DARESBURY LABORATORY

INFORMATION QUARTERLY for PROTEIN CRYSTALLOGRAPHY

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

Number 18 July 1986

Contents

Editorial 1

The Birth of a Protein Sequence/Structural Database 3

A Robust Method for the Superposition of many Protein Structures and its Application to Comparative Model Building 11

Ferritin Metals 19

Determination of the Wavelength Normalisation Curve in the Laue Method 23

An Improved Program Package for the Measurement of Oscillation Photographs 33

An Unusual Crystal Packing Arrangement in Phosphofructokinase 41

Changes to the MRC Version of Frodo since August 1985 45

CCP4 News and Network Addresses 47

Distribution of the Leeds Structure Prediction Suite 49

Editor:	Pella Machin	Science & Engineering Research Council, Daresbury Laboratory, Daresbury, Warrington WA4 4AD, England
Deputy Editor for Imperial College:	Dr. Alan Wonacott	Imperial College of Science & Technology, The Blackett Laboratory, Prince Consort Road, London SW7 2BZ
Deputy Editor for Birkbeck College:	Dr. David Moss	Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX

EDITORIAL

Our thanks once again go to the contributors to this newsletter and to Ian Tickle who was responsible for collecting the papers. A belated thank you is also due to Sheila Gover who did the hard work of coordinating the previous issue.

A recent development of interest is the birth of the European Association of Crystallography in Molecular Biology (EACMB) which has been formed to strengthen the already good connections between the various research groups active in the field in Europe. It largely follows the success of CCP4 and aims to provide a forum for the rapid exchange and exploitation of ideas and results amongst the crystallographic community, and to publicise the way in which crystallography underpins biology. Wim Hol is Chairman of the EACMB committee and Keith Wilson is Secretary. A EACMB newsletter will be published regularly and its first issue was in June 1986. I am sure CCP4 welcomes this extension into Europe and Gerard Bricogne will in future be attending CCP4 Working Group 2 meetings as the EACMB representative.

Pella Machin
22 July 1986

THE BIRTH OF A PROTEIN SEQUENCE/STRUCTURAL DATABASE
or
Reflections on a disc drive

Alan J Bleasby (Astbury Department of Biophysics, Leeds University)

This report, as you will shortly gather, is not of a purely crystallographic nature but, rather, concerns a project concerned with storing 'information' about proteins from several disciplines. This work is one of the projects supported by the Protein Engineering Club.

I. INTRODUCTION

The Protein Engineering Club was conceived and formed in 1984/85 as a collaborative scheme between Universities and industry. Essentially its purpose was to facilitate research into protein structure/function relationships directed at topics which bear directly on the ability to engineer proteins. Such funded research has included study of crystallisation techniques, investigation of target proteins and the setting up of a protein database; this communication is directed at the latter.

Two groups are involved

i) The LEEDS group is involved in setting up a 'sequence-related' database. Members involved at present include Tony North, John Wootton, John Findlay, Nigel Dix and myself.

ii) The BIRKBECK COLLEGE group is concerned with the production of a 'structure-related' database. Members involved include Tom Blundell, Mike Sternberg, Janet Thornton, Suhail Islam, Fiona Hayes and Francis Whitley.

II. WHY SET UP TWO MORE DATABASES?

There are plenty of databases around at present. On the protein side there are the PIR, Claverie and Doolittle databases: on the nucleic acid side there are the Brookhaven, EMBL and Genbank databases to name but a few. Why do we need more?

The databases mentioned above are collections of sequences (some verified, some not) with, mostly, only protein/DNA name, a specific database code which varies according to the database, source, function, literature references and, sometimes, freely supplied software to access the information. Such information is of course invaluable to a researcher wanting to match his sequence to others in the database, for example homology searches, alignments and the like. However, interrogation of such databases is necessarily confined to the level of primary or secondary structure. Further, any software supplied is of a restricted nature, 'one program one function' and there is little flexibility in the means of accessing the data. Generally you can look at one 'class' of data at a time. In a nutshell it is tedious to 'relate' one type of information with another.

The Leeds/Birkbeck databases are being designed to hold information about proteins both on the secondary structure aspect but also extensive information on the tertiary structure of the proteins. The databases will hold information at each level of protein structure (protein, domain, motif, residue, atom etc.).
Moreover, the data will be held in a 'relational' form. This requires the use of the so-called RELATIONAL DATABASE.

III. WHAT IS A RELATIONAL DATABASE?

All information in a relational database is stored in TABLES, each table having a distinct name e.g. PROTEINS, SSBONDS, MODIFICATIONS,

DOMAINS, FAMILIES, etc.. Each of the tables can have any number of COLUMNS. Into these columns are inserted ROWS of information, at least one row per protein.

The power of a relational database is that any value(s) in any column(s) of any row can be selected according to table name, column name and a 'key'. This key may be a single criterion such as the database code of the protein (e.g. get the protein names from table PROTEINS if the molecular weight (in column MOLWT) is greater than 300K), or may be composed of multiple criteria (e.g. selecting a motif sequence on the basis of protein family classification and type of motif/domain). Further, you are not restricted to selecting values from one table at a time, one command will suffice providing the two (or more) tables are linked by common columns. Other benefits of a relational database include the ability to select rows on the basis of matching character strings and to perform selection by arithmetic criteria.

These selections are performed using the relational database's QUERY LANGUAGE. These languages provide an 'English language-type' interface between the database tables and the user. This enables computer-illiterate users to access the database without developing too many neuroses.

IV. THE CHOICE OF COMPUTER AND RELATIONAL DATABASE

There is a bewildering variety of relational databases on the market. These include IBM's DB2 (MBS) and SQL (VM & DOS), CINCOM's SUPRA, CULLINET IDMS/R, ADR DATACOM DB and the DEC database. These are just some of the ones we haven't got. Birkbeck were delegated the responsibility of selecting the database software. They decided on a system called ORACLE (produced by the ORACLE CORPORATION U.S.A.). Amongst the apparant advantages of ORACLE are that it is not confined to the manufacturers hardware and it is indeed implemented on a wide

range of processors; it is also more economical in terms of space and speed than some rivals Of course, to maintain compatibility between both databases Leeds had to have ORACLE also. In my view this is a pity as, having experimented extensively with ORACLE I find it does not possess some valuable facilities one would expect of an expensive commercial database and also has a major drawback with regard to protein sequences, but more of that later.

Some of the good features of ORACLE are that the query language SEQUEL is one of the most 'human' available, it can handle very large databases, it can be linked to a host high level language and null column entries are supposed to take up no disc space.
Both Leeds and Birkbeck have ORACLE running on a VAX 11/750 cpu with >450Mb disc storage.

V. PROBLEMS WITH ORACLE

I have already mentioned that ORACLE can be linked to a high level computer language. Those supported are FORTRAN, C and COBOL; this came as an initial surprise as the two most database-like languages are LISP and PROLOG; perhaps these were considered too dynamic. FORTRAN and C have very limited character string handling ability whereas COBOL is restrictively business oriented. We have a FORTRAN compiler for the project. An associated problem is that although FORTRAN can talk to ORACLE the converse is not true - you cannot invoke a FORTRAN program from the query language.

ORACLE also lacks the ability to use dynamically derived column names and has limited flexibility in its use of wild-cards in character comparisons. Some supposedly supported features are bug-ridden - you've only got to think of using them and ORACLE dies a horrible death and does not allow you to shut it down in the normal way; it labours under the delusion that two users are still logged on and produces error messages which, when referenced in the manuals, say 'It is absolutely

impossible to get this error... contact ORACLE CORP. immediately'. All rather disconcerting. Looking on the bright side, all the above 'features' can be worked around. They do however cause a, hopefully negligible, loss of efficiency. However, there is one feature that cannot be circumvented and that involves the ORACLE datatypes.

ORACLE possesses four datatypes, these being INTEGER, DATE, CHARACTER and LONG. It is the CHARACTER and LONG datatypes that create the problem. Type CHARACTER can hold character strings of up to length 240, LONG can hold around 64000. The limitation is that you can only perform comparisons on type CHARACTER, type LONG supposedly being just for large amounts of text (e.g. references). Obviously putative users will want to perform comparisons on amino acid sequences and this could only be done on type CHARACTER. As the average protein has 300 residues you can see the dilemma, either you have the sequence spread over several columns (it would need 9 for Factor VIII!) or you forget about using ORACLE altogether. Owing to the difficulty of selecting part of a sequence from ORACLE (you could never know in what column the fragment of interest was stored) Leeds have decided to abandon ORACLE in any application involved with whole protein sequences. Fortunately, 240 characters is still enough to hold sequences of domains and motifs and therefore ORACLE will be used for this purpose. Birkbeck do not share our problem as it will be evident that a protein structure database will hold the sequences down columns rather than along rows.

It is now obvious why the host language interface (FORTRAN) is essential for the Leeds group. Each application involving whole protein sequences will need a purpose written FORTRAN program, and as far as the query element goes I've already shown this is a one-way affair. It is a case of 'one program one job'. These separate programs will all need to be gathered together in a menu driven master program.
Oh Birkbeck what have you done to me!

VI. THE PROPOSED STRUCTURE OF THE LEEDS DATABASE

A full description of the Leeds database is beyond the scope of this communication. In fact both databases are still at the developmental stage and it would be precipitous to give exact details at this time. However, here is a very brief summary of the type of data the Leeds database will hold.

Table PROTEINS: Leeds/Brookhaven/PIR/Claverie/Doolittle codes, title, species, function, classification, general details of X-ray/ coordinates/ligands/oligomers/precursors/families/domains/motifs etc. (referencing other tables).

Table SEQUENCES: DNA translation, open reading frames, how sequenced (if not DNA), inconsistencies, initiator MET, amino acid composition, percentage amino acid composition, overlaps etc.

Table PREPOLY: precursors/derivatives, residue numbers etc.

Table OLIGS: Oligomers/complexes/assemblies, relations etc.

Table MODS: Columns for exact types of modifications known,functional effects, residue number etc.

Table SSBONDS: residue number,crystallographic/chemical evidence etc.

Table LIGS: number and type of ligands etc.

Table STRUCTURES: how refined/resolution/thermal factors/solvent ions/space groups/molecules per unit cell/missing residues or atoms etc.

Table SECSTRUC: number of helices/sheets/turns etc. deduced from a range of predictive algorithms.

Table EVIDENCE: evidence for structure.

