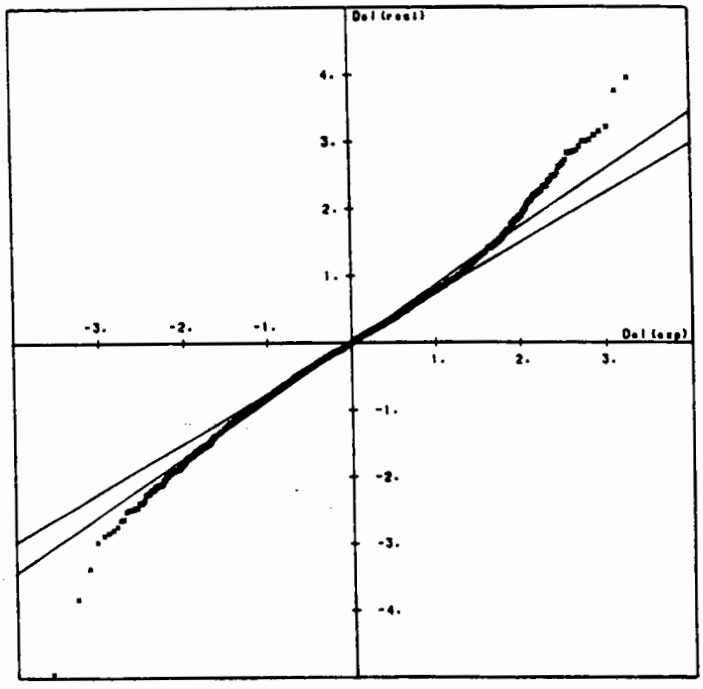


Figure 6. Normal probability plot of averaged Laue data against individual Laue data, measured at different wavelengths. The analysis was performed on F^2 using all data. The residual was 0.10 and the slope and intercept of the least-squares straight line for all data is 0.86 and 0.00, respectively.

The question of wavelength dependent errors still remains. However, the unmerged lambda output file from LAUENORM does contain numerous symmetry equivalents. If there are problems in particular wavelength ranges, one would expect to see significant differences if the unmerged data were compared to the DREAD averaged data. While this is not a comparison of two independent sets of measurements, it does provide some insight into potential problems; 489 of the averaged data were measured twice while 531

were measured three or more times. The resulting normal probability plot is shown in Figure 6. The plot is reasonably linear, with a slope of 0.86 and an intercept of zero. An analysis of the average deviation normalized to a slope of unity is illustrated in Figure 7. Although the largest deviations occur between wavelengths of 1.9 and 1.7Å and between 1.4 and 1.2Å, these deviations are not that much larger than the rest of the wavelength range and none are particularly large. From this analysis, one would conclude that differences in amplitudes, due to wavelength associated errors, are minimal.

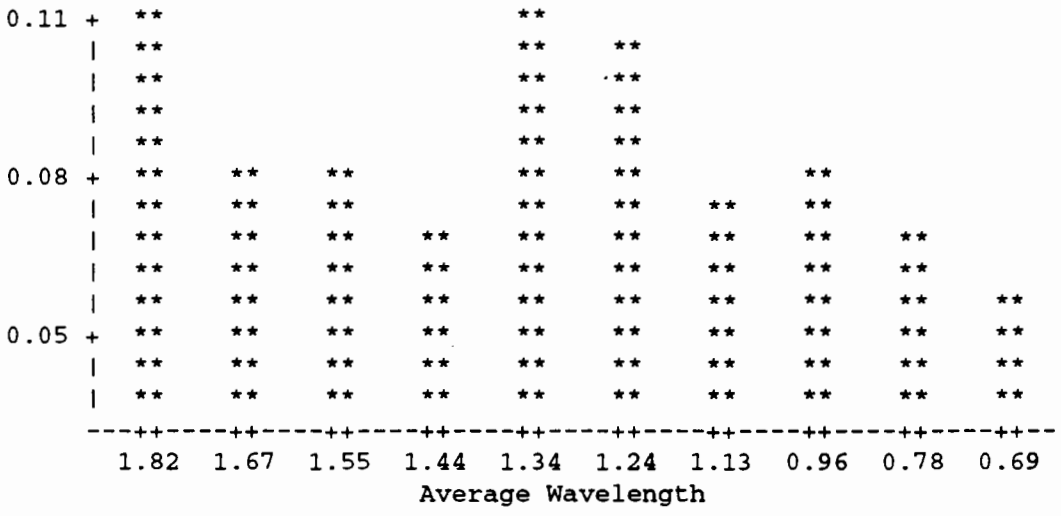


Figure 7. Average Deviation of each point normalized to a slope of unity from the equation of the least-squares straight line with slope 0.74 and intercept zero. Each bar represents 300 averages except for the one with an average wavelength of 0.69 which has 198 averages.

ACKNOWLEDGEMENTS: The author wishes to thank Drs. Robert Blessing and Lynne Howell for many helpful discussions, to Dr. Walter Pangborn for data measurement and processing, and to Dr. Patrick Van Roey for providing the data for macromomycin. This work was supported by a grant from the National Institutes of Health Grant No. DK-41387.

REFERENCES:

1. Howell, P.L. and Smith, G.D., J. Appl. Cryst., **25** (1992) 81-86.
2. Abrahams, S.C. and Keve, E.T., Acta Cryst., **A27** (1971) 157-165.
3. Blessing, R.H., Cryst. Rev., **1**, (1987) 3-58.
4. Rees, B., Israel J. Chem., **16**, (1977) 154-158; 180-186.
5. Van Roey, P. and Beerman, T.A. Proc. Natl. Acad. Sci., USA., **86**, (1989) 6587-6591.
6. Campbell, J.W., Clifton, I.J., Elder, M., Machin, P.A., Zurek, S., Helliwell, J.R., Habash, J., Hajdu, J. and Harding, M.M. in "Biophysics and Synchrotron Radiation", edited by A. Bianconi and A. Congiu Castellano, Springer Verlag (1987) 53-60.
7. Blessing, R.H., J. Appl. Cryst., **22**, (1989) 396-397.

DATA COLLECTION USING IMAGING PLATE SCANNERS

Zbigniew DAUTER

European Molecular Biology Laboratory (EMBL), c/o DESY, Notkestrasse 85,
D-2000 Hamburg 52, Germany.

At EMBL in Hamburg imaging plate scanners have been routinely used to collect X-ray diffraction data on macromolecular crystals for the last four years. The detectors in use are either the prototype constructed by Hendrix and Lentfer, or commercial MAR scanners with a plate radius of 90 or 150 mm. The results and conclusions presented here are therefore based on the experience with those machines. However they generally apply for other types of imaging plate detectors as well.

In some sense the imaging plate may be thought of as re-usable film. The active surface of the plate contains barium fluoride doped with Eu(II) ions which serve as colour centres, able to store the information about the incoming X-ray quanta in the form of a meta-stable excited state. Afterwards this information can be read-out by blue laser light which causes the colour centres to emit red light with intensity proportional to that of the recorded X-rays. The emitted red light can be measured by a photomultiplier system, pixel by pixel, and the resulting intensities stored directly on a computer disk for further processing after applying appropriate correction for individual pixel response. The image plate usually requires an additional erase by a flash of a hydrogen lamp and is then ready for the next exposure. The read-out of the imaging plate may be on-line and then takes place between exposures, or off-line using an independent scanner. The EMBL scanners work on-line, the image is scanned in a spiral manner with the plate rotating and read-out head moving radially. The spiral pixels need then to be transformed to cartesian coordinates for integration of reflection intensities. The dead time between exposures is about 2 min for small plates and about 4 min for the 150 mm scanner. The pixel size of our scanners is 150 μm . This allows the resolution of about 100 orders along the plate radius using the 90 mm scanner and well collimated synchrotron beam, and about 150 orders on the 150 mm scanner.

The properties of imaging plates make them in many ways superior to other types of detectors used in X-ray crystallography.

1) First of all, the relatively high sensitivity of imaging plates at short wavelength allows the use of radiation shorter than 1 Å on the synchrotron or molybdenum or silver Kα radiation (0.71 or 0.56 Å) from rotating anode or sealed tube. In fact imaging plate sensitivity drops to about 40 % below 0.5 Å but rises again almost to the same value at wavelength shorter than 0.4 Å, i.e.. the barium absorption edge. These characteristics are in sharp contrast to film, where using wavelength shorter than 1.5 Å requires much longer exposure times and additional foils between individual films within packs to attenuate recorded intensities. Due to this property of imaging plates the wavelength routinely used at synchrotron sites is below 1 Å. This minimizes absorption effects and radiation damage suffered by crystals. At such wavelengths it is very often possible to collect complete, high resolution data from a single crystal.

2) Imaging plates do not show any chemical fog effect and have very low intrinsic noise within their response.

3) Imaging plates display high linearity of response within a wide dynamic range of more than six orders of magnitude. This is in contrast to the logarithmic properties of film blackening measured by optical scanner. In practice the dynamic range of existing imaging plate scanners is limited by the number of bits transferred by the electronic read-out system. EMBL scanners have a dynamic range of 16,000 for the prototype and 64,000 or even 128,000 for the MAR scanners. However with strongly diffracting crystals on the synchrotron it is often necessary to record separately strong, low resolution intensities using short exposures. This procedure is recommended in order to ensure that strong, important reflections are not lost due to overloading of corresponding pixels.

4) The non-negligible scanning time between successive exposures is the reason that imaging plates are mostly used in wide oscillation mode. An extreme example of such application is the Weissenberg camera at the Photon Factory, which coupled with large plate size and highly collimated beam allows the use of oscillation ranges up to 20°.

Processing of the images, i.e. the integration of intensities of recorded reflections, can be done using existing standard software. In Hamburg we have used three programs: MOSFLM maintained by A. Leslie, DENZO written by Z. Otwinowski and XDS of W. Kabsch. Although these programs differ in detail, they each produce high quality data and we feel that selection between them should be left for users personal preference.

A point worth mentioning here is the treatment of standard deviations of measured intensities. In contrast to diffractometer measurements where σ is calculated directly from counting statistics, with imaging plates it can not be estimated so directly. The reason is that one quantum of red light measured from the plate does not necessarily correspond to one X-ray quantum stored beforehand. As a result the σ's may be at the wrong level. It is advisable to check after data merging by the so called t-plot if the σ's reproduce the error distribution with an average of zero 0.0 and standard deviation of 1.0 throughout all resolution ranges.

The examples given below illustrate some of the characteristics of the data collected. They are all from imaging plate scanners in Hamburg.

High resolution data collection on protein crystals

As mentioned before, imaging plates allow the collection of very high resolution diffraction data from macromolecular crystals. If short wavelength is used, below 1 Å, it is routinely possible to record complete data from a single crystal. All the examples of extra high resolution data listed in the Table below, were collected using a single specimen. All of them required at least two and some three sweeps to adequately cover very weak, high resolution intensities and very strong, low resolution reflections. The Table contains the information about the size of the protein, data quality and status of the refinement, where appropriate. The R values quoted are not final, as the work is not yet completely finished. Most of the examples refer to cooperation with other laboratories.

All these structures were solved earlier at lower resolution and are now being refined with anisotropic temperature factors using the program SHELXL of G. Sheldrick. The refinement is based on F^2 rather than F, as this allows the sensible use of even the weakest intensities. All reflections were included in the refinement without any sigma cut-off.

The preliminary results obtained so far allow some interesting observations.

The Luzzati plot for bacterial trypsin is shown for two methods in Fig. 1, for isotropic refinement of the model (with overall R value of 13.8 %) and for anisotropic refinement (R overall = 7.9 %). In both models hydrogen atoms were included at their calculated positions. At atomic resolution, close to 1 Å, the number of unique reflections exceeds the number of refined parameters (9 per atom in anisotropic case) by a factor of about 4. The anisotropic treatment of atomic vibration dramatically improves the agreement for the high resolution amplitudes.

