technical memorandum Daresbury Laboratory

DL/CSE/TM23

PROTEIN CRYSTALLOGRAPHY FILM MEASUREMENTS AND PROCESSING
(SEPTEMBER 1982 - FEBRUARY 1983)

BY

T.W. GREENHOUGH and P. MACHIN, Daresbury Laboratory.

MAY 1983

Science & Engineering Research Council
Daresbury Laboratory

CONTENTS

1. INTRODUCTION

This report describes the work of the Applications Group (CSE) in support of SRS protein crystallography data analysis.

A large amount of data processing is required for each protein data set collected on the SRS and while many users rely on their own in-house university systems, users such as Keele, Liverpool and Edinburgh are entirely dependent on DL for both densitometry and subsequent processing.

Densitometry is carried out by the Microdensitometer Service for which this group is responsible. Details of this service are given in Section 2 and a summary of the films scanned is presented.

Subsequent film processing has been carried out interactively on the SRS VAX 11/750 computer. This system was set up to replace an off-line system on the NAS which was unsatisfactory because of the slow turnaround enforced by heavy usage of tape drives and plotter, and the inability of the NAS to support interactive work, which greatly increases the speed, efficiency and accuracy of the data processing. Section 3 of this paper describes the VAX processing software and its implementation and summarises the processing accomplished to date.

Section 4 details program enhancements for synchrotron data. This is a new area which we are the first to investigate.

Future developments which may affect this protein film processing are described in Section 6.

This report covers the period September 1982 to February 1983 and the majority of the films were produced during SRS cycles 8 and 9 (September 1982 - December 1982).

2. MICRODENSITOMETER FILM SCANNING

2.1 Description of the Service

A general description of the SERC Microdensitometer Service including details of the hardware and associated applications can be found in the DL Publications

The SERC Microdensitometer Service (1981) (Elder, Machin, Browett)
Scandig-3. A Technical Memorandum. Sue Kennerley (in press).

The particular aspects which relate to protein crystallography film scanning are detailed here.

Films are scanned on a Joyce-Loebl Scandig-3 rotating drum scanner, which measures optical density by transmittance and results in optical density values (usually restricted to the 0-2D range) represented as numbers from 0 to 255. Films are attached to the transparent acrylic drum by masking tape. Digitization may be done in steps of 25, 50, 100 or 200 micron, the 50 micron option being used for nearly all of the protein work.

A typical scan of a film (12 x 12 cm^2) results in 2400 x 2400 digitized data values (one byte integers) and takes approximately 7 minutes to complete. Scans took twice as long as this before December 1982 when a machine modification successfully improved performance by a factor of two.

Data is output (via a computer) to magnetic tape at 800 bpi. Typically data from 3 films fit onto one magnetic tape at this density and subsequently the mainframe AS/7000 machine may be used to pack more film data onto tapes of higher density (1600 bpi).

Programs have been written for the computer (Data General Nova 3) controlling the scanner to carry out routine tests on the scanner hardware each morning and to diagnose possible faults.

2.2 Summary of Protein Data Scanned

This scanning has been carried out by an operator under the supervision of T Greenhough. The SRS cycles 8 and 9 produced a particularly large number of datasets for DL in-house scanning and it should be noted that most of this work was done before the modification to speed up the scanner had occurred [i.e. 60 mins/film pack].

In addition to the scanning, the operators routinely copied 800 bpi tapes to 1600 bpi tapes (or 6250 bpi for safekeeping) on the NAS AS/7000 in order to check read the data and to store it more efficiently.

It should be noted that film digitisation is a tediously repetitive task which nonetheless has to be performed with great care, and that the organization of tapes is also tedious but requires care to avoid confusion.