Table DOMAINS: domain information/sequence

Table MOTIFS: motif information/sequence

VMS SEQUENCE FILES: application programs for homology searches, alignments, HYDRA/FRODO linking etc.

I will concentrate on the database structure in any further correspondence but I hope the above gives you a flavour of things to come. Meanwhile AURICLE will always hold a special place in my heart.

A Robust Method for the Superposition of Many Protein Structures and its Application to Comparative Model Building

I. Haneef and M.J. Sutcliffe, Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.

Key Words: comparative model building
iteratively re-weighted least squares
multi-dimensional scaling
robust fit
unique solution

Abstract

The problem of simultaneously superpositioning more than two protein molecules is reviewed. A method which achieves this by firstly obtaining an "average structure" or "framework" and then least squares fitting all the molecules onto this "framework" is proposed. Investigation of the convergence properties of this algorithm in the case of both weighted and unweighted least squares leads to the conclusion that both give a unique answer and both are robust for an homologous family of proteins. Applications of the technique to comparative model building are discussed in conjunction with multi-dimensional scaling.

Introduction

There are many occasions when similar protein structures need to be superpositioned to enable their comparison. For example, in modelling a protein on the basis of its sequence and the known three-dimensional structures of homologous proteins, or in the comparison of conformers at different time points in the molecular dynamics simulation of a protein.

The simplest approach is to use one molecule and superpose all the others onto it using one of the standard pairwise least squares fitting algorithms (eg. Ferro and Hermanns, 1977; Kabsch, 1978; MacLachlan, 1979, 1982). These algorithms firstly superpose the centres of mass of the respective molecules and then minimise:

$$E = \Sigma\ w_i \mid \underline{X}_i - \underline{\underline{R}}\ \underline{Y}_i \mid^2 \qquad (1)$$

where E is known as the residual, the summation is over equivalenced pairs, w_i is the weight corresponding to the i^{th} pair of equivalenced atoms (usually set to unity), $\underline{X}$ and $\underline{Y}$ are the respective sets of coordinates, and R is the rotation for mapping $\underline{Y}$ onto $\underline{X}$. However, since in this approach one molecule is used in all the fits and the rest of the molecules only once, it is clear that this approach will bias the superposition towards the structure of the molecule used in all the fits.

It follows that a method whereby the superposition is performed without bias to any one of the structures is required. The method discussed herein involves the determination of the average position of each of the equivalenced atoms, the resulting set of positions being termed the "framework". It should be stressed that the "framework" does not have to comply with recognised protein geometries as it is simply a means to an end, rather than a structure in its own right.

Seven globins are cited as an example of an homologous family (see table 1). Twenty eight equivalences per molecule were selected from four residues in the centre of each of the A, B, C, E, F, G and H helices in the alignment according to Lesk and Chothia (1980). In this work, only alpha carbons were used in the superposition of structures.

Unweighted Derivation of the "Framework"

The simplest means of determining the "framework" is to perform an iterative, unweighted least squares procedure, as shown in algorithm 1. This algorithm produces a unique solution for the "framework" provided that the cut off value in step 3 is less than the lowest molecule - molecule root mean square error (RMSE).

The uniqueness of the "framework" produced by the algorithm was investigated by fitting the homologous proteins to a random structure, rather than to one of the structures, in step 1. Repeating this for a number of different random structures, the RMSE between all pairs of "frameworks" was between 10^{-6} and $3*10^{-5}$ A. Performing the pairwise fitting of all the molecules to each of these "frameworks" in turn resulted in a maximum difference of two digits in the fifth decimal place when comparing the RMSE's of a particular molecule to all the "frameworks". The negligible size of these deviations leads to the conclusion that the algorithm produces a unique "framework".

Weighted Derivation of the "Framework"

Unweighted least squares, by its very nature, weights all atoms equally. This is of no consequence if equivalenced atoms are in spatially similar positions, but if one of them differs in position from the others then this will draw the corresponding point on the "framework" towards itself by virtue of its distance from the other atoms. In such a case it would seem sensible to discount such an outlier, or at least make its contribution to the "framework" less, since the family fingerprint for this particular set of equivalenced atoms can be seen to lie within the larger cluster. In turn, this implies that weighted least squares, rather than unweighted least squares, should be used both in the determination of the "framework" and the subsequent fit of all the molecules onto this "framework".

The function now being minimised is:

$$E = \sum_{ni} \sum w_{in} \left| \underline{F_i} - \underline{\underline{R_n}}\ \underline{Y_{in}} \right|^2 \qquad (2)$$

where the summation over n is over all the molecules and that over i is over equivalenced pairs, F_i are the coordinates of the "framework". Now all the weights have to be determined iteratively, as well as the "framework".

The approach adopted for minimising equation (2) treats the weights and the "framework" separately (see algorithm 2). Firstly the "framework" corresponding to a given set of weights is determined (steps 1 - 5). Having determined this "framework", all the molecules are fitted onto the "framework" and the weights corresponding to the new orientations calculated (steps 6 - 9). This process is repeated until the residual in equation (2) is deemed to have been minimised (step 8).

The following philosophy was used to define the weighting function adopted, although it should be stressed that it is by no means a definitive weighting function. Indeed, each researcher will probably develop one to suit their own criteria. A function is required which:

(i) weights points inversely to their distance from the "framework".
(ii) tends to zero as the distance of a point from the "framework" becomes large.
(iii) is continuous as the distance of a point from the "framework" tends to zero.
(iv) reflects the error in the coordinates.

The error in crystallographically determined coordinates between independently determined solutions of the same structure is 0.3 - 0.5 A (Chothia and Lesk, 1986), the square of which is about 0.1 A^2. After considering these criteria, the form shown in step 6 of algorithm 2 was chosen.

The convergence properties of algorithm 2 are shown in figure 1, and the position of the "framework" with respect to the seven globins is shown in figure 2.

Algorithm 2 was extensively tested in order to see under what conditions, if any, it would break down. It worked successfully when a random structure was used in step 1, and also when a random structure was used in step 1 and thirty random points were introduced randomly into the seven molecules, implying that the algorithm is robust. If in the latter case random weights were introduced at step 1, the resulting "framework" was not unique. Repeating this for fourteen structures - the seven containing random points and the seven crystallographic structures - the resulting "framework" was again not unique. In this case, "framework" - "framework" RMSE's are less than 0.3 A and analysis of the results shows that the resultant "frameworks" are not significantly different from the corresponding unique "framework". This implies that, with a little modification, the algorithm can produce a unique solution, even under these extreme conditions.

The break down of the algorithm implies that the resultant "framework" is a function of the initial set of weights. If, however, one of the globin molecules is used as the first approximation to the "framework" and the seven structures containing random points assigned random weights in step 1, the resultant "framework" is unique. Thus it follows that, for an homologous family, the algorithm produces a unique solution and is robust.

Often, it may be necessary to sub group structures; this can be achieved using multi-dimensional scaling. Firstly, each of the N structures are fitted pairwise to each of the others. The RMSE's for the respective fits are read into a traceless, symmetrical N*N matrix and the covariance matrix corresponding to this matrix determined. The N eigen vectors and eigen values of the covariance matrix are determined. Plotting [(highest eigen value)$^{0.5}$ * (corresponding component of eigen vector)] versus [(second highest eigen value)$^{0.5}$ * (corresponding eigen vector)] for each molecule, a plot like that shown in figure 2 is obtained. The square root of the eigen value is taken as the eigen value is a measure of the variance. For a discussion of this particular type of multi-dimensional scaling, see for example Crippen and Havel (1978) or Jennings (1978).

Use of interatomic distances to determine a structure (Crippen and Havel, 1978) leads to another possible algorithm for determining the "framework". If the average distance corresponding to equivalenced atoms is determined, these entered into a distance matrix, the corresponding covariance matrix determined from which the thre highest eigen values and corresponding eigen vectors are determined, then the "framework" can be determined from:

$$F_{ij} = \lambda_j^{0.5} w_{ij} \qquad (3)$$

where j=1,2,3 and i=1,2,...,(number of points on "framework"), F_{ij} are the points on the "framework", λ_j are the eigen values and w_{ij} are the components of the three eigen vectors corresponding to the three highest eigen values.

Conclusion

The convergence properties of both algorithms for the determination of the "framework" imply that they produce a unique "framework" for homologous families, although that which uses weighted least squares is preferred as it picks out the structural "finger print" of the family corresponding to each equivalenced atom. Both this algorithm, and the use of multi-dimensional scaling to determine sub families, are likely to prove to be very useful in the field of comparative model building.

Acknowledgements

The authors would like to thank, amongst others, Professor Tom Blundell and Dr. Stephen Bryant for valuable discussions during the course of this work.

References

Chothia, C. and Lesk, A.M. (1986) Embo J. 5, 823-826.
Crippen, G.M. and Havel, T.F. (1978) Acta Cryst. A34, 282-284.
Ferro, D.R. and Hermanns, J. (1977) Acta Cryst. A33, 345-347.
Jennings, A. (1978) "Matrix Computation for Engineers and Scientists", Wiley, Chichester.
Kabsch, W. (1978) Acta Cryst. A34, 827-828.
Lesk, A.M. and Chothia, C. (1980) J. Mol. 136, 225-270.
MacLachlan, A.D. (1979) J. Mol. Biol. 128, 49-79.
MacLachlan, A.D. (1982) Acta Cryst. A38, 871-873.

Table 1

Protein Name	Resolution/A	Brookhaven Code
Erythrocruorin (deoxy)	1.4	P1ECD
Leghemoglobin (acetate, met)	2.0	P1LH1
Myoglobin (deoxy)	1.4	P1MBD
Hemoglobin alpha (deoxy)	1.7	P2HHB
Hemoglobin beta (deoxy)	1.7	P2HHB
Hemoglobin alpha (horse, aquo met)	2.0	P2MHB
Hemoglobin beta (horse, aquo met)	2.0	P2MHB

Algorithm 1

Unweighted Least Squares Superposition

1. Choose one of the structures at random as the first approximation to the "framework" and fit all the others to it pairwise.

2. Determine the new "framework" from:

$$\underline{F_i} = \frac{1}{NMOL} \sum_{n=1}^{NMOL} \underline{Z_{in}}$$

where: F_i are the coordinates of the i^{th} point on the "framework".
NMOL is the numberof molecules.
$\underline{Z_{in}} = \underline{R_n Y_{in}}$ = the fitted coordinates of the i^{th} atom in molecule n.