Protein	kD per asymm. u.	Resolution Å	R(I) merge %	Refinement aniso, %	Cooperation
Rubredoxin	6	0.92	3.8	R = 7.0	1
BPTI	6	1.08	3.9	R = 10.5	2
ROP	6	1.08	4.5	R = 10.6	3
Protein G	6	1.10	4.0	R = 11 %	4
RNase Ap1	10	1.08	2.7	-	5
RNase Sa	20	1.20	3.6	R = 10.0	6
Bacterial trypsin	21	1.10	4.4	R = 7.9	7

1) L. Sieker (Seattle) and G. Sheldrick (Göttingen)
2) T. Schneider (EMBL Hamburg, in-house)
3) M. Vlassi and M. Kokkinidis (Iraklion)
4) D. Wigley (Oxford) and J. Derrick (Leicester)
5) K. Polyakov (Moscow)
6) J. Sevcik (Bratislava)
7) NOVO/Nordisk (Copenhagen)

The inadequacy of the model without hydrogen atoms and anisotropy of atomic vibrations is illustrated in Fig. 2 (a) and (b). They show the difference Fourier density around a tyrosine residue of bacterial trypsin after isotropic refinement with and without hydrogens in the model. Fig. 2 (b) clearly shows the density corresponding to anisotropic off-plane vibrations of the carbonyl oxygen atom. The corresponding map after anisotropic refinement shows no density at all at the same level.

The program SHELXL offers several possibilities for distance restraints. Apart from the two extremes of unrestrained refinement and restraints taken from a geometrical library, the bond distances may be made ´similar´ for all corresponding bonds for the same or analogous residues. In such a way we realized that some of the distances in existing libraries are inappropriate. Also it seems that the planarity of the peptide group should not be strongly restrained, as the ω angles for some residues show a discrepancy of up to 20° from planarity.

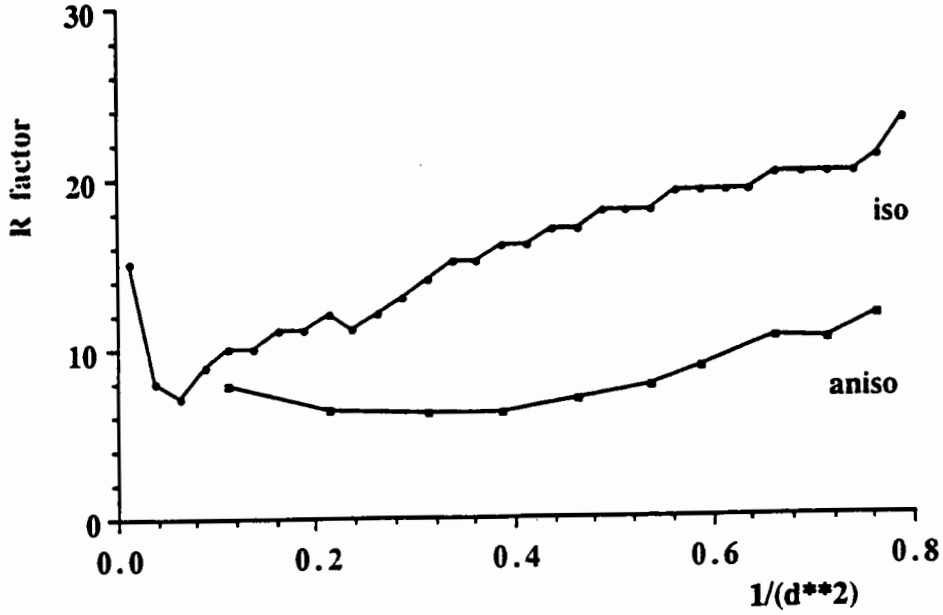
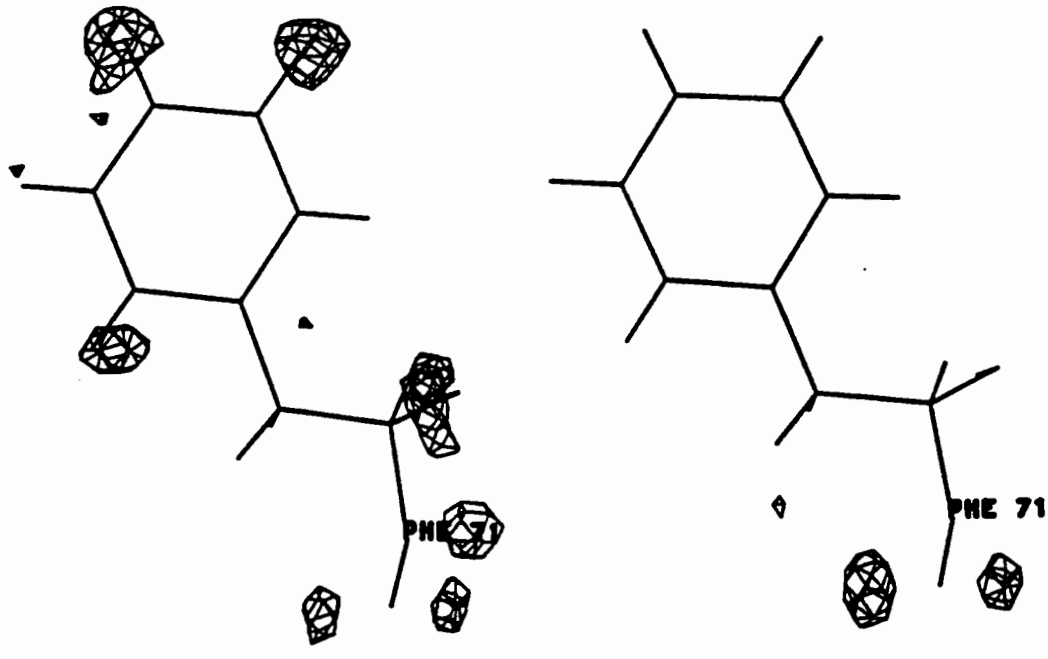


Fig. 1.R factor as a function of resolution for isotropic and anisotropic refinement.



(a) (b)

Fig. 2. Difference Fourier map around a tyrosine residue of bacterial trypsin contoured at 1.5 σ. (a) After isotropic refinement without hydrogen atoms contribution; (b) with hydrogens.

The data collected on rubredoxin crystals made it possible to solve this structure by direct methods (Sheldrick et al., 1993)

Small structures

To find how good data can be obtained from the imaging plate, we collected X-ray data on several small structure crystals. The size of structures ranged from 'supramolecular' cyclodextrin to urotropin with only three independent atoms. The small structure data were collected on a MAR scanner and Mo sealed tube source.

For crystals of small structures diffracting to a resolution of beyond 0.8 Å the effect of α_1: α_2 split of wavelength is significant. Fortunately the unit cells are so small that quite big raster box size can be used in integration without a danger of spots overlapping. Indeed, for small structures the accuracy of intensity measurements reach the theoretical limit of counting statistics; for strong reflections the R(I) merge values are in the range of 1.5 - 2.0 %. Taking into account that intensities measured from the imaging plate are in the range of I = 10,000 counts, their σ's can not exceed the value of $\sqrt{I}$ i.e.. 100. At this dynamic range of the detector, the expected limit of data accuracy is then about 1 %. This is confirmed by the results of refinement of such structures.

For β-cyclodextrin the refinement based on the 0.9 Å data collected in one day with imaging plate scanner resulted in an R factor of 3 %. On a diffractometer these crystals require several days to collect meaningful data and give an R factor of about 6 %. With data collected at low (100 K) temperature, it was possible to identify and refine the positions of hydrogen atoms including those on water molecules.

Urotropin (1,3,5,7-tetraazaadamantane) presents an extreme case. It consists of three independent atoms placed in special positions in the I43m cubic space group with a = 7.02 Å. The intensity data were collected at 100 K on an Ag tube to a resolution of 0.56 Å, resulting in a total of 122 unique reflections. The refinement of the structure (15 parameters including anisotropic thermal factors of hydrogen) resulted in an R factor of 2.0 % . This data will be subjected to multipole refinement of charge density distortions.

The results obtained clearly show the ability of imaging plate scanners to record very high quality data for proteins as well as for small molecule structures.

REFERENCES

Dauter, Z., Sieker, L.C. and Wilson, K.S. (1992), *Acta Cryst.*,**B48**, 42-59.

Kabsch, W. (1988), *J. Appl. Cryst.*, **21**, 916-924.

Leslie, A.G.W., Brick, P. and Wonacott, A.J. (1986), *CCP4 Newsl.* **18**, 33-39,

Otwinowski Z., (1991), DENZO: A film processing program for macromolecular crystallography. Yale University, New Haven.

Sheldrick, G.M., Dauter, Z., Wilson, K.S., Hope, H. and Sieker, L.C. (1993), *Acta Cryst.*, **D49**, 18-23.

Data reduction

P.R.Evans, MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH

I. Introduction

Once we have integrated intensities from any data collection instrument, we need to put all the data on a common scale, and derive a list of structure amplitudes |**F**| for a symmetry-unique set of reflections. This whole process may be called "data reduction", comprising:-

(a) placing reflections on the same scale (known & unknown corrections)
(b) rejection of outliers
(c) analysis for systematic errors
(d) averaging redundant measurements of intensity and estimating standard deviations
(e) compute |F| & σ(|F|), ΔFanom & σ(ΔFanom) from intensities

These various steps are discussed here. The literature on data reduction is surprisingly small, but a good review of many aspects of the process has been published by Blessing [1].

II. Putting data on same scale

The raw measured intensities are not all on the same scale, and the following reasons for this must be considered.

Calculable scale factors

- Lorentz factor - the relative speed of a reflection through the Ewald sphere. This correction is sometimes applied in the program (eg by MADNES [2]). It is straightforward to calculate from the diffraction geometry, but may be uncertain for reflections close to the rotation axis, since it is then very sensitive to any errors in orientation.

- polarization - correction for polarization of the diffracted beam by the crystal itself, and for the polarization of the incident beam. This is also sometimes applied in the integration program. The only uncertainty is for synchrotron radiation, where the polarization of the beam needs to be measured.

- partial recording of spots (unmatched partials). Partially recorded spots are usually assembled to full recording by adding together the two halves, but if there are a large number of unmatched partials, it may be worth scaling up partial intensities to the fully-recorded scale. This needs accurate orientation and cell parameters to predict the degree of partiality, and a good model of the rocking curve which relates angular fraction to intensity fraction. These parameters may be rerefined after assembly of the complete data (this is sometimes known as *postrefinement*).

Empirical scale factors

These are usually subsumed into a general scaling (see section III below)

- change of beam intensity - mainly on synchrotrons

- change of detector sensitivity - mostly for film or off-line image plates (different plates)

- different crystals

- illuminated volume, if the incident beam is smaller than the crystal (this is often the case for area detector data). This is indistinguishable from absorption in the primary beam

- absorption - not a problem at short wavelengths, but otherwise difficult to correct for satisfactorily (see below)
- radiation damage - difficult to correct for, should be minimized by quick data collection or cooling
- wavelength-dependent factors, for Laue diffraction

Absorption measurements.

Absorption corrections are often ignored (probably unwisely), or subsumed into the general scaling discussed below (section III), but for large crystals it may be worth measuring the absorption of the crystal by measuring the attenuation of the direct beam as a function of crystal rotation [3]. This can for example be done relatively easily with the FAST diffractometer. The beam must be smaller than the crystal in all orientations, so the method cannot be used for thin plate crystals. A special small collimator may be used. Alignment of the instrument so that the beam accurately intersects the rotation axis is essential. A weak beam is produced by turning the generator to low power, and the intensity measured for different rotations of the crystal about 2 axes, covering as many different directions as are physically accessible. This takes about 45mins on the FAST. Figure 1 shows the effect of applying such a correction, for a cubic crystal in which the high symmetry allows a simple scale and relative B-factor (in ROTAVATA) to account at least for the primary beam absorption. The effectiveness of the empirical absorption correction is shown by the scales and B-factors being much more uniform than if they have to correct for absorption as well as other effects.