TABLE 1

Oscillation Film Scanning (SRS Cycles 8 and 9)

Structure	User	Films	Scandig Hours
Pea Lectin Native	Papiz/Helliwell (Keele)	324	108*
" " Uranyl	Suddath (UAB)	130	43*
" " PCMBS	Greenhough (DL)	126	21
" " Mn	Ellis/Helliwell (Keele)	106	17
PNP-I	Rule/Helliwell (Keele)	150	50*
C4 Uranyl	Harding (Liverpool)	55	10
C4 Native	Sawyer (Edinburgh)	70	12
Apoferritin	White (Sheffield)	30	10*
Ferritin 2	White (Sheffield)	30	5
6-PGDH	Carr/Helliwell (Keele)	100	15
PNPDT	Greenhough (DL)	170	30
MnSOd	Parker (Oxford)	70	12
MnSOd (2)	Parker (Oxford)	75	13
PNPCC	Greenhough/Helliwell (DL)	150	to be scanned
PNPIG	Rule/Helliwell (Keele)	150	" " "
PNPHG	(Keele)	150	" " "

*Scanned before modification to speed up scanner.

3. FILM PROCESSING

3.1 Processing Software

Early in 1982 a policy meeting was held to determine how best to handle protein film data processing at Daresbury and it was agreed with Dr John Helliwell that the MOSCO software developed by Dr Alan Wonacott (Imperial College) should be implemented as an interactive program on the SRS VAX 11/750 computer.

This suite of programs allows the evaluation and analysis of integrated intensities on small angle oscillation photographs; it is recognised as state-of-the-art software; is well tested and is used heavily at Imperial College by the London protein groups.

The programs available in the MOSCO suite are:

i) STILLS - Get position of chosen spots on still photographs and prepare IDXREF file.

ii) IDXREF - Refine orientation and unit cell parameters.

iii) GENVEE - Generate all recorded reflections for a given orientation and rotation range.

iv) MOSFLM - Measure the sets of oscillation films.

v) POSTCH - Check agreement between calculated and observed degree of partiality.

vi) ROTCOR - Correct measured intensities for oblique incidence and absorption.

vii) PASCAL - Process the data from film packs, calculate inter-film scale factors, apply corrections and output h,k,l,m,I,S,D.

3.2 Implementation on the VAX 11/750 at Daresbury

The VAX was chosen as an ideal computer for running interactive work. It has sufficient disk space and magnetic tape facilities for the problem and is available to SRS users.

The film processing software was implemented on the VAX in July/August 1982 (by PM), requiring approximately one month of effort. The original FORTRAN code was very non-standard and had to be modified additionally to read pre-digitized data from disk, to run from a single terminal and to use new graphics routines. The resulting programs were tested using some film data which had previously been processed at Imperial College.

The programs were enhanced (December 1982 TG) to allow for effects relating to synchrotron data as described in Section 4.

3.3 VAX Processing Requirements

The system has been running routinely on the VAX since September 1982 (four months) and approximately 10 datasets have been processed by a variety of SRS users (see 3.4).

Input of digitised data is via the single 800/1600 bpi magnetic tape drive. Command files have been set up to allow users to submit a background job to stage these data on to a system disk (thus not competing with I/O to the user's disk), and then to move it to the user's own file area when required. Typically each magnetic tape contains data from three films (ie one film pack), totalling about 18 Mbytes. This method of operation allows users to input data for one film pack while processing data from the previous pack without undue interaction or any reduction in processing throughput.

The current PX disk allocation (of 150K blocks = 75 Mbytes) is sufficient to store the various program codes (10K blocks) and about 8 digitised data files (100K blocks) in addition to general data files (40K blocks) which accumulate during the processing for a given user. [The minimum disk space needed would be to hold 6 digital data files.]

A graphics terminal with adjustable cursor (Tektronix or VT100 with retrographics) is required for most of this work.

The time taken to process a given film pack depends on several factors, in particular the number of spots on the film. A 3 film pack with 2700 spots per film may be processed on average in 30 minutes real time.

One of the advantages of this processing package is that it allows the user to interact with the processing, for example to display and adjust the relative differences between predicted and observed spot positions on the film.

Such adjustments will necessarily lengthen the time taken at the terminal but they significantly improve ones chances of obtaining a set of accurately measured data.

3.4 Summary of Datasets Processed

Small protein crystallography groups such as Keele, which do not have their own film scanners, rely upon the DL service for film processing, and it is these groups which the service has aimed to help first.

Groups such as Sheffield, process data at their own university, but use the DL microdensitometer service for some film scanning.