3. IF (RMSE between consecutive "frameworks") $< 10^{-5}$ A THEN STOP.

4. Fit all the molecules onto the "framework" pairwise.

5. GOTO 2.

Algorithm 2

Weighted Least Squares Superposition

1. Choose one of the structures at random as the first approximation to the "framework" and fit all the others to it pairwise using unit weights.

2. Determine the new "framework" from:

$$\underline{F}_i = \frac{\sum_{n=1}^{NMOL} w_{in}\, \underline{Z}_{in}}{w_{in}}$$

3. Fit all the molecules to the "framework" pairwise.

4. IF (RMSE between consecutive "frameworks") $< 10^{-5}$ A THEN GOTO 6.

5. GOTO 2.

6. Calculate the weights from:

$$w_{in} = \frac{1}{0.1 + d_{in}^2}$$

where d_{in} is the distance between atom i on molecule n and point i on the "framework".

7. Calculate the residual from:

$$E = \sum_{n=1}^{NMOL} \sum_{i=1}^{NATOMS} w_{in} \left| \underline{F}_i - \underline{Z}_{in} \right|^2$$

where NATOMS are the number of atoms in a given molecule to be superposed.

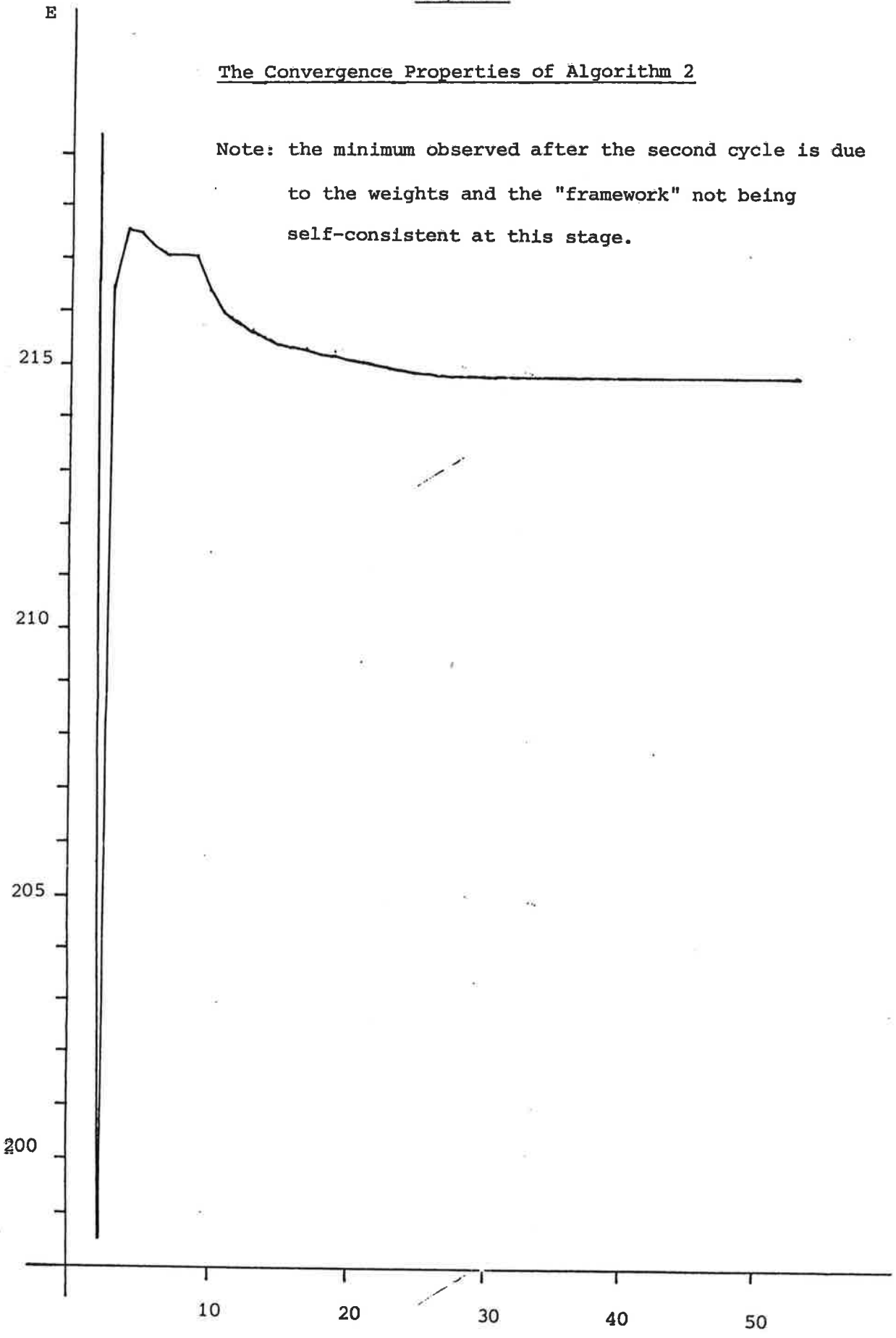
8. IF (difference between E's on consecutive cycles) $< 10^{-5}$ OR 10^{-5} (RMSE between "frameworks" on consecutive cycles) $< 10^{-5}$ A THEN STOP.

9. GOTO 2.

Figure 1

The Convergence Properties of Algorithm 2

Note: the minimum observed after the second cycle is due to the weights and the "framework" not being self-consistent at this stage.



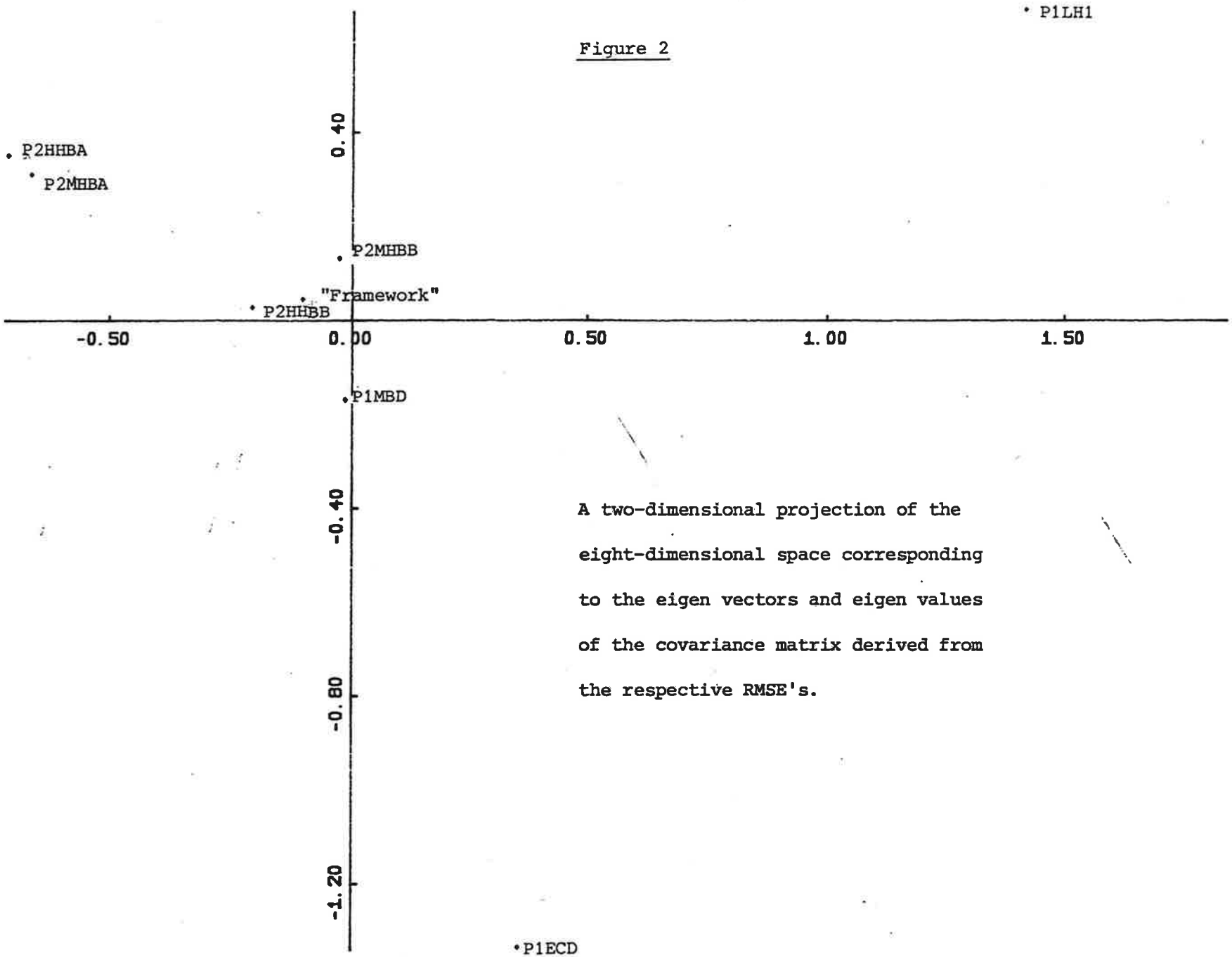


Figure 2

A two-dimensional projection of the eight-dimensional space corresponding to the eigen vectors and eigen values of the covariance matrix derived from the respective RMSE's.

FERRITIN METALS

We have endeavoured to provide a structural explanation for the effects of metal ions on the rate of accumulation of the central iron core within the hollow protein shell of apoferritin. Figure 1. provides an overview of the apoferritin structure. The molecule consists of 24 subunits associated together by 432 symmetry into a shell of about 20 A thickness around a central cavity of 80 A diameter. The 174 amino acid chain folds into a 4-helical bundle, A, B, C and D, with helix E lying at about 45° to the axis of the bundle and a long loop, L, connecting B and C which forms a short stretch of antiparallel β-sheet with its 2-fold related self L'. Passageways through the shell exist at the interfaces around 3- and 4-fold symmetry axes, the 3-fold channel being completely hydrophilic while the 4-fold channel is entirely hydrophobic.

The kinetics of iron uptake have been modelled by a slow nucleation phase, in which a few iron atoms assemble at an initiation site on the inside surface of the shell, followed by a more rapid growth phase in which the oxidation of Fe^{2+} to Fe^{3+} occurs on the surface of the expanding iron core. Metal ions may effect iron uptake by blocking entry, or by inhibiting nucleation or growth.