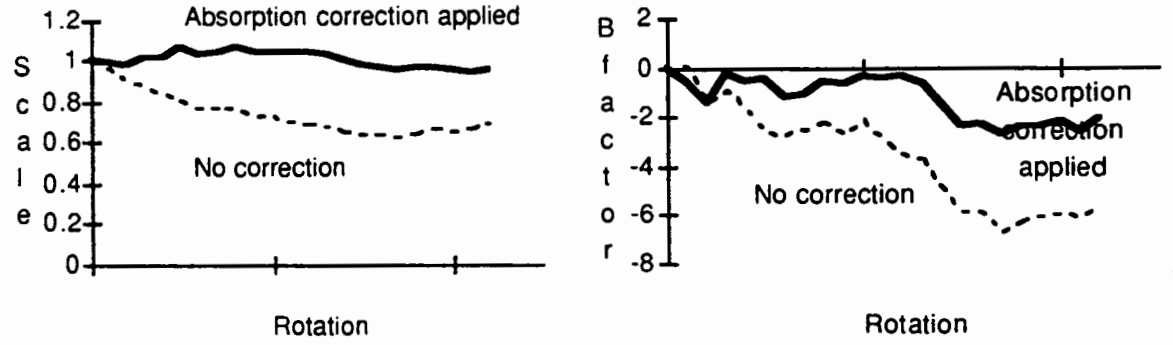


Figure 1. Application of an empirical absorption correction results in more uniform scale factors and relative B-factors. Data from a cubic crystal (spacegroup group $I4_132$), FAST area detector, 2.8Å resolution, 90° rotation.

III. General scaling.

In order for the scale factors to be determined, repeated measurements of the equivalent reflections need to be brought together: these include observations from different runs or different crystals. In the CCP4 package, this is achieved by reducing measured reflection indices *hkl* to an asymmetric unit (preserving the number of the symmetry operator used), then sorting the reflections on *hkl*. Each observation is also assigned a "batch" number, indicating its provenance: this can then be used as an index to a scale or set of scales. Typically, the batch number indicates the image number or a rotation range.

A scale factor k_{hi} is required for each observation I_{hi} (ie the i'th measurement of reflection **h**) which will put it on the same scale as all the others. This scale factor k_{hi} will be a function of other parameters. The usual practice is to determine scales from the internal redundancy of the data, but sometimes external reference data are available, eg derivative data may be scaled to a native set. External reference data may be treated either as fixed, or as a set of observations with the same status as the other observations. Scaling derivative data to the native set allows

the scaling parameters to minimize systematic differences between the data sets, leading to more accurate differences, and also allows merging of derivative datasets with insufficient internal redundancy for internal scaling, such as partial datasets for derivative surveys. Similarly, multiple data sets (eg from a MAD experiment) may be scaled simultaneously.

The scale factors are determined by minimizing the function

$$\Phi = \sum_h \sum_i w_{hi}\left(I_{hi} - \langle I_h \rangle / k_{hi}\right)^2$$

with respect to k_{hi} [4]. In this equation

I_{hi} is the i'th measurement of reflection h
k_{hi} is the scale factor belonging to that observation
w_{hi} is the weight for that observation, = $1/\sigma^2(I_{hi})$ (see next section for a discussion on weighting)
<I_h> is the current best estimate of the intensity $\quad \langle I_h \rangle = \dfrac{\sum_i w_{hi} I_{hi}/k_{hi}}{\sum_i w_{hi}/k_{hi}^2}$

Different forms of k_{hi} are discussed below, but the simplest case is when there is one scale factor per "batch", where a batch is for instance one image or film in a "rotation" data set.

Main data (rotation around b*) Cusp data

1099	122	3	0	0	0	0	0	0	0	0	0	0	0	0	0	4	337	104	60	34	0	0	5	35	61	138
122	1072	110	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	82	70	40	0	0	12	29	52	135
3	110	1159	107	0	0	0	0	0	0	0	0	0	0	0	0	0	0	117	91	54	7	0	16	43	49	142
0	1	107	1129	87	0	0	0	0	0	0	0	0	0	0	0	0	0	104	91	60	12	0	25	34	42	149
0	0	0	87	1156	71	0	0	0	0	0	0	0	0	0	0	0	0	104	109	85	37	5	45	23	52	157
0	0	0	0	71	1204	43	0	0	0	0	0	0	0	0	0	0	0	100	119	116	64	20	64	55	69	133
0	0	0	0	0	43	1192	26	0	0	0	0	0	0	0	0	0	0	70	108	148	139	116	111	51	46	65
0	0	0	0	0	0	26	1191	1	0	0	0	0	0	0	0	0	0	23	62	121	325	392	73	19	8	2
0	0	0	0	0	0	0	1	1253	17	0	0	0	0	0	0	0	0	5	29	138	409	394	14	0	0	0
0	0	0	0	0	0	0	0	17	1167	41	0	0	0	0	0	0	0	78	129	202	148	17	95	40	21	10
0	0	0	0	0	0	0	0	0	41	1176	69	0	0	0	0	0	0	113	150	165	46	0	41	22	43	78
0	0	0	0	0	0	0	0	0	0	69	1180	46	1	0	0	0	0	126	136	105	15	0	32	19	11	103
0	0	0	0	0	0	0	0	0	0	0	46	864	107	1	0	0	0	101	96	51	4	0	9	33	26	76
0	0	0	0	0	0	0	0	0	0	0	1	107	1104	132	2	0	0	113	93	51	5	0	6	45	47	100
0	0	0	0	0	0	0	0	0	0	0	0	1	132	1106	138	7	0	113	78	39	1	0	2	43	51	127
0	0	0	0	0	0	0	0	0	0	0	0	0	2	138	1067	142	3	99	76	29	0	0	1	33	50	137
4	0	0	0	0	0	0	0	0	0	0	0	0	0	7	142	1055	141	99	64	28	0	0	3	35	57	139
337	2	0	0	0	0	0	0	0	0	0	0	0	0	0	3	141	370	1122	83	24	0	0	0	38	74	177
104	82	117	104	104	100	70	23	5	78	113	126	101	113	113	99	99	122	845	0	0	0	0	3	39	164	231
60	70	91	91	109	119	108	62	29	129	150	136	96	93	78	76	64	83	0	850	0	1	8	51	119	149	2
34	40	54	60	85	116	148	121	138	202	165	105	51	51	39	29	28	24	0	0	820	11	68	156	144	1	0
0	0	7	12	37	64	139	325	409	148	46	15	4	5	1	0	0	0	0	1	11	914	439	199	1	0	0
0	0	0	0	5	20	116	392	394	17	0	0	0	0	0	0	0	0	0	8	68	439	815	0	0	0	0
5	12	16	25	45	64	111	73	14	95	41	32	9	6	2	1	3	0	3	51	156	199	0	416	0	0	0
35	29	43	34	23	55	51	19	0	40	22	19	33	45	43	33	35	38	39	119	144	1	0	0	310	0	0
61	52	49	42	52	69	46	8	0	21	43	11	26	47	51	50	57	74	164	149	1	0	0	0	0	379	0
138	135	142	149	157	133	65	2	0	10	78	103	76	100	127	137	139	177	231	2	0	0	0	0	0	0	932

Figure 2. Matrix of number of overlaps, 10° intervals, monoclinic, detector offset to measure on one side of beam only. The left group of columns & top rows refer to the main data, with the crystal rotated by 180° around b*, the right & lower groups to cusp data, with the crystal rotated by 90° around an axis perpendicular to b*

Whatever the form of k_{hi}, it is important that the scale factors are well defined by the redundancy in the data. In the simple batch scaling case, this means that each batch should overlap with a reasonable number of other batches, and not just its neighbours. Thus the strategy of data collection must be designed for good scaling, as well as complete coverage of reciprocal space. In particular, rotation about one crystal symmetry axis will not produce good definition of scale factors in space-groups of monoclinic or lower symmetry, particularly if the detector is offset to one side of the beam, as is common with electronic area detectors, unless a reference data set is available (eg scaling derivative to native). Rotation of the crystal around at least one other axis will generally provide "batches" which intersect all (or most) of the batches about the first axis, but in some low-symmetry cases data about more than two axes may be required. Even greater redundancy is required if secondary-beam absorption corrections are attempted. Figure 2 shows the scaling overlaps in a bad case, a monoclinic crystal rotated around the unique b^* axis, with the detector positioned to collect on one side of the beam only. The upper left part of the matrix shows the overlaps clustered around the diagonal, with no information relating the reflections 180° apart. The cusp data (from a perpendicular rotation axis) provides scaling information which cuts across the whole main data set, so the two sets combine to give well-determined scale factors. Figure 3 shows the overlaps for an orthorhombic crystal rotated by 90° around a principal axis, collecting data on both sides of the beam. The scales are much better determined than for the monoclinic case. There is still not much overlap between data 90° apart, but the scales are probably well enough determined. Note that merging data collected about more than one axis may increase R_{sym} but nevertheless *improve* the data, because reflections which are not related by symmetry (and therefore do not affect the R_{sym}) are more accurately put on to the same scale. However, poor merging between data from different rotation axes indicates an inadequate scaling model.

2944	1524	606	120	57	39	28	19	19	14
1524	2691	1105	230	43	16	17	14	16	12
606	1105	2709	808	316	42	22	16	31	24
120	230	808	2232	667	353	83	41	35	62
57	43	316	667	2398	1003	432	150	111	149
39	16	42	353	1003	2617	1112	560	303	337
28	17	22	83	432	1112	2589	1147	728	638
19	14	16	41	150	560	1147	2683	1334	1263
19	16	31	35	111	303	728	1334	2803	1993
14	12	24	62	149	337	638	1263	1993	3215

Figure 3. Matrix of number of overlaps, 10° intervals, orthorhombic symmetry, detector measuring on both sides of beam

Of the corrections listed above, most are functions of the primary beam direction (illuminated volume and absorption in the primary beam), or of time or batch number, which is directly equivalent (beam intensity, detector sensitivity, radiation damage). Only absorption (and indirectly radiation damage, since it depends on resolution) is a function of the secondary beam direction. With an area detector, a large number of reflections are measured for each primary beam direction (or at any one time), so these scale factors are generally well determined. Absorption in the secondary beam is a function of the direction of the diffracted beam relative to the crystal axes, and there are many fewer reflections measured with the same beam directions, so extracting reliable absorption corrections from data redundancy is much more difficult. This is equivalent to considering the primary beam corrections as a function of one angle (or time), but the secondary beam corrections as a function of two angles.

Alternative formulations of the scale:

- batch scale factor

- batch scale factor & relative temperature factor $k_j \exp(-2B_j \sin^2\theta/\lambda^2)$. This is probably the most common formulation, and is used for instance in the CCP4 program Rotavata. The temperature factor is largely a radiation damage correction, allowing for the greater effect on high resolution reflections.

- Xengen uses a batch scale factor of the form $k_j + A_j s + B_j s^2$, but also has an option to use (a batch scale factor)×(a factor expanded in spherical harmonics on the diffracted beam vector) [Xengen documentation]

- Kabsch [5] determines scale factors on a grid of positions on the detector and in rotation angle, such that each observation contributes to scales at several positions, weighted according to its distance from them, and the applied scale is interpolated between the scales.

- Absorption corrections may be expressed as an analytical function parameterized on the primary and secondary beam directions, and then the coefficients determined in the least-squares scaling. Suitable functions include Fourier series [6,7,8], spherical harmonics [9], or a quadratic function [10]. Of these, the spherical harmonic expansion is probably the most physically reasonable. These functions tend to be ill-determined at the edges of the data, unless restrained in some way.

In all cases, some of the parameters (or combinations of them) are undetermined or poorly determined. In the simple batch scaling case, for instance, all the scale factors may obviously be multiplied by any factor. Thus in the determination of the scale parameters, the normal matrix may be singular or ill-determined, and needs to be filtered to remove small eigenvalues, avoiding parameter shifts in undefined directions [11]. Note that the scale factors are part of the *model*, not part of the data.

IV. Averaging intensity measurements.

Usually we want to average redundant observations of reflections, but it should be noted that such averaging represents a loss of information, particularly about the difference in systematic errors between different observations. There is a case that refinement, for instance, should be done against unaveraged data, and systematic errors included as part of the model. The average intensity is also required for all scaling methods (see above), but the process of averaging multiple measurements is not entirely straightforward.