Interest in the DL scanning and processing service has been shown recently by the larger groups such as Oxford, which have film scanners of their own, but which cannot necessarily process data as fast as they would wish.

We therefore hope and believe that the service here may benefit all of the protein crystallography community.

Both TG and PM have been involved with the processing carried out here and TG's expertise in this field has contributed significantly to the success of the work so far.

The datasets which have been processed are detailed in Table 2.

3.5 Experiences

We have gained useful experience in film processing techniques. Our involvement with this data processing has led us to observe that some problems arise as a result of decisions made at the time of data collection. The following three points illustrate difficulties we have had to deal with.

i) High resolution data collected on flat film (rather than Vee) cassettes may be difficult to process because of the large expansion and therefore diffuse nature of some of the reflection data.

ii) Choice of exposure times. Over exposure may lead to a sharp increase in thermal diffuse scattering as well as increased radiation damage to the crystals and film saturation producing spots which can't be measured with any accuracy.

TABLE 2

Summary of VAX Processing of Oscillation Films (September 82 - February 83)

User	University	Data	Film Packs (no/pack)	R Sym
P Carr	Keele	6PGDH Native	30 (2)	30.0
M Papiz	DL/Keele	Native Pea-Lectin 1.8	90 (3)	8.5
S Rule	Keele	PNP-I	45 (3)	8.0
F Suddath	UAB/Keele	Pea-Lectin Uranyl	44 (3)	?
P Carr	Keele	Recollected data as above 6PGDH	20 (3)	11.5
T Greenhough	DL	PNP-I reprocessed with Synchrotron Geometry program modifications	10 (3)	not produced
M Harding	Liverpool	C4 Uranyl	18 (3)	being processed
G Ellis	Keele	Mn Pea Lectin 2.3A	53 (2)	5.2
M Parker	Oxford	MnSOd (V. Cassettes)	34 (2)	6.9
L Sawyer	Edinburgh	C4 Native	20 (3)	to be processed

iii) Care taken in film handling and film developing is reflected in the accuracy of the resulting data. Attempts are made in processing to allow for film distortion but clearly it is better to avoid such problems in the first place.

TG has recently (February 1983) become Station Master of port 7.2 (protein crystallography) and we are therefore in a good position to feed back information gained from processing to the experimenters.

3.6 Implementation of STILLS/IDXREF on a PERQ Computer

The program IDXREF is used to refine the misorientation angles and cell parameters of the crystal in order to get good values on which to base the prediction of spot positions for each film. This program requires input data which includes the actual spot positions of a series of spots on Still photographs - as provided by the program STILLS.

The original version of STILLS as implemented on the VAX, displays a digitised image of the film on a Tektronix and relies upon the crystallographer to cursor in selected spots seen on the screen.

We have found that an alternative method of spot position input is to place the actual film on the digital tablet of the PERQ computer, and to use the puck device to enter spot positions. This has proved to be a much quicker method which gives as good results.

A program on the PERQ was written to achieve this input and the existing IDXREF code was transferred to the PERQ and implemented (PM). Since the result of this process is to produce 4-6 refined parameters there is no problem in transferring the answers to subsequent processing programs on another computer.

In summary it would appear that this processing on the PERQ is easier and as successful as the parallel process on the VAX.

4. PROGRAM ENHANCEMENTS FOR SYNCHROTRON DATA

4.1 Description and Formulae

The inclusion of synchrotron geometry (Greenhough and Helliwell, J. Appl. Cryst. (1982) 15, 493-508) into the prediction and refinement routines in MOSCO has led to major improvements in the processing of SRS film data. In some cases post-refinement via POSTCHK proved to be impossible or meaningless without the correct SR geometry, while in all cases the predictions are greatly improved. Furthermore, in the majority of cases a refinement of mosaic spread has given a narrowly defined value leading to much better prediction and post-refinement.