Since our last contribution to the CCP4 newsletter sites for Zn^{2+} and Tb^{3+} have been located in horse spleen apoferritin. Both are inhibitors of ferritin formation, although they act in somewhat different ways. We find Zn^{2+} occupies two sites in the 3-fold intersubunit channels which pass through the apoferritin shell. The outermost Zn^{2+} has three glutamic acid ligands (one from each of three symmetry related subunits) and three water molecules. The inner Zn^{2+} has three aspartic acid ligands and no water visible at 2.6 A resolution. These acidic residues are conserved in all known ferritin sequences. In the Tb^{3+} derivitive there is only a single Tb^{3+} in these channels bound by all six carboxy groups, necessitating a movement of these side chains as compared to their positions in crystals containing Zn^{2+} or Cd^{2+}.

Other metals (Mn^{2+} and VO^{2+}) have been shown to bind with stoichiometry 1/3 or 2/3 per subunit and may also be located in these channels, which are probably also the channels used by iron as Fe^{2+} or Fe^{3+}. Inside the molecule there are a number of metal sites, Table I. Site 5 which binds Cd^{2+}, Zn^{2+}, Tb^{3+}, or UO^{2+} has only one generally conserved ligand. In an electron density map of the isomorphous rat liver apoferritin calculated with horse spleen apoferritin phases, there is no Cd^{2+} at this site. We think the conserved glutamic acid residues binding Tb^{3+} at site 8 may be the ligands binding

Fe^{3+} and forming the nucleation centre for the formation of ferritin's iron core. This core is a microcrystalline hydrous iron oxide mineral called ferrihydrite, and is the form in which up to 4500 iron atoms may be stored within the apoferritin shell.

A plot of temperature factors vs sequence number is shown in Figure 2. As may be expected, high B values are found for residues in inter-helical regions. The B helix also has rather high B's. Due to interdigitation of loop residues between helices A and A' around the 2-fold axis, B and B' are rather widely separated forming a long groove on the inside surface. Many of the acidic residues projecting into this groove have disordered side chains which may provide some flexibility for the nucleation site. Alpha carbons of residues 1, 7-15, 79-87, 115-121, and 156 are at a radius greater than 60 A from the center of the molecule and thus protrude from the nearly spherical surface. Three of these high B protruding regions have been shown to contain epitopes. Antibodies to peptides from these regions are reactive to native ferritin.

Refinement of the structure is being improved with a new version of the Hendrickson-Konnert program (Univac version from J Priestle, Biozentrum, Basel converted to IBM by GC Ford) which takes account of intersubunit and intermolecular contacts. Ferritin's structure with 24 subunits related by 432 symmetry and intermolecular cadmium bridges makes ignoring these contacts very unsatisfactory as a significant fraction of the surface is involved.

Biochemistry, Sheffield

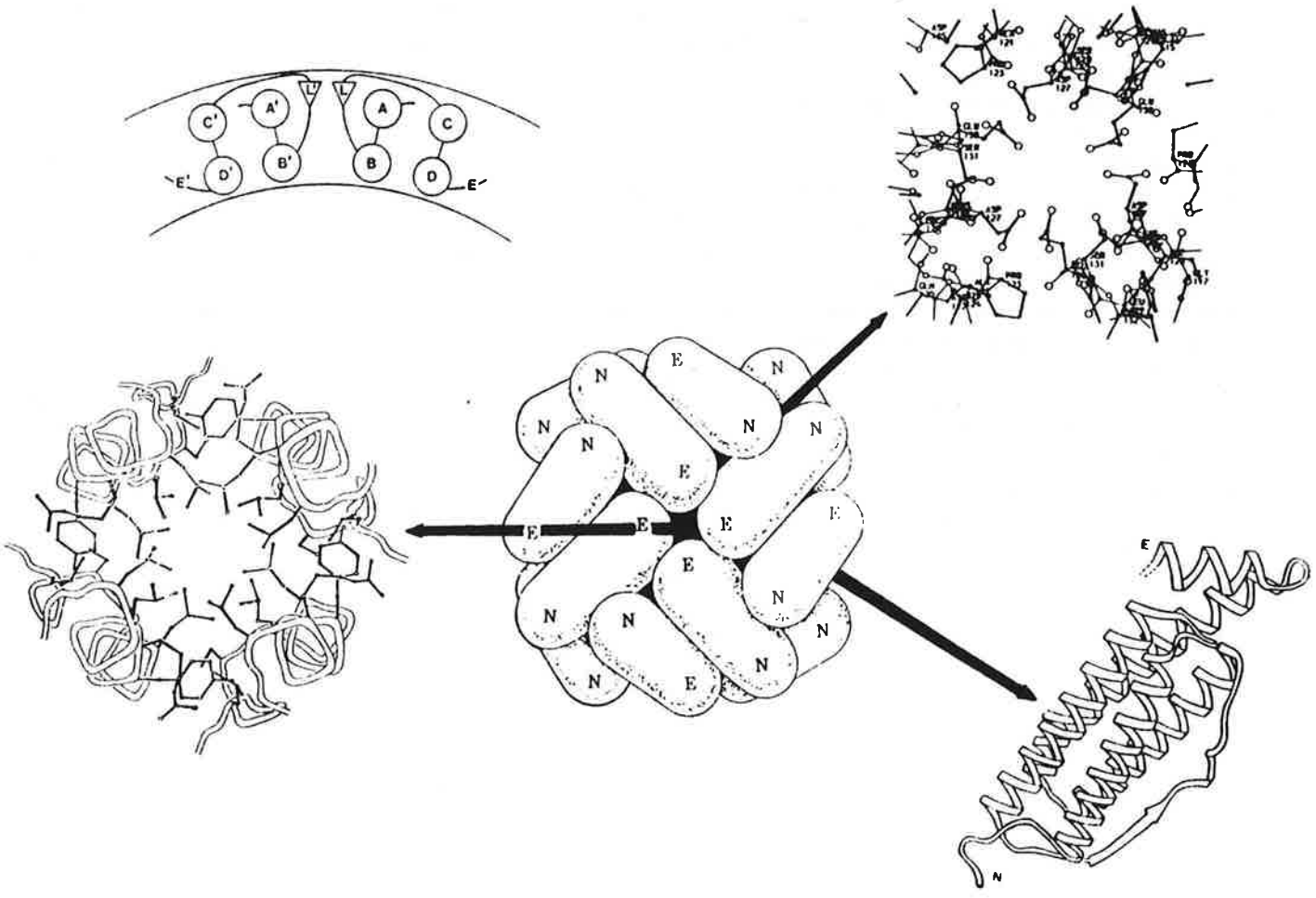


TABLE I

PRINCIPAL METAL BINDING SITES ON APOFERRITIN

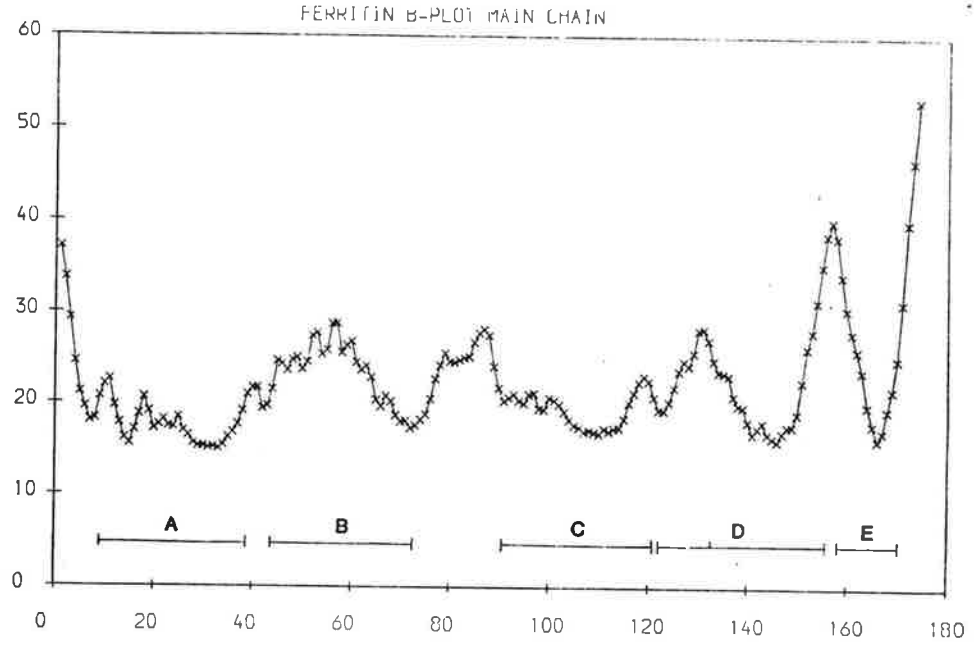
#	RESIDUES	METALS	LOCATION
1	ASP 80, GLN 82',	Cd,Tb,(Zn),(UO2)	Outside on 2-fold
2	ASP 127, ASP 127', ASP 127''	Cd,Zn,(Hg)	Inner 3-fold channel
3	GLU 130, GLU 130' GLU 130'', 3 H_2O	Cd, Zn	Outer 3-fold channel
4	HIS 132, ASP 135'	Cd,Zn,(Tb),(SO4)	Inside near 3-fold
5	ASP 38, GLU 45, CYS 48	Cd,Zn,(Tb),UO2, Hg	Inside near 2-fold
6	GLU 130, GLU130', GLU 130'', ASP127, ASP 127', ASP 127''	Tb	Middle 3-fold channel
7	GLU 53, GLU 56	Tb	Inside in 2-fold groove
8	GLU 57, GLU 60	Tb	Inside on B helix
9	CYS 126	Hg	Outer lip 3-fold channel
10	GLU 63, ARG 59, (GLU 56'),(ARG 52')	UO2	Inside in 2-fold groove
11	GLU 57, H_2O, GLU 136	UO2	Inside surface at BD interface