We have from the integration program a series of estimates of the intensity of reflection **h**, I_{hi}, each with an estimate of its standard deviation σ_{hi}. Any of the integration programs in normal use will give a good estimate of the intensity, but the error estimate is much less reliable, for two reasons. Firstly, the usual estimate that Var(count) = count from Poisson statistics is a *biased* estimate, since an accidently high count gives a high σ^2, and accidently low count gives a low σ^2. In a set of true replicate measurements, all arise from the same underlying distribution and should have the same variance, equal to the "true" count. Secondly, σ_{hi} is usually an *underestimate*, since there are errors other than counting statistics, many of which lead to an additional error approximately proportional to I. It is common to adjust the initial estimate of σ_{hi} to allow for the higher errors on strong intensities, for instance as $\sigma'_{hi} = A\sqrt{\sigma_{hi}^2 + B\,\mathbf{I}^2}$ where *A* & *B* are parameters to be determined, but such inflation of σ should be based on <I> not the individual I_{hi}, otherwise this correction will also be biased.

Suppose we know (or have determined) the scale factor k_{hi} which puts this observation on a common scale, then we can estimate the mean intensity $\langle I_h \rangle$ as $\langle I_h \rangle = \sum_i u_{hi} k_{hi} I_{hi} \Big/ \sum_i u_{hi}$, where u_{hi} is a weight. If all the observations were true repeated measurements, then they would all be equally accurate (ie would arise from the same underlying distribution), so should have equal weights u_{hi}. However, usually not all observations of a reflection are equally reliable: they may come from strong & weak crystals for instance. It is usual to weight observations by their estimated variances, ie $u_{hi} = 1/(k_{hi}\sigma_{hi})^2 = w_{hi}/k_{hi}^2$, but because of the bias in σ_{hi}, this leads to a systematic (though small) underestimate of $\langle I_h \rangle$. Two alternatives ways of calculating a weighted mean are worth considering:-

(a) Hammersley (unpublished) has shown that the bias in the weighted mean, with weights from Poisson statistics, is -1 count, so he suggests adding 1 count to each observation before taking the weighted mean. This requires that the observed intensity is a true photon count and allowance has been made for the detector gain.

(b) a "scale-weighted" mean may be derived as follows [12]:- if the errors arise purely from counting statistics, and the X-ray background is zero, then if the true (unknown) intensity is J_h

$$\mathrm{Var}(I_{hi}) = \frac{J_h}{k_{hi}} \quad \text{then} \quad w_{hi} = \frac{1}{k_{hi} J_h}$$

$$\langle I_h \rangle = \sum_i \frac{I_{hi}}{J_h} \Big/ \sum_i \frac{1}{k_{hi} J_h} = \sum_i I_{hi} \Big/ \sum_i \frac{1}{k_{hi}}$$

ie observations from strong batches have a higher weight than those from weaker batches, which is intuitively reasonable. Even if the assumptions of zero background and pure counting stastistics are invalid, then this estimate of the mean may be more reliable than the conventional one.

We may calculate two estimates of the standard deviation of an observation from n observations.

a) from the weighted mean of the individual variances (*external*)

$$\sigma^2_{ext} = \frac{\sum_i u_{hi} (k_{hi}\sigma_{hi})^2}{\sum_i u_{hi}}, \quad = \frac{n}{\sum_i 1/(k_{hi}\sigma_{hi})^2} \quad \text{if } u_{hi} = 1/(k_{hi}\sigma_{hi})^2$$

(b) from the weighted scatter of individual observations (*internal*)

$$\sigma^2_{int} = \left(\frac{n}{n-1}\right)\left[\frac{\sum_i u_{hi} (k_{hi} I_{hi} - \langle I_h \rangle)^2}{\sum_i u_{hi}}\right]$$

These are estimates of the standard deviation of the *population* of measurements. The standard deviation of the mean of the population is given by $\sigma^2(\langle I \rangle) = \sigma^2/n$, ie this is smaller for reflections measured more times, provided that the measurements are statistically independent. Blessing [1] suggests that because measurements are *not* independent, it is better to use the larger population standard deviation and not allow for the multiplicity of measurement, but a reflection measured several times *is* more accurate than one measured only once, so this should be allowed for, and $\sigma^2(\langle I \rangle) = \sigma^2/n$ used. However, a reflection measured 100 times is not 10 times more accurate than one measured once: in practice, this is only a problem for reference reflections from 4-circle diffractometer data, which should be omitted from the final data.

σ_{int} is not an accurate measure of error unless the multiplicity is very high, but analysis of the mean value of $\sigma_{int}/\sigma_{ext}$ against intensity and perhaps $\sin\theta$ can suggest how σ_{ext} should be adjusted to give a better error estimate (eg to determine A & B in $\sigma'_{hi} = A\sqrt{\sigma^2_{hi} + B\,I^2}$. This is done in the CCP4 program AGROVATA).

V. Rejection of outliers

An ideal set of data will have no outliers, but in practice a few measurements may be wrong because of detector defects, blips in the source, etc. If a large number of reflections are deviant, there is something badly wrong with the data collection or processing, and that should be put right (see section on systematic errors). You should not normally reject more than perhaps 1% of measured data. A common cause of outliers is reflections behind the backstop: ideally, this should be dealt with by a detector mask in the integration program, rather than attempting to reject reflections from statistical tests. There is *no* justification for the unfortunately common practice of rejecting weak or negative measurements. The information that a reflection is weak is just as important in the structure determination as the information that it is strong. Rejecting negative observations while retaining positive measurements of the same reflection also introduces bias.

Outlier rejection, and detection of systematic error, is much easier if the data have a high redundancy. This also improves the accuracy of the averaged data, by averaging out systematic errors. Rejection algorithms need care: in a small group of 3 or 4 observations, it is not usually clear which is the odd one out, and if there are only 2 measurements it is impossible. Sensible decisions can only be made with some understanding of the physical reasons for deviant observations. The usual test is deviation from the mean as a fraction of the standard deviation. The test can be against the scale-weighted mean, which avoids a spurious measurement with a very low standard deviation causing rejection of all other observations, or can be against the mean of all other observations. It is often informative to look carefully at the list of rejected or monitored reflections for patterns which are not analysed specifically in the tests for systematic errors. A procedure has been proposed [12] for weighting down outliers according to their deviation from the initial estimate of the mean, rather than rejecting them. This has the advantage of treating all reflections in the same way, without arbitrary cutoffs.

VI. Systematic errors

In looking for systematic errors, we are looking for trends in the data scale which cannot be corrected by the scaling function we have chosen. This analysis may indicate that more elaborate scaling is necessary. The mean deviation (R_{sym}) can be analysed against various data collection parameters, such as batch number (or crystal number, rotation value, etc), resolution, position on the detector, fraction recorded, or anything else we can think of. This analysis then allows us to decide on the any parts of the data which should be rescaled, or rejected. For instance, the batch analysis may show that one crystal may disagree with another, or that a crystal should be considered as dead beyond some point in the data collection. The analysis against resolution shows how far the data really extend: the shell R_{sym} should not go much above 0.2 for reasonable data. Note that the overall R_{sym} is effectively weighted by the multiplicity of data measurement, so that if you have a lower multiplicity at high resolution, as with a asymmetric detector (such as the Sakabe Weissenberg camera), the overall R_{sym} will be lower than it would be with uniform coverage.

Data sets should be *complete*, ie they should not have large regions of reciprocal space missing, nor should the strong or weak reflections be systematically missing. Incomplete data will lead to distortions in the maps and refinement.

VII. Amplitude F from intensity I

The simple calculation of F from I is by taking the square-root.

$$F = \sqrt{I}$$

$$\sigma(F) = \frac{\sigma(I)}{2F} \text{ for large F, or } \sigma(F) = -F + \sqrt{I + \sigma(I)}$$

The second expression for $\sigma(F)$ gives more sensible values for small F: it arises from setting $[F + \sigma(F)]^2 = [I + \sigma(I)]$.

However, there is a problem of how to deal with negative observations. These will arise as a legitimate measurement of weak reflections, and should not be rejected since they do indicate that the reflection is weak. A sensible statistical treatment was suggested by French & Wilson [13], and is programmed in the CCP4 program TRUNCATE. This estimates F & $\sigma(F)$, assuming the observation of I is normally distributed with standard deviation $\sigma(I)$, and using the expected distribution of intensities, which is always positive. Thus we take the mean of only the positive part of the Gaussian probability distribution of I (centred on the measured <I>), weighted by the expected distribution (determined from all the observations in a resolution shell). Specifically, if $p_I(I|J)$ is assumed to be a normal distribution around the unknown true intensity J, and $p_J(J)$ is the expected distribution of intensities (the prior distribution), based on Wilson statistics and the observed mean intensity for each resolution shell, then F can be estimated by integrating over the unknown intensity:-

$$p_J(J|I) \propto p_I(I|J)\, p_J(J) \qquad \text{(Bayes theorem)}$$

$$E_J(F|I) = \int_0^\infty \sqrt{J}\ \ p_J(J|I)\, dJ \qquad \text{var}(F|I) = \int_0^\infty \left[\sqrt{J} - E_J(F|I)\right]^2 p_J(J|I)\, dJ$$

The effect of this procedure is to inflate the values of the weakest reflections to above $\sqrt{I}$, since they are likely to have been underestimated (this is obviously the case for negative observations), but it has a negligible effect on reflections stronger than about $I = 3\sigma(I)$. For negative observations, integration over the positive part of the distribution will lead to F close to but greater than zero.

Anomalous differences, ΔF, may be taken from F^+-F^-, F^+ & F^- each being taken from I^+ & I^- by the above procedure. A similar procedure for improving ΔF estimates has been suggested by Lewis & Rees [14], but as far as I know, has not been used.

In order to calculate F, the program TRUNCATE examines the cumulative distribution of intensities in resolution shells, $N(I/\Sigma)$, where Σ is the mean intensity in that shell. Most proteins follow the theoretical curves quite closely, and major deviations indicate either a serious problem with the data collection, or a marked pseudo-symmetry, or indeed an error in the indexing or the spacegroup.

References

1. Blessing, R.H. (1987), *Cryst. Rev.*, **1**, 3-58

2. Messerschmidt, A. & Pflugrath, J.W. (1987), *J. Appl. Cryst.*, **20**, 306-315

3. Schwager, P., Bartels, K. & Huber, R. (1973), *Acta Cryst .*, **A29**, 291-295

4. Hamilton, W.C, Rollett, J.S. & Sparks, R.A. (1965), *Acta Cryst.*, **18**, 129-130

5. Kabsch, W. (1988), *J. Appl. Cryst.*, **21**, 916-924

6. Stuart, D. & Walker, N. (1979), *Acta Cryst.*, **A35**, 925-933

7. Walker, N. & Stuart, D. (1983), *Acta Cryst.*, **A39**, 158-166

8. Schutt, C.E. & Evans, P.R. (1985), *Acta Cryst.*, **A41**, 568-570

9. Katayama, C. (1986), *Acta Cryst.*, **A42**, 19-23

10. Takusagawa, F. (1987), *J.Appl.Cryst.*, **20**, 243-245

11. Fox, G.C. & Holmes, K.C. (1966), *Acta Cryst.*, **20**, 886-891

12. Schwarzenbach, D., *et al.* (1989), *Acta Cryst.*, **A45**, 63-75

13. Blessing, R.H. & Langs, D.A. (1987), *J. Appl. Cryst.*, **20**, 427-428

14. French, S. & Wilson, K.S. (1978), *Acta Cryst.*, **A34**, 517-525

15. Lewis, M. & Rees, D. (1983), *Acta Cryst.*, **A39**, 512-515

PROBLEMATIC DATA COLLECTIONS: GIVE UP OR PERSIST?

Elspeth F. Garman.
Laboratory of Molecular Biophysics, University of Oxford.

1. Introduction.

I will discuss some of the problems encountered during data collections, but since the emphasis will be very much on difficulties, I would like to start with a word of encouragement. Of over 450 data sets collected in Oxford during the last 5 years, only about 5% gave any problem at all, so the odds are that your data collection should not be problematic! It is rare for us to abandon data, and very few data sets remain unprocessed. Once good quality and trustworthy data are obtained, the problems reduce to usually tractable software ones, but if the original data are not reliable, there is not much that can be done about it afterwards. Thus it must be stressed that it is well worth taking care to optimise your data collection.