The synchrotron geometry is included in data processing by replacing the conventional reciprocal lattice volume radius:

$$eps_c = \frac{d^*\cos\theta}{2} \left[\eta + \gamma + \left(\frac{\delta\lambda}{\lambda}\right) \tan\theta \right]$$

where η is mosaic spread and λ beam crossfire (Greenhough and Helliwell, J. Appl. Cryst. (1982) 15, 338-351) by:

$$eps = \frac{1}{2}\left[(\delta d^{*2} + z\gamma_H)^2 + \gamma_V^2 y^2\right]^{\frac{1}{2}} + \frac{d^*\cos\theta}{2}\left[\eta + \left(\frac{\delta\lambda}{\lambda}\right)\tan\theta\right]$$

where δ is the correlated component of spectral dispersion, γ_H and γ_V are the horizontal and vertical cross fire angles, $\delta\lambda/\lambda$ is the conventional type spectral dispersion and y and z are the crystal axis coordinates perpendicular to the beam with z the rotation axis.

Stills refinement where all partialities are assumed to be ½ is of course unaffected by the size or shape of the rocking curve. Post refinement however depends on both functions; using only reflections nearly ½ recorded reduces the dependence on the shape of the rocking curve but increases the uncertainty about the width. Since the majority of SRS data is collected near the Guinier position for 1.488 Å the most important factor to be considered is the highly asymmetric beam cross fire. A small component of the correlated spectral dispersion term is however present; in the case of optimised anomalous dispersion experiments a large contribution often needs to be accounted for.

The usual SRS set up yields:

γ_H = 0.201°
γ_V = 0.013°
$(\delta\lambda/\lambda)_{conv}$ = 0.00151
δ = 0.00043

but users should check for changes which occur from time to time.

4.2 An Example Application

The Manganese superoxide dismutase data (Colin Blake, Mike Parker; Oxford) provides a good example of the improvements gained with the use of synchrotron geometry during data processing. Collected at the Mn edge and very near to the Guinier position, these data are currently being processed on the VAX at Daresbury. The crystals seem to suffer quickly from radiation damage in the early exposures, and this is reflected in a steep rise in the value of the mosaic spread parameter required for good prediction.

Using a conventional geometry description of mosaic spread and beam cross fire and cycling through prediction, integration and post refinement for the first two packs leads to a 'refined' value of 0.27° for $(\eta + \gamma)$. The refinement is carried out by varying the value of $(\eta + \gamma)$ in repeated runs of IDXREF and plotting this value against the minimized quantity

$$\sum_I ((R_o - R_c)/d^*)^2$$

Convergence is considered to have been achieved when two successive runs of prediction, integration and post refinement yield the same misorientation angles and the same minimum residual in IDXREF for the same value of $(\eta + \gamma)$.

The process was repeated for the same two packs using the correct SR geometry for this case (asymmetric cross fire γ_H = 0.15°, γ_V = 0.013°) and a refined value of the mosaic spread η of 0.185° was obtained, as shown in the figure.

TABLE 3

			Conventional Geometry	SR Geometry
$\eta + \gamma$ (refined)			0.27°	-
η (refined)			-	0.185°
γ_H			-	0.15°
γ_V			-	0.013°
$(\delta\lambda/\lambda)_{conv}$			2.5×10^{-3}	0.77×10^{-3}
Prediction:	Pack 1	Full	492	671
		Partial	989	632
	Pack 2	Full	492	672
		Partial	984	613
Post refinement:	NREF		72	56
	In range		46	41
	Out of range		26	15
	$\sum_i ((R_o - R_c)/d^*)^2$		5.2×10^{-7}	4.0×10^{-7}
	r.m.s. residual(°)		0.008	0.007
	ϕ_x		0.286°	0.286°
	ϕ_y		-0.672°	-0.672°
	ϕ_z		0.503°	0.514°

MnSOD data processing. Vee Cassette data to ~ 2.2 Å resolution at λ = 1.89 Å

A comparison of the two results is given in the table which shows that there are large differences between the two methods in crucial areas. While the final values of the misorientation angle ϕ_z differ by 0.01°, the prediction of fulls and partials differs dramatically. It was clearly seen that the conventional prediction gave far too many partials towards the vertical axis and too few as the cusp was approached. The $(\eta + \gamma)$ value of 0.27° is a compromise between the $(\eta + 0.013)$ needed in the vertical and the $(\eta + 0.15)$ needed towards the horizontal. This is a simplified view since the situation is somewhat complicated by spectral dispersion, but the value of η obtained from the synchrotron geometry refinement gives a good prediction over the whole film. The somewhat sudden appearance of split spots on the data films in later packs is reflected by a sharp rise in the refined value of η to around 0.35°.