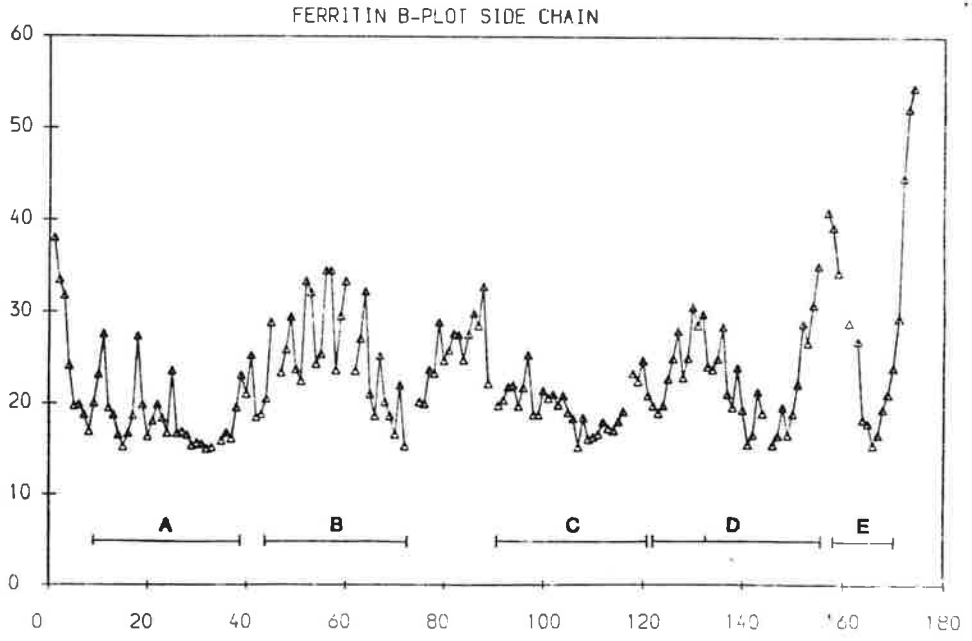
Hg = PCMB = $^{+}Hg-\phi-COO^{-}$ UO2 = $(UO_2F_5)^{3-}$ = UO_2^{2+}

Cd = Cd^{2+} Zn = Zn^{2+} Tb = Tb^{3+} SO4 = SO_4^{2-}

( ) = Tentative or low occupancy

___ = Residues conserved in known sequences





Determination of the wavelength normalization curve in the Laue method

J. Campbell[1], J. Habash[2], J.R. Helliwell[1,2] and K. Moffat[2,3]

[1]SERC Daresbury Laboratory, Warrington WA4 4AD, UK

[2]Department of Physics, University of York, Heslington, York YO1 5DD, UK

[3]On leave from Section of Biochemistry, Molecular & Cell Biology, Cornell University, USA

Introduction

The overall processing strategy for broad wavelength range Laue diffraction patterns, on which the current Laue package at Daresbury is based, has been given by Helliwell (1985). Details of processing results on data sets from pea lectin and a small organic crystal were given in issue 15 (1985) of this Newsletter. Zurek et al also reported in issue 16 (1985) on the unscrambling of doublet and triplet reflections.

The recorded Laue intensity (strictly, an integrated power) is given for the reflection $\underline{k}$ by (Kalman, 1979; Moffat et al 1984):

$$I_L(\underline{k}) = \left(\frac{e^2}{mc^2}\right)^2 \frac{dI}{d\lambda} \lambda^4 \cdot \frac{1}{2 \sin^2 \theta} \cdot \frac{V}{V_0^2} \cdot P.T.S. \left|\underline{F}(\underline{k})\right|^2 . \qquad (1)$$

Here, $dI/d\lambda$ denotes the spectral intensity distribution of the incident X-ray beam; V is the volume illuminated; V_0 is the unit cell volume; θ is the Bragg angle for the reflection $\underline{k}$; P is the polarization factor; T is a transmission factor arising from absorption in the sample, capillary, diffracted

beam path and detector window; and S is a detector sensitivity and obliquity factor. Quantities such as P, T and S may vary with any or all of λ, θ and $\underline{x}$, the position of the diffracted beam on the detector; the spectral intensity distribution is in general not precisely known; and the detector may suffer from spatial distortion and non-uniformity. Thus equation (1) may be written as

$$I_L(\underline{k}) = K\, G(\lambda,\theta,\underline{x})\, \left|\underline{F}(\underline{k})\right|^2 \qquad (2)$$

where K is a constant. Assume that all quantities which depend on more than one variable are factorable. Then $G(\lambda,\theta,\underline{x}) = f(\lambda)g(\theta)h(\underline{x})$, and

$$\left|\underline{F}(\underline{k})\right|^2 = \left[K\, g(\theta)h(\underline{x})\right]^{-1} \left[f(\lambda)\right]^{-1} I_L(\underline{k}). \qquad (3)$$

We make the reasonable assumption that K, $g(\theta)$ and $h(\underline{x})$ are known, to a very good approximation. Thus, the quantifying of Laue diffraction patterns depends almost entirely on determination of $f(\lambda)$, known as the wavelength normalization curve or λ-curve.

Several strategies for this are possible. The λ-curve may be measured experimentally, on the sample itself (for example, via an unscreened oscillation or precession photograph of low angle; Moffat et al 1984, and unpublished Cornell results) or on a reference sample such as a Si crystal (Wood et al 1983); it may be derived by reference to a monochromatic data set, the approach used hitherto (Helliwell, 1985; D. Szebenyi, B. Smith and K. Moffat, unpublished); or it may be derived internally from the sample Laue data set alone (a strategy which has been most vigorously advocated by the Cornell group).

The last strategy depends on the stimulation and recording of symmetry-related reflections by different wavelengths. Each set of such reflections then samples different points on the λ-curve and hence are recorded with different intensities. The deviations from equality allow the λ-curve to be deduced. We present here a preliminary application of this strategy.

Method

A single Laue pattern contains a large number of reflections in the protein crystal case, which in the initial process of refinement are assigned to different wavelength intervals or bins. It is possible therefore to use reflection intensities measured in one wavelength interval or bin to serve as a reference and one can then scale all other wavelength bins to that one. The use of a short λ reference bin should automatically lead to a flat absorption surface for the crystal. This is an important advantage.

A program has been written for future inclusion in the CCP4 suite and is called LAUENORM which uses an iterative curve fitting procedure to determine the λ-curve. A program description can be obtained from Daresbury.

Results

The Laue data recording conditions for pea lectin were reported previously (Hails et al.,1984; Helliwell, 1984). The crystal axes were miss-set from the principal axes of the camera by no more than 0.2°. Across the mounting axis, this produces a wavelength difference between symmetry-related reflections of roughly 0.03 Å, for θ = 10° and λ = 1.5 Å for example.

Table 1 shows the matrix of symmetry-related reflection overlaps between wavelength bins, for one of the pea lectin Laue film packs. Because the crystal was nearly perfectly set, the matrix has a diagonal band form. With

Table 1.

An upper diagonal matrix showing the number of overlap or equivalent reflection intensities falling in different λ bins. The mean wavelength of each bin number can be found by reference to the figure.

		0.5 Å ———														2.0 Å
λ bins		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0.5 Å	1	11	7	19	7	2	0	0	0	0	0	0	0	0	0	0
	2		55	50	43	14	0	0	0	0	0	0	0	0	0	0
	3			97	44	33	6	0	0	0	0	0	0	0	0	0
	4				122	36	28	0	0	0	0	0	0	0	0	0
	5					109	63	13	0	0	0	0	0	0	0	0
	6						158	62	14	0	0	0	0	0	0	0
	7							153	85	14	0	0	0	0	0	0
	8								182	80	14	0	0	0	0	0
	9									178	92	13	0	0	0	0
	10										193	63	13	0	0	0
	11											142	54	15	0	0
	12												113	43	11	0
	13													92	32	7
	14														65	19
2.0 Å	15															38

Table 2

As for Table 1 but showing the R-factors (on I) between equivalents in different λ bins BEFORE (above the diagonal) and AFTER (below the diagonal) λ NORMALIZATION. Note especially the improvement in the off-diagonal terms after normalization.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.058 .058	.070	.367	.409	.384	0	0	0	0						
2	.059	.063 .056	.226	.352	.371	0	0								
3	.170	.140	.059 .051	.155	.268	.183									
4	.063	.106	.086	.063 .054	.152	.135									
5	.035	.061	.077	.113	.053 .052	.120	.161								
6			.072	.063	.096	.067 .056	.183	.183							
7					.073	.097	.050 .047	.123	.097						
8						.081	.076	.051 .049	.105	.108					
9							.054	.080	.057 .056	.135	.109				
10								.101	.123	.066 .066	.132	.142			
11									.117	.135	.057 .057	.109	.146		
12										.129	.105	.058 .056	.121	.195	
13											.124	.115	.083 .082	.173	.278
14												.106	.109	.084 .075	.218
15													.193	.087	.087 .074

$$\text{R-factors on I} \quad \frac{\left(\Sigma \left|(I_{hk\ell}(i) - I_{hk\ell}(j)\right|\right)}{\left(\Sigma(I_{hk\ell}(i) + I_{hk\ell}(j))\right)}$$

a more miss-set crystal, the off-diagonal elements would also be well populated. Clearly, this would be a more desirable situation, which seems likely to yield an even more precise λ-curve.

Figure 1 shows the λ-curve for one pack determined via LAUENORM, compared with the λ-curve determined for the identical data using a monochromatic pea lectin reference data set and the program LAUESCALE. With the exception of wavelength bin 14 (where there is an unexplained glitch in the data), the results are very comparable.

A separate λ-curve was determined for each of the 5 film packs, by both strategies, and the data for each strategy were merged via the program ROTAVATA/AGROVATA. To 3.0 Å resolution, the overall merging R-factor on intensities is 12.6% via the LAUENORM strategy using internal scaling, and 12.1% via the LAUESCALE strategy using a monochromatic reference set.

Conclusions

Despite the fact that the pea lectin data was derived from a nearly perfectly set crystal the results are very similar by both strategies. The external scaling strategy is obviously not affected by any residual systematic errors in the monochromatic data, such as inadequately corrected absorption effects. Furthermore, the internal scaling strategy may be applied in cases where the crystals are not strictly isomorphous with those which yielded the monochromatic data. Such may often be the case in kinetic crystallography experiments (Moffat et al 1986; Hajdu et al 1986); this is an important advantage.