Difficulties divide into roughly three classes: insurmountable (denoted by a * in the following text), surmountable with special measures (**), and surmountable if the experimenter acquires more knowledge (***). The question I will address is how to recognise that you have a problem, and whether you should proceed once you have identified it.

Taking the data collection in chronological order, problems may arise in any of three areas: the crystals, the data acquisition system, and lastly the data processing. Aspects of each of these will now be considered.

2. Problems with crystals.

Having obtained a crystal suitable for a diffraction test, find out before mounting it if there are any physical limitations on the length of the crystal mounting tube and position of the crystal in the tube. For instance for the Raxis II on 9.6 at the SRS, the crystal must be within 1 cm of the bottom of the tube, whereas on the Siemens area detector with a 3 axis Supper goniostat, the crystal must be between 1 and 2 cm from the base of the tube.

Once the X-ray shutter is opened, there are many problems that may arise with your crystals:

* a) 'Protein' crystal may be salt! If so, no low resolution reflections will be observed, and there will be spots around and beyond the solvent ring. Note that if your detector is at a crystal to film distance for, say, 5$\AA$ data at the edge, you must move the detector nearer or swing it in 2θ in order to confirm that the crystal is salt, rather than disordered protein. It is well worth knowing which, since you can then adjust your crystallization conditions accordingly.

* b) Crystal obviously twinned. Those which look twinned before mounting are bound to give difficulties at the data processing stage, unless one member of the twin has a very much smaller volume than the other. If at all possible, avoid suspect crystals.

** c) Crystal internally twinned. Sometimes crystals which look single under the microscope have parts of their volume with the lattice rotated relative to the rest, and data processing can be very difficult but is not always impossible. Inspection of raw data images at an **early** stage (i.e. **during** data collection) is **essential** to check for twinning, which can become evident by adding successive images together to build up, say, a 2° picture. Inspection of the indexing error in *hkl* can reveal that internal twinning is present, as shown in Table 1. Here 235 reflections are correctly indexed, and 135 have an index almost half an integer away, indicating that either the cell length should be double, or the crystal is twinned. The distribution for a well indexed crystal is presented in Table 2.

	error (h)									
Index	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
h	235	27	25	12	13	8	0	6	20	135

Table 1: Typical indexing error distribution for a twinned crystal.

	error (h)									
Index	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
h	412	60	8	1	0	0	0	0	0	0

Table 2: Typical indexing error distribution for a good crystal.

If twinned, it is advisable to take fresh data from a different crystal if at all possible. If no more crystals are available, the data should be processed with great care, trying to ensure that only one lattice is selected and used for refinement during integration. However these data may not be reliable.

d) Crystal is disordered:

* i) Along one axis. Figure 1 shows a 0.25° oscillation Siemens frame from a crystal of a DNA hexamer with porphyrin (M.Sanderson, C.R.C., Sutton) taken in Oxford. Here the situation is hopeless and it as a waste of time to collect data from such a crystal.

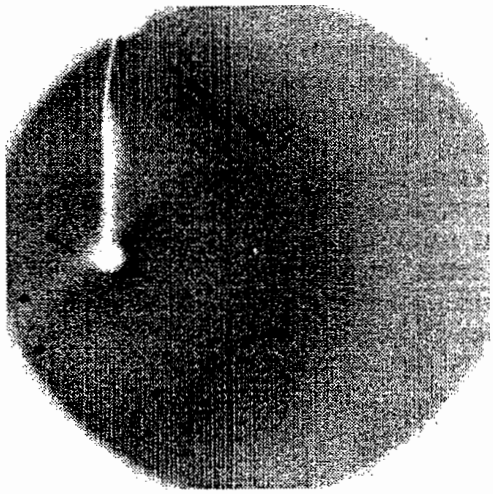


Fig.1 DNA hexamer with porphyrin showing disorder along one axis.

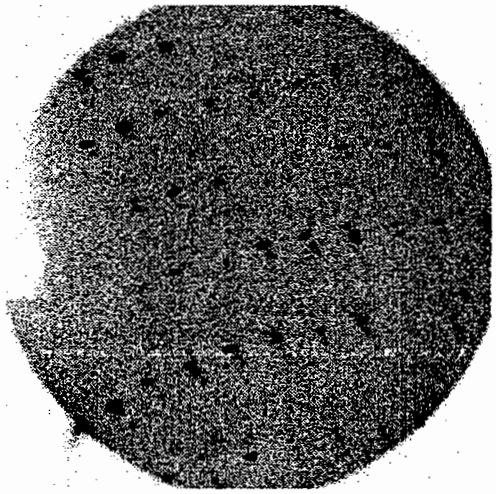


Fig.2 Nitroreductase showing high mosaic spread.

** ii) High mosaic spread. Figure 2 shows a 0.25° oscillation Siemens frame from a crystal of nitroreductase (cell: $57\mathring{A}\times60\mathring{A}\times258\mathring{A}$) (J.Skelly, M.Sanderson, C.R.C., Sutton) taken at Oxford. The mosaic spread of the spots was over 1°, and the data were unprocessable. In this case the problem was overcome, as it was found that freshly grown crystals exhibited reasonable mosaic spread, whereas those which had sat in the crystallization trays for more than 6 weeks were unuseable. In addition, even fresh crystals sometimes reacted badly to the journey from Sutton to Oxford on the motorway, but the acquisition of an in-house detector at Sutton solved that problem! Excess or rough handling of crystals can also induce an increase in mosaic spread.

** iii) Statistical or 'systematic' disorder. This phenomenon was first observed and analysed in 1954 in crystals of imidiazole methaemoglobin by W.Bragg and his colleagues (1-3). Other experimental reports and extensions of the original theoretical analysis have since been reported (4-6). The disorder is characterised

by a periodicity in the sharpness and diffuseness of the spots, as shown in Figures 3a and b.

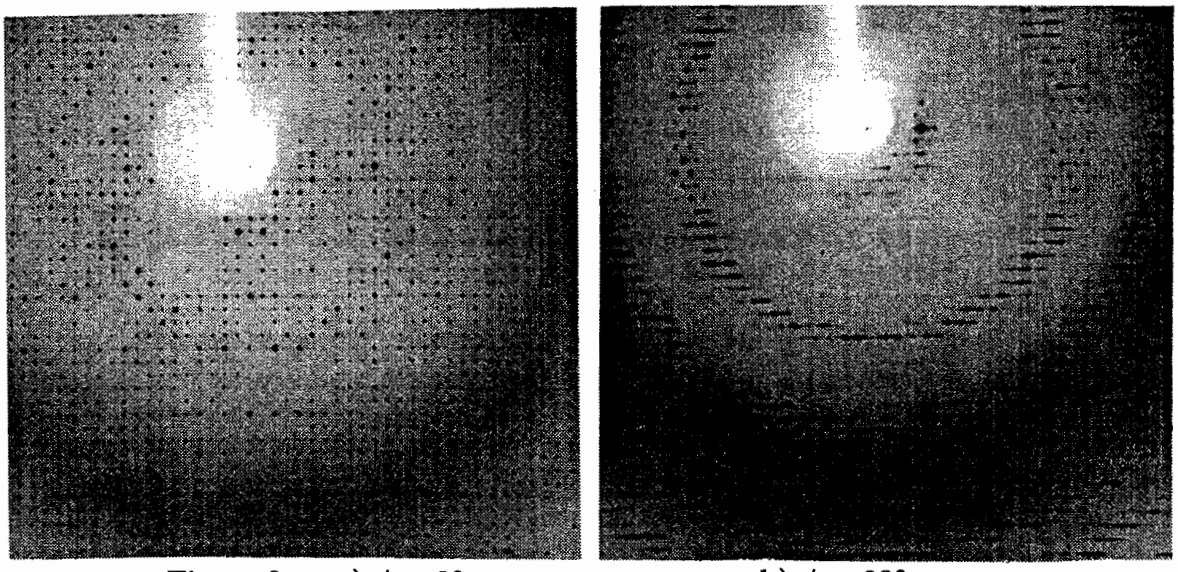


Figure 3 a) $\phi = 0°$ b) $\phi = 90°$
Activated spinach rubisco showing statistical disorder.

This is a classic example of such disorder in a crystal of activated spinach rubisco (I. Andersson and T.Taylor, Raxis SRS 9.6 $\Delta\phi = 1°$). The pattern at $\phi = 0°$ (3a) looks reasonable, apart from the rather high mosaic spread of some spots. However at $\phi = 90°$ (3b), some reflections are smeared out, and a periodicity in this is evident. This can be explained if layers in successive planes (say, the a-b plane) are systematically displaced along a or b. The displacements of successive layers are random. Can data from such crystals be useful? A recent analysis of data from statistically disordered crystals of p21 Ras protein (7) in Oxford led to a clear molecular replacement solution, stable to refinement, but which gave a poor quality electron density map. Inspection of the indices of measured reflections revealed gaps in the data along k, where reflections were too smeared to be accepted during processing, and it has not been possible to improve the map.

In principle it should be possible to 'correct' such data for the faulty layer stacking, but as yet, there is no record of this in the literature.

It is thus essential to check that your new protein crystal does not suffer from this sort of disorder by rotating it through 90° to take an image before embarking on a lengthy data collection.

*** e) Crystals are small. Data collection from crystals smaller than 0.1 mm in the largest dimension is best undertaken at a synchrotron. Down to 0.1 mm, a well collimated rotating anode source can be sufficient, provided that the crystals are reasonably well ordered.

*** f) Crystals have large unit cell. It is often inefficient to collect data on large unit cells in the laboratory, due to the high λ of the Cu_{K_α} radiation available (which means that high resolution reflections are at larger 2θ than if using lower λ radiation) and due to the comparatively small physical extent of some detectors (e.g. FAST, 4.7×6.3 cm^2, Siemens, diameter 11.5 cm) which limits the resolution of data collected in one sweep about the oscillation axis. If the crystal supply is limited it is better to collect data on crystals with cell lengths greater than about 260$\AA$ at a synchrotron, where lower λ X-rays and image plate detectors (diameter 18 cm and larger) are available.

** g) Crystal are weak diffractors. Again collection at a synchrotron is advisable, since as mentioned elsewhere in these proceedings, it is important to collect statistically

significant data.

** h) Crystals are very susceptible to radiation damage. As a rule, merging alot of small sections of data taken from different crystals is rather problematic, so ideally a data set should be obtained from a minimum number of crystals. As already described in these proceedings by Steve Gamblin, crystal lifetime can be made more or less infinite by flash freezing to liquid nitrogen temperatures, and if possible this is obviously the first thing to try.

For room temperature collection, keep a thermometer inside the radiation housing of your data collection instrument and monitor it, since many protein crystals cannot tolerate changes in temperature. Only one, in my experience, has actually lasted longer if warmed up (phosphoglycerate kinase, for which all data were taken at or above 21°C!!).

For collection at lower than room temperature, a column of silicon oil can be placed on either side of the crystal after the excess mother liquor has been removed. This seals the crystal in a small volume of the tube, keeping it hydrated and minimising distillation in the tube due to temperature gradients.

In all cases take care to **minimise** draughts around your crystal, again to avoid distillation of mother liquor in the tube.

Another important factor affecting radiation damage can be X-ray wavelength as generally damage is less at lower λ. Thus for susceptible crystals, collection on a synchrotron should be considered.

When is your crystal so damaged that you should stop data collection? This depends critically on how many crystals are available, how desperately the data are required, and why the data are needed. When the average reflection intensity has fallen by 30%, it is probably time for a new crystal, but bear in mind that the high resolution reflections disappear first, so if these are critical for your purpose, the crystal should be changed sooner.

*** i) Crystals dry out. Figures 4a and b show 2 images from a crystal of salmonella sialadase (1.7Å at edge of the image plate) which dried out during collection due to a cracked glass capillary. Observe the salt reflections appearing in the second image.