Other datasets which particularly benefit from SR geometry are bacterial 6-PGDH (Carr/Helliwell; Keele) which does not respond to post refinement in a sensible way with the conventional geometry, and native pea lectin (Papiz/Suddath; Keele/UAB) which gives a better post refinement with SR geometry. Obviously all datasets will benefit from much improved prediction.

4.3 Future Improvements

Future improvements to the data processing facilities at Daresbury include the calculation of individual spot shapes and sizes for integration as outlined in Greenhough, Helliwell and Rule (J. Appl. Cryst., in press). This is particularly important for high resolution data collected on flat films, due to the spot inclinations introduced by the asymmetric SR cross fire.

5. SUBSEQUENT PROCESSING

The film processing described in Section 3 results in indexed intensity data (~ 100,000 measurements) stored on magnetic tape.

Subsequent processing may then be performed on any computer. The CCP4 (protein crystallography collaborative computing project) suite of programs on the NAS AS/7000 at Daresbury include specific film processing programs as follows:

PASCLCF - Input reflection data from magnetic tape, sort, and prepare a standard disk data file.

ROTAVATA - Calculate scale factors and/or temperature factors between overlapping batches of data by the method of Fox and Holmes.

AGROVATA - Apply scale factors and average data.

Further documentation is available on request.

6. FUTURE DEVELOPMENTS

6.1 Data General Nova Computer Enhancements

In response to Mike Elder's paper "Enhancements to the Data General Nova 3/12 controlling the Scandig Microdensitometer" the Laboratory agreed to the requested computer enhancements and the new system comprising floating point hardware, increased disk storage and Fortran V software is being installed (February 1983).

These enhancements will allow us to run the MOSCO film processing program on the Data General computer on-line to the microdensitometer. Previous attempts without the Fortran V and floating point unit were prohibitively slow (2 hours for a single film).

In normal operation there will be a choice between film digitization and processing on the VAX, as at present, and on-line processing on the Nova. Circumstances will dictate which method is used for any particular dataset, but total througput will certainly be increased and the added flexibility welcomed.

6.2 The VAX Implementation of these processing programs has not only been of use to those at Daresbury but considerable interest has been indicated from other sites. The system has already been distributed to Berlin, Cambridge and Leeds and requests for copies have been received from Paris, Japan and India.

6.3 The Development of Area Detectors

The TV detector will obviously play a very important role in data collection in the future, but it is difficult to assess when (and if) it will completely replace film.

A lot will depend on how long it takes to install the detector and to implement software, and there must then be a trial period of checking out the system to ensure accurate data are obtained.

Film methods are well tested and are known to be reliable and accurate. Much expertise in film processing has been built up, and data collection on film has the advantage that it maximises use of the SRS beam.

We would estimate that film methods will continue to be important for at least another two years.

7. CONCLUSIONS

7.1 The microdensitometer service can now meet the demands of DL, Keele, Edinburgh and Liverpool scanning requirements and can additionally service other universities.

7.2 The film processing system on the VAX computer is now well tested, efficient and reliable. We hope that the facility will remain available to the protein crystallography community whilst film methods continue to be used at DL.

7.3 With the introduction of the enhancements to the Data General Nova system allowing film processing on-line to the microdensitometer, we envisage increased flexibility and potential throughput of data. We look forward to an increased interaction with more university groups.

7.4 We have already enhanced the processing programs (Section 4) and have plans for further refinements which, if successful, will be reported in the scientific literature. The combination of expertise in data collection and data processing is proving most fruitful.

FIGURE CAPTION

Fig. 1 Mosaic Spread refinement for MnSOD using SR geometry with $\gamma_H = 0.15^o$, $\gamma_V = 0.013^o$, $(\delta\lambda/\lambda)_{conv} = 0.77 \times 10^{-3}$ in POSTCHK and IDXREF