In the future, data will be obtained on more miss-set crystals. One of us (K.M.) suggests that the strategy may be extended to include

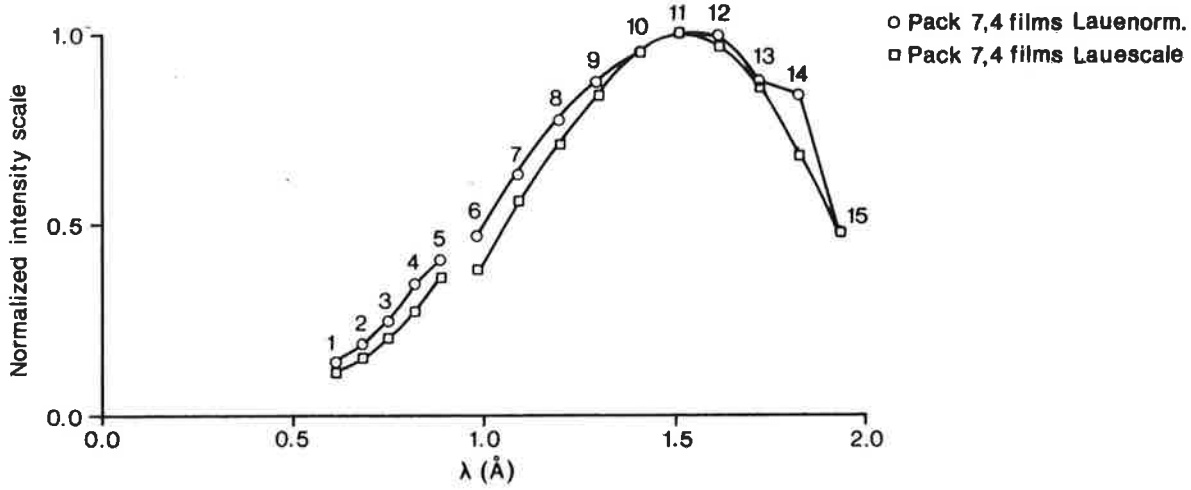


Fig. 1

Fig. 1. Typical wavelength normalization curve: λ-curve for pack 7 via LAUESCALE and LAUENORM.

(LAUESCALE involves comparison with monochromatic data)

(LAUENORM is an internally derived λ-curve using equivalent reflections)

(the number refers to the λ-bins - see Tables 1 and 2)

position-dependent effects through $g(\underline{x})$ (equations 2 and 3), the measurement of the same reflection at different wavelengths through a known crystal rotation for the triclinic case, the further refinement of cell and orientation parameters and hence λ, and as an alternative idea for the unscrambling of doublets and triplets to the one used by Zurek et al (1985). Another alternative to determining the λ-curve internally is using the mean intensity in different λ bins (J.C.).

References

Clifton, I.J. et al., (1985) J. Appl. Crystallogr. 18 296-300.

Hails, J. et al., (1984) Daresbury Preprint, DL/SCI/428E

Hajdu, J., Machin, P., Campbell, J.W., Clifton, I., Zurek, S., Gover, S. and Johnson, L.N. (1986) Information Quarterly of CCP4, No.17, Daresbury Laboratory.

Helliwell, J.R. (1984) Rep. Prog. Phys. 47, 1403-1497.

Helliwell, J.R. (1985) J. Mol. Struct. 130 63-91 (Proceedings of the Cruickshank Symposium, held September 1984).

Information Quarterly of CCP4, No.15, Daresbury Laboratory, July 1985.

Kalman, Z.H. (1979) Acta Crystallogr. A35, 634-641.

Moffat, K., Szebenyi, D.M.E. and Bilderback, D.H. (1984) Science 223, 1423-1425.

Moffat, K., Schildkamp, W., Bilderback, D.H. and Volz, K. (1986) Nucl. Instrum. Meth., in press (Proceedings of the SRI Symposium, held July-August 1985).

Wood, I.G., Thompson, P. and Mathewman, J.C. (1983) Acta Crystallogr. B39, 543-548.

Zurek, S. et al., (1985) Information Quarterly of CCP4, No.16, Daresbury Laboratory.

An improved program package for the measurement of oscillation photographs

Andrew G.W. Leslie, Peter Brick & Alan J. Wonacott,
Imperial College, London.

Approximately one year ago the processing of oscillation photographs at Imperial College was transferred from the Data General Nova minicomputer to the VAX-11/750. Subsequently a considerable programming effort has been invested in improving the ease of use of the MOSFLM suite of programs. This has included the automatic processing of 'still' photographs, a new algorithm for the generation of reflection lists, profile fitting, and the implementation of keyworded input and extensive machine-readable documentation. A brief description of the various programs involved is given below.

1) SCANFILM

This program controls the Scandig microdensitometer and writes a digitised film image to disk. It makes use of a VAX specific device driver written by Nigel Gammage of Joyce Loebl. The image is written as a direct access file with one 'stripe' of optical densities (2400 bytes for a 50 micron scan) per record. At 50 microns, each image is thus 6 Mbytes in size. All films within one film pack are written to the same direct access file, which is used as input to either STILLS or MOSFLM. The program is run interactively, and it takes about 9 minutes (elapsed time) to scan one film with a 50 micron step size. The disk space currently available allows between 30 and 40 film images to be stored at any one time.

2) STILLS

This program is used to create a file containing the coordinates of spots present on 'still' photographs that can be used by IDXREF to determine the crystal orientation and cell parameters.

Up to three 'stills' are digitized using SCANFILM and written to a disk file. The program reads this file and finds the coordinates of the fiducial marks and direct beam (if present). It then searches for diffraction spots on the film by looking for pixels with an optical density above a threshold level. Those pixels above threshold that are adjacent to one another or within some pre-defined distance from the spot centroid are considered to belong to a single spot. Spots whose size differs significantly from the median size are rejected. The algorithm is very robust, and will cope with split spots without difficulty. The film is then divided into bins and the coordinates of the strongest spots in each bin are

written to the output file.

3) IDXREF

This program is used to determine the cell parameters, crystal orientation and the crystal-to-film distance from the position of spots on 'still' photographs. The basic algorithm remains the same but the users interface has been rewritten to allow key-worded input. The program is completely general, and in particular the user no longer need supply a space-group specific subroutine (MATS) to define the derivatives of the orientation matrix. It is now also possible to constrain any of the cell parameters to their input values (using code taken from Phil Evans's post-refinement program) and to refine mosaic spread or beam divergence where appropriate.

4) POSTCHK

The program enables one to check for crystal slippage following the measurement of film packs by MOSFLM, making use of the common partial reflexions recorded on adjacent contiguous packs. A file of suitable reflexions with partial fractions is prepared from one or more 'generate' files which is then passed to IDXREF for refinement of orientation or crystal parameters. Only minimal input is required since all relevant variables are stored in the 'generate' file.

In using the actual measured partial fractions, rather than the assumed half-recorded reflexions as in STILLS, the final rms. angular residual in IDXREF is significantly reduced with typical values 0.01 degress or less.

5) OSCGEN

The program uses information about the crystal and beam geometry provided by the user to create a binary file containing the indices and film coordinates of all reflections present on an oscillation film. This file is used by the program MOSFLM. Alan's original algorithm has been abandoned and replaced by one proposed by George Reeke and coded by Bob Sweet. The user no longer receives the message: "Lost in outer reciprocal space".

The new program has two further modes of operation. In "TESTGEN" mode it can be used to determine the number of overlapped reflections on each film for a given total rotation range as a function of the oscillation angle. In "UNIQUE" mode the program will write a file (in lcf format) containing the Miller indices (reduced to the appropriate unique portion of reciprocal space) of all reflections recorded by a specified total rotation. The rotation is generally divided up into "blocks" of, say, 5 degrees. This file can then be used in

conjunction with a file containing all possible unique reflections out to a given resolution limit in order to determine the total rotation required to obtain a complete dataset. This analysis is done by a separate program (COMPLETE), which analyses both the percentage of the unique data recorded and the multiplicity of observations as a function of the number of "blocks" included. The percentage of anomalous pairs present as a function of resolution is also calculated. This option can be very useful if the crystal is rotated about an axis other than a crystal symmetry axis.

6) MOSFLM

The MOSFLM program (which was based on MOSCO (Nyborg & Wonacott, 1977)) has been extensively restructured, since the restrictions originally imposed by the limited memory of the Nova (32K) are no longer relevant. The Nova version of MOSFLM was first modified by Pella Machin at the SERC Daresbury Laboratory to run on a VAX. The Daresbury version served as a starting point for the current program, in which all unnecessary scratch files (previously required for communication between program overlays) have been replaced by data transfer via common blocks. Keyworded input greatly simplifies the task of running the program, since most parameters have (hopefully) sensible default values. Another major change allows the program to be run on-line or as a batch job. On-line use is normally restricted to processing the first one or two films from a given crystal in order to determine an appropriate size for the measurement box and to check that there are no serious errors in cell parameters or crystal orientation. When running on-line, graphical output is available for Sigma or Tektronix terminals (or any terminal capable of working in Tektronix emulation mode). This output includes:

a) A display of the residual vectors between the observed and calculated positions of reflections used in the refinement of the transformation from the ideal film coordinates to scanner coordinates of the digitised film image.

b) If required, a display of the central region of the film image with the calculated positions of the reflections superimposed. This is invaluable if there are uncertainties in the camera constants.

c) A display of residual vectors between calculated and observed positions of all fully recorded reflections on the film (above a low threshold intensity). This display is useful in detecting errors in cell parameters or orientation angles, which give rise to characteristic types of patterns.

The graphical output can be extremely useful in dealing with

'problem' films.

Parameters included in refining the film to scanner coordinate transformation include the camera constants (ccx,ccy,ccomega), the crystal to film distance, a scale factor applied to the Y scanner coordinates (around the drum) to allow for the fact that because of the finite film thickness the true sampling distance is 0.15% greater than the nominal scan increment, and three distortion parameters, tilt, twist and bulge. Tilt and twist represent rotation of the plane of the film about horizontal and vertical axes and thus allow for misalignment of the cassette on the camera, while bulge models the non-planarity of the film which is frequently observed in practice (CEA and Kodak films seem to be much more susceptible to bulging than Ilford G).

Following refinement the final rms residual (for sixty reflections distributed over the entire film) is typically less than 20 microns for an 'A' film. This means that it is unnecessary to allow the measurement box to move for individual reflections, as is done in some film-measurement programs. In addition, no attempt is made to redefine background and peak areas for individual reflections.

The final major change to the program has been the incorporation of a profile-fitting algorithm. The program calculates a set of empirical 'standard' profiles for different areas of the film and uses this information to calculate profile-fitted intensities using equations given by Rossmann, 1979. The approach used in determining the standard profiles is rather different to that employed by Rossmann, and is described below.