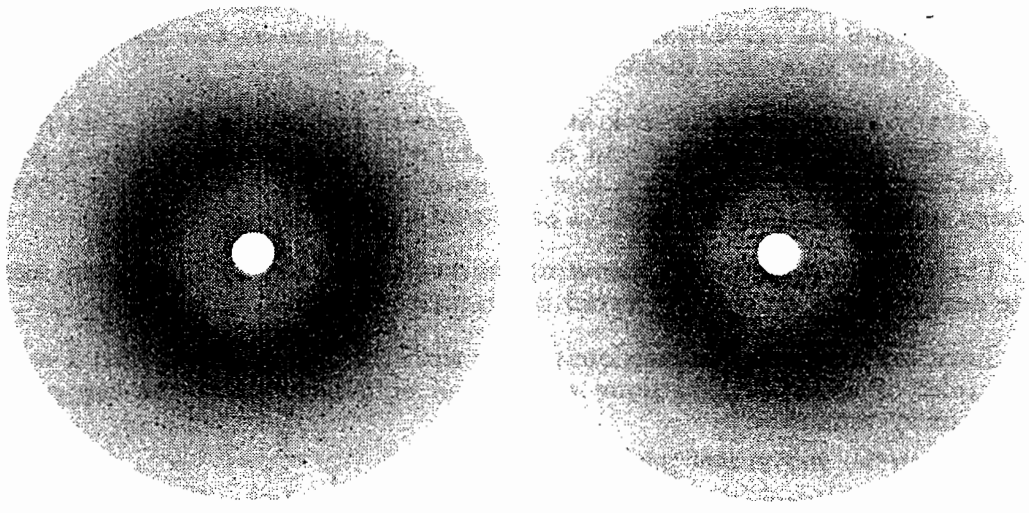


Figure 4 a) 10th image b) 29th image
Salmonella sialadase showing crystal drying out.

As protein crystals contain anything between 29% (edstin) (8) to 95% (tropomyosin) (8) solvent, if they are allowed to dry out they shrink. Usually (but not always!) this decreases their 3-D order, and hence their diffraction quality. The crystal must have some mother liquor near it and around it, but be dry enough so that it does

not move during data collection, since even a small slippage can make a data set unprocessable.

*** j) Crystals slip. If, in spite of your best drying efforts, your crystals still slip, leave your mounted crystal in the orientation that will be used for data collection to settle for a day or two before starting. Alternatively, use a flattened mounting tube to increase the contact area, and hence the adhesion, of the tube to the crystal. Note that we find the tube flattening block illustrated in (9), which holds the tube at both ends in a metal clamping arrangement, puts too much stress on the tube as it cools, and it is better to hold just one end down with a strip of plasticene.

** k) Crystals are very fragile. Some crystals are extremely fragile and difficult to mount (e.g. thin plates (0.05 mm thick) of beta-lactamase), while others (e.g. lysozyme) will withstand an amazing amount of handling and still diffract afterwards. For very thin crystals it is again worth flattening the mounting tubes, so that the crystal does not have to support its own weight. For very fragile crystals or if only a tiny amount (5-50μl per experiment) of protein is available, consider growing the crystals in the capillary tube (10).

3. Instrumental problems.

The following general points regarding data collection seem very obvious, but I include them because of bitter experience!

It is really vital to record your experimental parameters: crystal to detector distance, oscillation range per image, overall direction of oscillation, and main beam position. If refinement of the cell of a new protein crystal is problematic, the camera constants can be determined using a crystal of known cell, and then fixed in the refinement of the unknown cell, so reducing the number of uncertain parameters.

Incoherent background can be substantially reduced by minimising the collimator to backstop distance. For instance the collimation supplied with our Siemens detector had an exit aperture 38mm from the backstop. An extended collimator and a new backstop were designed and built which reduced the distance to 11 mm (4mm from the collimator end to the crystal, and 7mm from the crystal to the backstop), and we immediately obtained a 60% reduction in background. However, be warned that this close geometry can result in errant plasticene! Plasticene on the backstop can obscure useful low resolution reflections, and plasticene on the collimator can severely attenuate the X-ray flux incident on the crystal.

It really is important to look at a sequence of raw images during data collection to check that they are sensible and different. With the increasing automation of X-ray diffraction data collection, it is now common not to inspect the data beyond the first image until processing of the data afterwards gives a problem!

There is no substitute for knowing your own data acquisition system and learning the danger signs, as this can avoid much waste of time and crystals.

Since my experience is predominantly with the Siemens detector, I will give some examples of what can/has gone wrong, but equivalent problems can arise with any detector system.

a) Goniometer movement problems:

** i) Oscillation axis (ω) did not turn as expected. It is always worth noting ω at the beginning of any data collection, and then checking at the end that it is at the expected setting for the number of images that have been collected. When new, the motor speeds on our system were mis-set, and instead of consecutive frames of $\Delta\omega = 0.2°$ at 0.2° intervals, a typical sequence of frames might be at omegas of 0.0°, 0.0°, 0.4°, 0.6°, 0.6°, 1.0°... The motor would always end on the correct ω, but had been at unpredictable places inbetween, resulting in

data which were truly unprocessable. The Weizmann Institute system has also suffered from this problem. Our motor speeds are now regularly checked, since this behaviour is difficult to spot during data collection and has to be avoided! Inspection of the (x,y) coordinate of the brightest pixel is a good way of seeing it, since the same pixel will remain brightest and contain the same number of counts (within statistical errors) if ω is the same for consecutive frames.

** ii) ω and 2θ became correlated so that the detector, and thus the diffraction spots, moved in the x direction between frames. Adding 4 frames together resulted in a picture which could be mistaken for a sort of Weissenberg image! An overheating camera control box was the culprit in this case, and the box top is now left open to the air permanently.

b) Detector problems.

* i) No X-rays are detected (N.B. even in the absence of generated X-rays, cosmic rays give a flux of around 1 count/pixel/day). One likely diagnosis is a broken wire grid inside, and if this is so, the detector must be returned to the manufacturer for repair.

* ii) A dark strip corresponding to approximately a seventh of the active area appears on the detector with a thick bright line of excess counts next to it. This means one preamplifier is dead, which can be fixed in situ by a Siemens engineer. Data collection is inadvisable until the repair is carried out.

* iii) 64 bright and dark lines (or any other number $=2^N$) appear on the frame. This is caused by a loose connector or badly seated integrated circuit in the position decoding box, and can be remedied by reseating the connectors and pushing the integrated circuits in gently. No useable data can be taken with the detector in this state.

* iv) A thin (about 3 pixels) bright line is evident on the image. This can easily be discharged and thus cured in situ by an engineer, but the detector system is perfectly capable of quality data collection with the line present.

** v) A small dark spot appears in the centre of the detector. The main X-ray beam has been incident on the detector face, and there is now a localised deposit of hydrocarbons on the grids. This spot will not detect X-rays again until the grids are cleaned by the manufacturer. Data can be collected as usual, but to enable the main beam position to be measured, the detector must be shimmed up 1 or 2 mm so that a live part can be exposed to the attenuated main beam.

** c) Detector bias drift. The optimum detector bias should be checked every 8 weeks or so, since the 4 atmospheres pressure of Xenon gas in the detector chamber gradually leaks out, and a higher bias is required to maintain the detector signal. A re-gas is required approximately every 22 months. The highest bias setting for Cu_{K_α} X-rays at which useable data can be collected is around 9.45 on the potentiometer. Here spots at the detector edge become unacceptably smeared due to the increase in the path length along which the X-rays lose energy at reduced pressure, which increases the parallax effect. Large errors between fitted and observed pixel position ($\geq$0.5 pixels for rms x and y and $\geq$3 pixels for the maximum deviation) for the brass plate calibration indicate that the bias needs checking and adjusting.

Note that using the wrong bias can result in huge variations in the sensitivity of the detector across its face e.g. if the bias setting suitable for ^{55}Fe X-rays (5.9 keV) is used for data collection with Cu_{K_α} X-rays (8 keV), there can be a variation of 65% in sensitivity between the centre and the edges, instead of the $\pm$1% quoted by the manufacturer. Obviously this will ruin your intensity measurements, and the scaling of the data will either fail catastrophically, or worse, appear normal but result in seriously distorted data.

** d) Count rate considerations. The maximum global count rate for acceptable dead time is about 35 kHz in the Siemens detector. However, local saturation effects in

the detector can depress intensities for strong (usually low resolution) reflections because the dead time for them is locally higher than for the rest of the frame. This is only a problem for small unit cell crystals such as DNA fragments, (11) where the diffraction strength is shared between only a few reflections. The generator power must be decreased and the frames each collected for a longer time to ensure accurate data.

Note that for the Siemens PC acquisition system, data should **never** be collected with the real time accumulation of the display 'on', since this significantly increases the overall dead time.

*** e) The last class of instrumental problems are due to the 'random knob twiddlers' who roam most laboratories. Try to keep your data acquisition system in a restricted access room to minimise their chances!

4. Processing problems.

Data processing has been extensively covered by other contributors to this volume, so I will only mention a few common problems that we have encountered, and which will be all too familiar to most experimenters.

a) Problems autoindexing unit cell. This can be due to any number of reasons: crystal disorder, crystal twinning, crystal slippage, weak diffraction, misbehaving goniometer motors, and incorrect camera parameters and main beam position can all result in ill definition of the orientation matrix.

** i) Crystal slippage, if it is small or slow, can be tracked in the refinement during integration. For large or jerky movements, the data may have to be split into smaller sections between jerks, and merged after integration. Sometimes the slippage is so severe that the data have to be abandoned.

** ii) Unknown cells. For problematic cases when XDS (12) fails, we commonly use XENGEN Version 1.3 (13) for autoindexing and refinement of the camera parameters and cell, and then feed these into XDS for subsequent integration of the data.

** iii) Limited oscillation range of data (in ω or ϕ). If autoindexing fails on the first data frames, it is worth trying several subsequent sections and/or a larger section of data. For unit cells where one length is significantly longer than the other two, the orientation with respect to the long axis can remain uncertain until a wider oscillation range of data are included, or a section where the long axis is adequately sampled is found.

** b) Problems determining space group. The main message here is not to jump to conclusions! For example, an autoindexing solution for a new crystal of N6 viral neuraminidase gave a solution with low spindle and pixel errors of 106.85Å, 106.65Å, 74.4Å, 90.25°, 89.21° and 89.58° respectively. Since neuraminidase is a tetramer, space group P4 immediately suggested itself, and the data were processed accordingly. However the space group was eventually found to be P2 with unit cell 107.4Å, 74.6Å, 107.2Å, and β=89.6° (14) and the clue was not the relative R_{merge}s for the two space groups, but the number of rejected observations during scaling, which was 12% for P4 and 0.6% for P2.

Looking at the intensity weighted reciprocal lattice vectors on the display can be very helpful in space group determination (e.g. Program LATTICE (15) to convert .LCF, .MU or .HKL files for inspection with FRODO).

c) Problems merging data. Again there can be many causes for this.

** i) For P3, P32, P23, F23, F32, and I23 space groups there is an ambiguity in index choice, so this is the first check to make if, when merging two data sets in any of these space groups, the combined R_{merge} is significantly worse than those for the two separate data sets.

*** ii) Unit cell changes between data sets can give rise to poor merging statistics, since a 0.5% change in all cell dimensions will result in an average change of approximately 15% in reflection intensities within a 3Å sphere (16).

** iii) Merging data processed using different processing packages should not be problematic but sometimes is. Check the definition of the standard deviations output by each package, and also the rejection criteria for observations.

** d) Problems scaling data. Lack of redundancy can result in rejection of useful observations and the geometry chosen for the data collection can be a very important factor in this respect. For example, diffraction data for glucose 6-phosphate dehydrogenase ($P3_121$ with a=105Å, c=225Å (17)) were always collected with $2\theta = 7.5°$ to obtain a maximum resolution of 5Å. Crystals are tapered rods which were mounted with their long (c) axis approximately parallel to the capillary, which itself was mounted at 45° to the beam. This geometry resulted in a noticeable, and at some rotation angles dramatic, difference in the redundancy of observations in the top and bottom halves of the detector. Also, marked differences in the distribution of scaling information in reciprocal space (sometimes reasonably isotropic, sometimes very much dominated either by $h^2 + hk + k^2$ or by l^2) were exacerbated for this protein by anisotropic radiation damage. Since the scaling programme we then used divided the detector into a top and bottom half, treating these as separate shifts, observations made in the half with fewer redundancies could be rejected due to lack of information on how to scale them, or assigned scale factors inconsistent with those used for observations in the other half. Scaling without dividing the detector (there was no scientific reason to do so) resulted in much more reliable data.