The size and shape of a diffraction spot on the film varies as a function of Bragg angle as a result of the obliquity of the diffracted beams, the finite film thickness and the change in the projected shape of the diffracting volume of the crystal. In addition, reflections close to the oscillation axis are drawn out in a more complicated way, but this is not considered seperately (a complete description of the diffracted spot size for synchrotron geometry is given in Greenhough et al. (1983)). In the original version of MOSFLM the size of the peak component of the measurement box is automatically expanded in the X and Y directions (in steps of two pixels) as a function of Bragg angle to allow for this effect. The total number of expansions required depends on the beam geometry, collimator and/or crystal size, and the maximum Bragg angle. For example, 2.5 Å resolution films collected on the Wiggler station at Daresbury (0.3 mm collimator, 0.88 Å wavelength) generally require only one expansion in each direction, giving a total of nine different measurement boxes across the entire film. By contrast, 2.5 Å resolution films collected using graphite monochromatised copper radiation (wavelength 1.542 Å, beam divergence about 0.2 degrees), require three expansions in X and Y giving a total of twenty-five different measurement

boxes.

A separate empirical profile is calculated for each measurement box by summing the optical densities of all fully recorded reflections above a (very low) threshold intensity. This approach differs from that used by Rossmann, but better reflects the errors associated with individual optical density measurements. A best least-squares plane is then fitted to the backgound region of the summed optical densities, and an rms residual calculated. Any pixels with an optical density which deviates by more than three times this residual are then rejected and the background plane re-evaluated. The background-subtracted profiles are placed on a common scale, and two rejection criteria applied. The first is simply the number of reflections contributing to the profile (default minimum is 10), and the second is based on the rms variation in the background, which gives an indication of the 'noisiness' of the profile. Any profiles failing either test are 'improved' by forming an average profile over all the immediately neighbouring measurement boxes. Proper account is taken of the different sizes of the adjoining measurement boxes when forming this average.

In order to improve the quality of the profiles, particularly in the outer regions where there are fewer, weaker reflections, the program allows the formation of the standard profiles using a number of successive film packs (the only limitation being that they must be in the same 'generate' file). This is valid providing the crystal has not slipped appreciably during the recording of the films, and there is no significant change in spot shape (eg due to radiation damage). At Imperial we normally accumulate the profile over four successive packs. The profiles are usually determined only from the 'A' films, and these profiles are then used to evaluate all the films within the filmpack.

When an acceptable set of standard profiles has been determined, the spot intensities are calculated in a final pass through the film image. The profile-fitted intensity and standard deviation are calculated using the equations given by Rossmann (after correcting errors in the published equation for the standard deviation). In this integration pass, a best least-squares background plane is fitted to each reflection, and outliers are rejected in the same way as when forming the standard profiles. For each reflection (including partials) the integrated and profile-fitted intensities and their standard deviations are calculated and stored. Trevor Greenhough (personal communication) has demonstrated that the systematic error introduced when using a profile derived from fully recorded reflections to measure the intensity of partially recorded reflections vanishes when the two contributions to the partially recorded reflection are summed. Thus partials are treated in the same way as fully recorded reflections, except that a standard deviation based on the fit of the profile is not calculated, but set equal to the standard

deviation of the integrated intensity.

Results

Since both integrated and profile fitted intensities are stored for each reflection it is possible to evaluate the improvement in the quality of the data resulting from profile fitting. When scaling together films within one pack (A,B,C scaling), the overall R factor is generally about 1% lower for the profile fitted intensities (ie a reduction from 4% to 3%). However, the overall merging R-factor for a full three-dimensional dataset is usually identical for profile fitted and integrated intensities. The reason for this becomes apparent when the R factor is analysed as a function of intensity (table 1). Although profile-fitting significantly improves the R-factor for weaker reflections (intensity less than the overall mean), for stronger reflections the integrated intensities actually agree better. Because the overall R-factor is dominated by the stronger reflections, the final numbers are identical for profile fitting and simple integration. The reason for the poor agreement between the stronger reflections is currently under investigation, but it may be associated with the finite sampling size of the profile and the spot optical densities. When using a 50 micron step size, there can be a positional error of up to 25 microns in X and Y between the centroid of the standard profile and the centroid of any given reflection (ignoring errors in the calculated positions of the spots). This will introduce a systematic error in the profile fitted intensity whose magnitude will be proportional to the intensity of the reflection. One way around this problem would be to calculate a separate interpolated profile for each strong reflection, but this would be computationally expensive. Another difficulty is that the profile itself will be artificially broadened because of the finite sampling increment, so this too will introduce a systematic error which will be more important for strong reflections. A pragmatic solution to the problem would be to calculate a weighted average of the profile-fitted and integrated intensities, in such a way that the profile-fitted value dominates for weak reflections and the integrated value for strong reflections. Such an implementaion would require changes to both the ABSCALE and AGROVATA programs and is under consideration.

table 1.

The variation of the overall merging R factor for profile-fitted and integrated intensities as a function of intensity. The data are for a 2.5A resolution dataset from chloramphenicol acetyltransferase collected using monochromatised copper radiation (a total of 21 films). The mean multiplicity is 2.0. The intensity is given as multiples of the overall mean intensity of the dataset. Typically 10-12% of the total number of reflections on an 'A' film were overloads (greater than 2 OD). The number refers to the number of independent reflections in each bin.

INTEN	0	0.3	0.7	1.0	1.3	1.7	2.0	2.3	2.7	3.0	3.3
NUMBER	2766	1521	862	570	343	291	235	176	122	98	87
R(prof)	12.0	5.6	4.5	4.2	3.9	3.9	3.9	3.7	3.8	3.6	3.5
R(int)	15.3	5.9	4.4	3.9	3.7	3.4	3.5	3.3	3.3	3.4	3.2

R overall (profile) 4.4%
(integ.) 4.4%

7) IMPLEMENTATION

The programs are written in FORTRAN 77 suitable for a VAX machine. Limited use is made of VAX specific system calls, but extensive use is made of INTEGER*2 and BYTE variables, and many variable names exceed six characters in length. The subroutine libraries CCPLIB and FORTRAN 77 LCFLIB are required.

The source code and documentation is available on request.

References.

Greenhough, T.J., Helliwell, J.R. & Rule, S.A. (1983). J. Appl. Cryst. 16, 242-250.

Nyborg, J. & Wonacott, A.J. (1977). In The Rotation Method in Crystallography, (Arndt U.W. & Wonacott, A.J., eds) pp139-152, North Holland, Amsterdam.

Rossmann, M.G. (1979). J. Appl. Cryst. 12, 225-238.

An unusual crystal packing arrangement in phosphofructokinase

Wojtek Rypniewski, Phil Evans

MRC Laboratory for Molecular Biology
University Postgraduate Medical School
Hills Road
Cambridge CB2 2QH

We have crystallised the allosteric enzyme phosphofructokinase (PFK) from E.coli with 14% polyethylene glycol, 1M NaCl at pH 7.7, in the presence or absence of the allosteric inhibitor 2-phosphoglycolate. We hoped that these crystals would show us the structure in the allosterically inhibited T-state. They belong to space group C2 with cell dimensions 177.0, 66.4, 154.0 Å, β=118.8°. Native data were collected to a resolution of 2.4Å on the synchrotron in Daresbury. The merged data set contained 60000 unique reflections with an overall R_{sym} 0.104 after post-refinement.

We have solved the structure using the known structure of the same molecule in the active R-state. Location of this model in the observed data shows an interesting packing of molecules in the crystal.

From the unit cell volume, the crystals were estimated to contain one PFK tetramer in the asymmetric unit. Native Patterson and rotation function were calculated using low resolution data (7-5 Å) and examined to identify the non-crystallographic symmetry elements.

In general the effect of a non-crystallographic rotation axis is to produce a pseudo-Harker plane though the origin and perpendicular to the local symmetry axis. The rotation function transforms a Harker plane in the reciprocal space to produce a more easily identifiable peak in the direction of the symmetry axis. In some special cases however it is instructive to examine the native Patterson separately. In particular, when the local and the crystallographic axes have the same order (in this case both are two-fold) and are parallel, the combination of the two symmetry elements results in a pure translation. The Harker plane should have a clearly identifiable peak where the intermolecular vectors coincide, and the position of this peak indicates the position of the molecular axis relative to the crystal axis. If the two axes are not exactly parallel the vectors do not coincide and the peaks become smeared out.

In this case, the kappa=180° section of the self-rotation function showed several peaks at the periphery, which requires a peak at the origin to complete a 222 set. This implies that one of the molecular dyad axes is parallel to the crystallographic dyad along b. The native Patterson was then examined, but no large peak was found on the Harker section (v=0). This can be explained if the local axis is coincident with the crystal axis, and the self-rotation function can be explained by two tetramers each sitting on different crystallographic dyad axes, with the asymmetric unit containing two half-tetramers. Figure 1 shows the two orthogonal sets of peaks for the different tetramers at 45° to each other. Between them there are two strong peaks relating the two molecules. These peaks are at double weight as there are twice as many vectors in superimposing one tetramer onto the other as in superimposing two half-tetramers. The mirror planes in 180° section of the self-rotation function can also

be seen in the precession photograph of the h01 zone.

The cross-rotation function calculated using the known R-state model confirms this arrangement and shows which molecular axes (pqr) correspond to the non-crystallographic axes. Figure 2 shows the 180° rotation section of the cross-rotation function showing that the model tetramer can be superimposed on the two tetramers in this crystal form by a rotation of 180° about c or about an axis at about 22° to c (half way between 0 and 45°)

The relative position along y of the two tetramers was found by a translational R-factor search using data to 3Å. It involved moving one tetramer relative to the other since the overall origin is not fixed in space group C2. The R-factor dropped to 0.44 from an average of 0.51 at 0.69b (figure 3).

In summary a rather unusual packing arrangement is revealed of alternating, centred layers of PFK tetramers with the molecules in different layers rotated to each other by about 45°. Such arrangements of oligomeric proteins expressing crystallographic symmetry in different ways is not unknown, but is rare. The structure has now been refined from this starting position, at first using Corels treating each subunit as two rigid domains, and later using the Hendrickson-Konnert program with the FFT matrix calculation (Tony Jack's program DERIV). Unfortunately, the structure turns out to be very similar to that in the liganded R-state crystals, so this crystal form probably does not show us the inactive T-state conformation.