** e) Ice rings on data from flash frozen crystals. These can look pathological, but we have rescued information from two such data sets by cutting the resolution to 4.6Å, and processing enough data to conclude that the substrate we had soaked into the crystal was not bound in it.

5. Conclusions

In spite of the catalogue of potential problems outlined above, most data collections go smoothly and the processing is routine. However, to avoid problems, there is no substitute for monitoring the raw data, understanding the data acquisition system, and actually inspecting the output of the processing programs.

I would like to thank all the users of the Oxford Siemens system over the last 5 years (more than 50 different experimenters) for letting me gain my experience on their data, and for permitting me to use some of that data in this presentation.

6. References.

1. Bragg, W.L. and Howells, E.R. Acta Cryst. *7* (1954) 409

2. Howells, E.R. and Perutz, M.F. Proc. Roy. Soc. *A225* (1954) 307

3. Glauser, S. and Rossmann, M.G. Acta Cryst. *21* (1966) 175

4. Cochran, W. and Howells, E.R. Acta Cryst. *7* (1954) 412

5. Pickersgill, R. Acta Cryst. *A43* (1987) 502

6. Luo, M., Laver, W.G. and Air, G. Acta Cryst. *A48* (1992) 263

7. Grimes, J. (1991) First Year Research Report, LMB, Oxford. Unpublished.

8. Matthews, B.W. J.Mol.Biol. *33* (1985) 491

9. Rayment, I. 'Methods in Enzymology' (Eds. H.W.Wyckoff, C.H.W.Hirs, S.N.Timasheff) *114* (1985) Pp. 136. Academic Press.

10. Phillips Jr., G.N. 'Methods in Enzymology' (Eds. H.W.Wyckoff, C.H.W.Hirs, S.N.Timasheff) *114* (1985) Pp. 128. Academic Press.

11. Clark, G.R., Brown, D.G., Sanderson, M.R., Chwaliinski, T., Neidle, S., Veal, J.M., Jones, R.L., Wilson, W.D., Zon, G., Garman, E., and Stuart, D.I. Nucleic Acids Research *18* (1990) 5521

12. Kabsch, W. J.Appl.Cryst. *21* (1988) 916

13. Howard, A. XENGEN Version 1.3 (1988) Unpublished.

14. Taylor, G., Garman, E., Webster, R., Saito,T., and Laver, G. J.Mol.Biol. *230* (1993) 345

15. Taylor, G.L. Program LATTICE. (1988) Unpublished.

16. Crick, F.H.C. and Magdoff, B. Acta Cryst. *9* (1956) 901

17. Adams, M.J., Levy, H.R. and Moffat, K. J.Biol.Chem. *258* (1982) 5867

Simple example of the molecular replacement technique: The structure determination of 3-phosphglycerate kinase from *Bacillus stearothermophilus*

by

Gideon J. Davies

Department of Chemistry, University of York, Heslington, York, YO1 5DD

Molecular replacement is now an extremely widely used method for the determination of protein structures from experimentally determined structure factor amplitudes. Before going on to examine some of the more complicated applications of the molecular replacement method, described later in this booklet, it is, perhaps, useful to study a simple "test case" and analyse the factors which can lead to success or failure of the method.

1. INTRODUCTION

The introduction of the molecular replacement method revolutionised protein structure determination. Given a suitable model structure, gone was the need for derivative screening, gone was the need for nasty heavy-metal reagents and gone were the problems of non-isomorphism. Whilst these facts are more-or-less true, what replaced these problems was a whole new set, which in the hands of the inexperienced or simply unlucky could lead to the failure to solve the protein structure, or worse still, to a completely incorrect structural determination. The structure determination of 3-phosphoglycerate kinase (PGK) from *Bacillus stearothermophilus* was not such a case, but analysis of the the molecular replacement method, as applied here, gives some insight into the use and applications of the method as well as suggesting reasons why the method can often fail to give the correct solution.

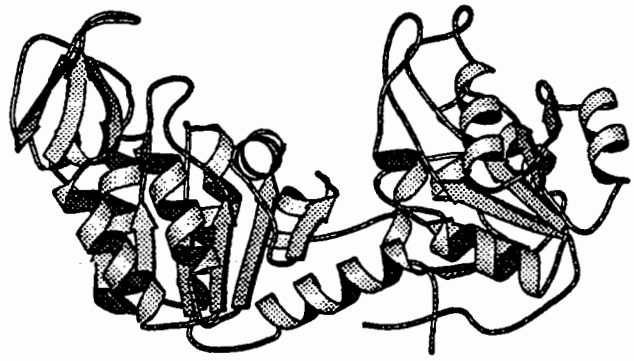
The molecular replacement method as used to solve the structure of similarly related proteins is a six-dimensional problem, normally split into two three dimensional calculations, the *rotation* and *translation* functions. As demonstrated by David Blow and Eleanor Dodson in the study weekend of 1985, these rotation and translation functions are most easily visualised by trying to superimpose two similar coffee cups ! In this paper we examine how two simple aspects of the rotation and translation function calculations, the accuracy of the trial structure and the completeness of the

observed data, can lead to failure to correctly determine the correct rotation and translation parameters when proteins rather than coffee-cups are involved.

2. THE SYSTEM

The work described in this paper was performed as part of a study into the factors involved in the thermal stability of proteins in Herman Watson's laboratory in Bristol. Phosphoglycerate kinase is a monomeric glycolytic enzyme which catalyses the first substrate level phosphorlylation in glycolysis. The enzyme consists of two domains (Figure1) which are believed to undergo a conformational change during catalysis. Two PGK structures had previously been solved at the time of this study, those from horse muscle (Banks *et al.*, 1979) and yeast (Watson *et al.*, 1982).

FIGURE 1. The structure of phosphoglycerate kinase



The structure of 3-phosphoglycerate kinase (PGK). The two domains of PGK are thought to change their orientations depending on the state of ligation, and the presence of ions such as $(SO_4)^{2-}$.

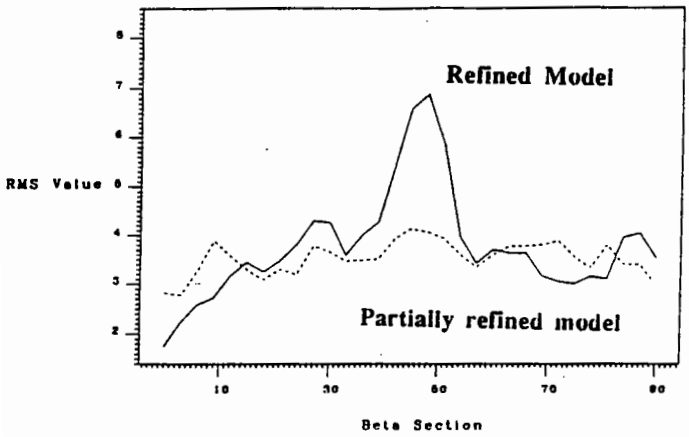
Crystals of *B. stearothermophilus* PGK were grown from polyethylene glycol in the presence of the nucleotide substrate MgATP. The space group was $P2_1$, with cell dimensions a=40.5Å, b=74.0Å, c=68.5Å and β=99.8°. There was one monomer of PGK in the asymmetric unit giving a solvent content of about 50%. The simplicity of this system made it an ideal candidate for structure solution by molecular replacement.

3 (a) THE ROTATION FUNCTION: *choice of model structure*

Initial attempts at molecular replacement with a completely unrefined model, the first build into the MIR map of yeast PGK, failed to give any significant features in the rotation function calculations. As beginners to molecular replacement this came as a surprise to us, for whilst the model obviously contained errors, these were generally small and randomly located across the whole molecule and we were only using data between 8 and 5Å. It is only when we performed trial calculations and found that the errors in an initial MIR built structure are often sufficiently large to prevent that model being used to solve *its own* rotation function that we began to appreciate the

need for a well refined model when trying to solve homologous structures ! Often, an initial MIR build will show a great discrepancy between the low resolution F_{calc} and F_{obs} which can be overcome with just a few cycles of refinement. An example of the effects of a partially refined model can be seen below in Figure 2. Two models were used, one is the refined Yeast PGK model (Watson *et al.*, unpublished) R=19% at 2.0Å and the other is a partially refined model with an R-factor of about 38% at 2.5Å. The two models are not vastly different, the differences are generally quite small but they are sufficient to make a large difference in the rotation fuction calculations.

FIGURE 2.



Rotation function output for *B. stearothermophilus* PGK. Calculations were performed with the MERLOT program (Fitzgerald, 1988). Data from 8-4Å were used together with an outer radius of Patterson integration of 25Å. This radius was chosen so as to maximise the number of self-vectors but limit the inclusion of Patterson origins from adjacent unit cells.

3(b) THE ROTATION FUNCTION: *data quality and completeness*

The rotation function, as calculated, measures the degree of overlap between two Patterson functions, one calculated from the model structure and the other calculated from the observed structure factor amplitudes. Patterson functions, calculated using F^2 terms, are obviously dominated by the stronger F's. The need for complete and accurate observations of the strong terms is clear, but is one that is often overlooked in data collection for molecular replacement calculations. It is all too easy to collect the highest resolution data possible from a crystal whilst neglecting the stronger and lower resolution F's, which often saturate the detector if good high resolution data are being collected. Even with the increased dynamic range of modern detectors, such as image-plates, measuring complete data from a well diffracting crystal may need as many as 3 data collection runs with varying degrees of exposure.

Data on the *B. stearothermophilus* PGK crystals, which diffract to beyond 1.6Å resolution, was collected in three stages. Initially a film dataset to 3.4Å was collected on station PX9.6 at Daresbury. Care was taken to ensure that there were no overloads, *i.e.*, that all the strong diffraction terms were measured. Two datasets were then collected at the EMBL Hamburg outstation using the Hendrix-Lentfer imaging-plate

scanner. A Table showing the completeness of the data, in the range used for molecular replacement calculations, is shown below:

Table 1 Completeness and quality of the *B. stearothermophilus* PGK data

RESOLUTION (Å)	Completeness	R_{merge}	
10.0	89%	0.030	
7.0	96%	0.028	
5.8	95%	0.028	
5.0	95%	0.029	
4.5	95%	0.031	
4.0	96%	0.032	(overall to 4Å, 0.030)

In order to assess the importance of the strong F's, and to simulate the effect of "overloads" (those reflections not measured due to detector saturation) different numbers of the highest intensities were removed from the rotation function calculations. The results are shown in Table 2 below:

Table 2. The effect of omitting strong observed amplitudes from the rotation function calculations

% of the strongest terms removed	Position of CORRECT answer in the peak list
0%	1st (!)
8%	not in top 100
6%	not in top 100
3%	20th
1%	6th
0.5% (only 15 reflections absent)	3rd

These results are quite shocking ! Even with a simple system, such as this, the absence of as little as 0.5% of the strongest observed intensities caused the correct answer to start to disappear into a sea of incorrect solutions. What is slightly more perplexing is

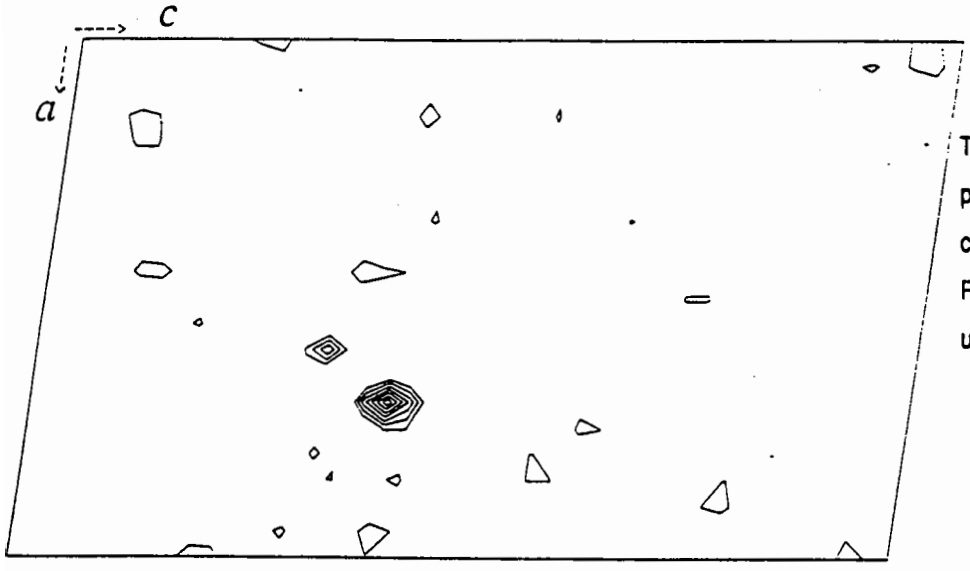
that the rotation function calculations carried out with the incomplete data are not featureless by any means. They contain a selection of *incorrect, but quite convincing* solutions, many of which could lead to a great deal of work running translation functions and the possibility of an incorrect structure.

The importance of these strong observed intensities cannot be over-stressed. Indeed using the, not insubstantial, benefits of viral point-group symmetry Rossmann and colleagues have demonstrated that it is possible to determine the orientation of a virus particle with less than 1% of the observed, *but strong*, diffraction data (Tong and Rossmann, 1990).

4. THE TRANSLATION FUNCTION

The *B. stearothermophilus* PGK crystals, described here, are monoclinic space group $P2_1$ with a single molecule in the asymmetric unit. The undefined origin along the crystallographic *b* axis means that a translation function search is limited to one over the *a-c* plane. With a single molecule in the asymmetric unit a simple R-factor serach was used using the CCP4 program RSEARCH (Figure 3). It gave one answer with an R-factor some 5% lower than at other positions on the search grid.

FIGURE 3. Translation function for *B. stearothermophilus* PGK



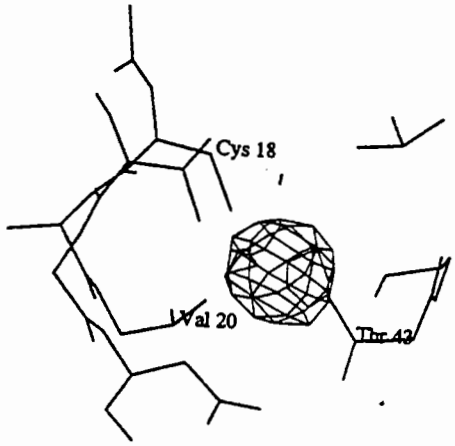
Translation function map in the *a-c* plane for *B. stearothermophilus* PGK calculated using the CCP4 program RSEARCH. Data from 10-5Å were used.

5. CONFIRMATION OF SOLUTION AND INITIAL REFINEMENT

It is useful to have independent means of assesing the validity of a molecular replacement solution. These are discussed in more detail in Eleanor Dodson's article later in these proceedings but those which were applicable to the *B. stearothermophilus* PGK structure determination are discussed here. Upon

obtaining the initial solution, the correctness of the packing of the molecules was assesed both on the graphics using FRODO and computationally using DISTANG (CCP4). There were no particularly bad steric clashes. We were fortunate during the *B. stearothermophilus* PGK structure determination to have a single heavy metal derivative dataset. This was a single-site mercury acetate derivative. A Hg difference Fourier, calculated with phases derived from the molecular replacement solution, gave a single large peak. Not only was this in the same position as the peak found in the difference Patterson synthesis, but superimposition of the model coordinates showed it to be only 4Å away from the sulphur atom of the single cysteine of *B. stearothermophilus* PGK suggesting that the solution was essentially correct (Figure 4). It should be noted, however, that peaks in the difference Fourier can be obtained when the molecular replacement solution is only partially correct, so they should be treated with some caution (Evans, 1985). For instance, in the example discusssed here an translation that was misplaced by 1/4 of a unit cell would give rise to a set of phases half of which would be correct.

FIGURE 4. Hg difference Fourier calculated with phases from the molecular replacement model.

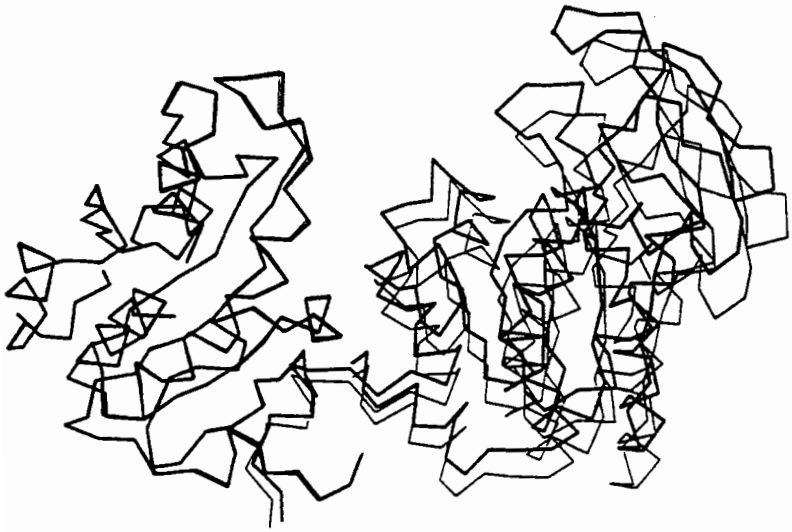


Section of the PGK Hg difference Fourier. The density peak shown is 14σ higher than any other feature in the synthesis. The cysteine sulphur to Hg distance is approximately 4.5Å.

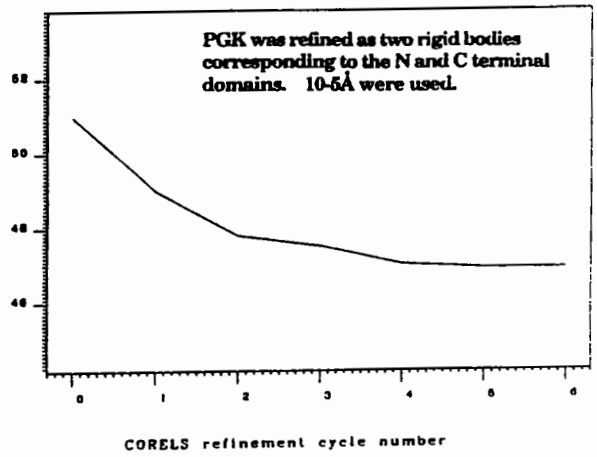
Having arrived at a sensible solution to the rotation and translation function problems, what is the most appropriate refinement method to choose ? For a simple two domain protein such as PGK a simple application of rigid body refinement seemed sensible. In addition, there was independent evidence to suggest that the domains would have a changed orientation due to the ligation state of the enzyme. Rigid-body refinement can now be performed by a number of programs (CORELS, XPLOR, TNT, *etc*) but in this study constrained rigid body refinement was carried out using CORELS, written by Joel Sussman (Sussman *et al.*, 1977, and for review see Leslie, 1985). *B. stearothermophilus* PGK was treated as two independent domains linked by a single "peptide" restraint in the centre of the helix linking the two domains. Initial rigid body refinement of the rotation and translation parameters for

the two domains using data between 15 and 7Å proved to be unstable, so data from 10-5Å were used. Seven cycles of refinement gave an R-factor drop of some 4% and gave a movement of the two domains of approximately 4.5° (Figure 5).

FIGURE 5. Rigid body refinement of *B. stearothermophilus* PGK.



Rigid-body movement of the two domains of PGK after refinement with CORELS. The pre-CORELS structure is shown in faint lines and the structure after CORELS in bold.



R-factor fall during the 7 cycles of CORELS rigid-body refinement.

It is our experience in this laboratory that initial rigid-body refinement of the molecular replacement solution is a most sensible way to proceed. Not only does the refinement leads to a better solution but it helps to "filter-out" incorrect molecular replacement solutions (see Derewenda 1985, 1990). It is often the case with protein structure refinement that initial difficulties with conventional refinement turn out to be due to a global change in domain orientation that could, perhaps, have been avoided with some sensible initial rigid-body refinement (see Swift *et al.*, 1991).

5. DISCUSSION

The advantage of molecular replacement is that it should be simple ! What we have tried to show here that it can be simple, but that errors in the model and the data can lead to molecular replacement becoming rather more tricky. If a couple of lessons are to be learned then they would be (1) always start with the best model and (2) measure all the strong intensities. Film and image processing programs such as DENZO and MOSFLM allow for the inclusion of a best estimate of the intensities of "overloads" in the final output file. Whilst one may not wish to use these estimates for refinement

purposes, it might well be worth including them for the purpose of determining the structure. At the study weekend itself, David Blow suggested that all data missing due to detector saturation should be included in the final dataset with "guestimated" intensity equivalent to at least the average intensity for the whole dataset, so that approach may well be worth trying.

ACKOWLEDGEMENTS

The work described here was carried out during my Ph.D studies in Herman Watson's laboratory in Bristol. The contributions of Herman Watson, Jenny Littlechild and Steve Gamblin to this work are gratefully acknowledged. Eleanor Dodson, Johan Turkenburg, Dale Wigley, Keith Wilson and Zbigniew Dauter are thanked for useful discussions.

REFERENCES

Banks, R.D., Blake, C.C.F., Evans, P.R., Haser, R., Rice, D.W., Hardy, G.W., Merrett, M., and Phillips, A.W. (1979) Sequence, structure and activity of phosphoglycerate kinase: a possible hinge-bending enzyme. *Nature*, **279**, 773-777.

Derewenda, Z. (1985) Some experiences with haemoglobin refinement in *Molecular replacement, proceedings of the Daresbury study weekend*, P.A. Machin ed., CCP4, Daresbury, U.K.

Derewenda, Z. (1990) Notes on the errors of phase determination in the multiple isomorphous replacement method and the molecular replacement method in *Accuracy and Reliability of Macromolecular crystal structures, proceedings of the CCP4 study weekend*, K. Henrick, D.S. Moss and I.J. Tickle eds. CCP4, Daresbury. U.K.

Evans, P.R., Farrants, G.W., Lawrence, M.C. and Shirakihare, Y. (1985) Low resolution structures of two forms of phosphofructokinase in *Molecular replacement, proceedings of the Daresbury study weekend*, P.A. Machin ed., CCP4, Daresbury. U.K.

Fitzgerald, P.M. (1988) MERLOT, an integrated package for the determination of protein structures by molecular replacement. *J. Appl. Cryst.*, **21**, 274-278.

Leslie, A.G.W. (1985) CORELS - How rigid is your molecule ? in *Molecular replacement, proceedings of the Daresbury study weekend*, P.A. Machin ed., CCP4, Daresbury, U.K.

Sussman, J. L., Holbrook, S.R., Church, G.M. and Sung-Hou, K. (1977) A structure-factor least-squares refinement procedure for macromolecular structures using constrained *and* restrained parameters *Acta Cryst.*, **A33**, 800-804.

Swift, H, J., Brady, L., Derewenda, Z., Dodson, E.J., Dodson, G.G., Turkenburg, J., and Wilkinson, A.J. (1991) Structure and molecular model replacement of *Aspergillus oryzae* (TAKA) α-amylase: an application of the silulated annealing method. *Acta Cryst.*, **B47**, 535-544.

Tong, L. and Rossmann, M.G. (1990) The locked rotation function *Acta Cryst.*, **A46**, 783-792.

Watson, H.C., Walker, N.P.C., Shaw, P.J., Bryant, T.N., Wendell, P.L., Fothergill, L.A., Perkins, R.E., Conroy, S.C., Dobson, M.J., Tuite, M.F., Kingsman, A.J., and Kingsman, S.M., (1982) Sequence and structure of yeast phosphoglycerate kinase. *EMBO. J.*, **1**, 1635-1640.