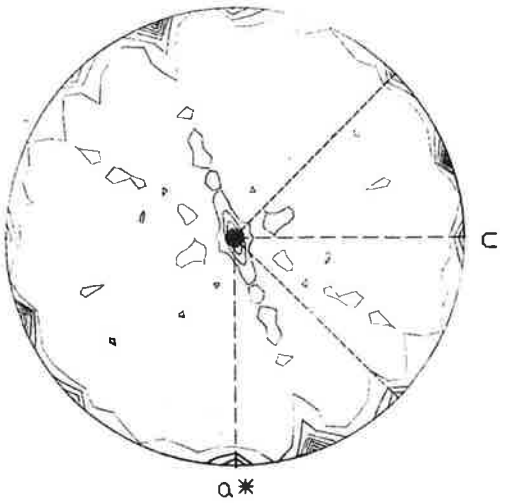


Figure 1

Stereographic projection of the 180° section of the self rotation function, (7 – 5Å resolution, integration radius 29Å) viewed down the b axis. The two sets of dashed lines show the peaks related by the 222 symmetry of the two tetramers. The large peaks on the periphery between these represent the vectors between one tetramer and the other.

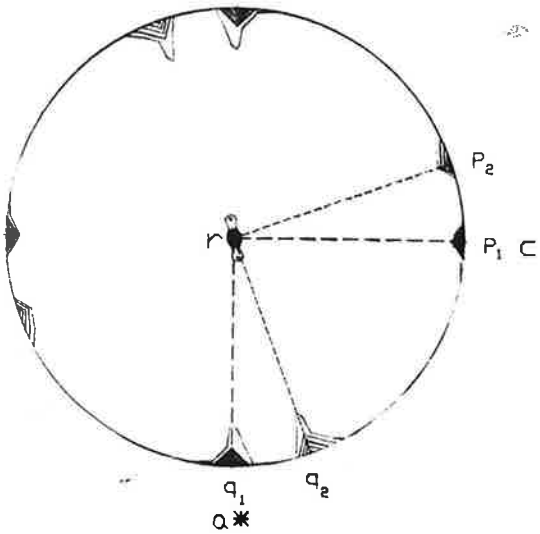


Figure 2

Stereographic projection of the 180° section of the cross rotation function, (7 – 5Å resolution, integration radius 29Å) viewed down the b axis. The dashed lines connect the peaks representing the rotations of a model tetramer on to the two different tetramers in the crystal.

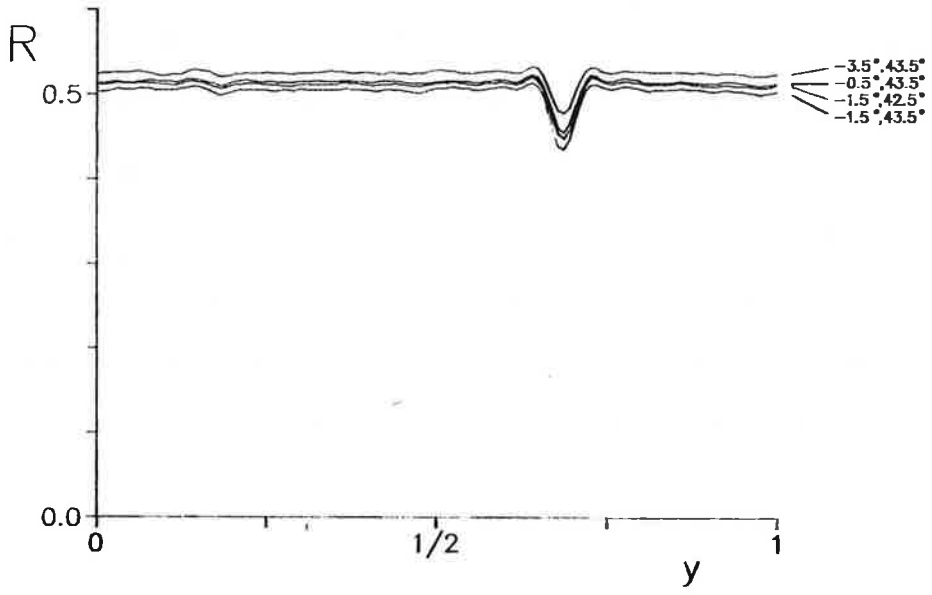


Figure 3

R-factor as a function of the relative translation along y between the two tetramers on different crystallographic dyads. Each line is for a different combination of rotations of the two tetramers about the y axis.

Changes to the MRC version of Frodo since August 1985

This list may be useful to current users of my version of Frodo.

Phil Evans, MRC Laboratory of Molecular Biology, Cambridge

12/6/86 E2.4 Changed READIT so that Update mode in SAM>READDIAM/READPDB updates pseudo-sphere of residues that are updated

22/5/86 E2.4 VAXSUBS should have been compiled /check=nooverflow
Various released versions have not had this done. If so,
MAP 4 blows up (overflow). To fix, in directory [.frodo.source]

```
$fortran vaxsubs/check=nooverflow
$libr frodolib vaxsubs
$@frlink
```

19/5/86 E2.4 increased MOLMEM array sizes to 16384 atoms in virtual memory

9/5/96 E2.4 fixed CREWAT so that when you insert water residues, the internal residue number is updated on all atom records for subsequent residues. If this is not done, when you SAVE coordinates of residues beyond the insert, the wrong residue gets its pseudo-sphere updated. Note that the SAM command INSERT also does not update this internal residue number so the same problem happens after an INSERTion. However its effects are not serious.

2/5/86 E2.4 increased some array dimensions: can now take maps of up to 30000 bricks, with up to 3000 displayed at once.

1/5/86 E2.4 Substituted Ian Tickle's corrected TOR: this will search properly to the end of a molecule to activate it all as rotatable.

26/4/86 E2.4 Fixed bug in keyboard C to centre picture: LOADA .DIST/.ANGL can now hit point

24/4/86 E2.4 .CENT now can hit point
Sets "Fkeys always" mode on first entry to graphics bit, so that the keyboard function keys are always active even in Chat

22/4/86 E2.4 Converted all function networks to GSR calls using Ian Tickle's PSPARS program (see fro$fun:pspars.txt, download.com etc) This speeds up down load even on serial lines. DOWNLOAD.COM needs to be changed if you have a parallel interface.

22/4/86 E2.4 PSINIT clears out old bricks, MOLn objects etc,on entry

16/4/86 E2.4 fixed bugs in CREWAT and ADDWAT

4/4/86 E2.4 New commands to handle water molecules: SAM>WATER allows insertion of water residues and setting of pointers to next free water, menu option .Watr (replaces .Del) either adds next available water at the position of the point, or if .No is set deletes the next hit water atom (by setting it to a dummy position). This command is at present experimental!
Changes in DATASTRUC.FUN, CURSOR.MAC (menu item .del->.watr), SAM, GETPUT, RINGIO(PUTSEQ), NUPSEU, INTER, P3GSRIO, CHAT, new bits CREWAT, ADDWAT.

24/3/86 E2.4 The point can now be hit, and used by the menu commands .ORIG and .NAYB (not .DIST and .ANGL yet). An active point will be automatically accepted.

19/3/86 E2.4 New SAM io routines for Brookhaven and Diamond format READDIAM and READPDP/PDB now call routine READIT instead of READRD and PDB, MAKEDIAM and MAKEPDB now call MAKEIT instead of routines MAKERD and MAKEWH (old routines still in file in case they are needed). Differences:
Input: options to make, update or append to existing file
Brookhaven: uses chain identifier (if not blank) to append to residue name, doesn't prefix H for HETATM
Output: defaults zero B values to defined value [20]
Tries to set weights if zero on real atoms
Brookhaven: writes chain identifier if last character of residue name is not a digit.
Diamond: tries to extract residue number from residue name, either as 4-digit number or as 3-digit number+chain identifier

19/3/86 E2.4 MOL SYMGEN can now optionally use just crystallographic or non-crystallographic symmetry (SYMTRY, MOLCUL, MOLSUBS)

5/3/86 E2.4 foreground SYMM can now use all symmetry operations (as before), or just crystallographic or non-crystallographic operations.
Changes in CHAT, LOADA, SYMGEN(SYMTRY): 13/3/86 corrected CHAT

26/11/85 E2.3 Chat/MUVA command now updates pseudo-sphere radius of residue

10/9/85 E2.3 Changed MAPSUBS so that maps with more than 9999 bricks can be used (maximum currently 15000): brick names now are of the form BR_mlnnnnn where m is the map number, l the level number, and n (5 digits instead of 4) is the brick number

10/9/85 E2.3 Changed HBND so that the h-bond search will find no h-bonds within the same residue, if no hydrogens are present. If hydrogens are present, the search is safe within a residue.

This facility is available on Vax DLVB at Daresbury and is accessed by typing CCP4NEWS. The program allows the user to read a file of messages already present or add new messages. At the present time the facility is in a very simple form, but it is hoped to improve this if users find the system useful.

The current file contains a list of names and network addresses of protein crystallographers who can be reached on the JANET network. A copy of this list is also included below. If anyone would like to be added to this list would they please get in touch with me. We have received requests for network addresses of UK protein crystallographers from Alwyn Jones and Martha Teeter. Alwyn is compiling a list of network addresses of european crystallographers and Martha is compiling a similar list to include crystallographers from all over the world. I shall send them a copy of the following list unless anyone has any objections to their name being included.

Sue Bailey.

DON@UK.AC.LEEDS.BIOVAX	! Akrigg, Don
CCP4@UK.AC.DL.DLVD	! Bailey, Sue
UBCG11J@UK.AC.BBK.CR	! Bryant, Steve
WBTC@UK.AC.CAM.PHX	! Cruse, William
XTAL@UK.AC.YORK.CHEMVAX	! Dodson, Eleanor
UBCG03D@UK.AC.BBK.CU	! Driessen, Huub
MA6@UK.AC.DL.DLVD	! Glover, Ian
UBCG08A@UK.AC.BBK.CU	! Goodfellow, Julia
OK@UK.AC.CAM.PHX	! Kennard, Olga
EBM028@UK.AC.BRISTOL.BSA	! Lyle, Andrew
CDS@UK.AC.DL.DLVB	! Machin, Pella

UBCG05M@UK.AC.BBK.CU ! Moss, David

MUIRHEAD@UK.AC.BRISTOL.BSA ! Muirhead, Hilary

KUM@UK.AC.DL.DLVD ! Papiz, Miroslav

SEVP@UK.AC.LEEDS.BIOVAX ! Phillips, Simon

KUS@UK.AC.DL.DLVD ! Rule, Steve

EOBC12@UK.AC.ED.EMAS ! Sawyer, Lindsay

XBSH10@UK.AC.SHEFFIELD.SHGA ! Smith, John

GARRY@UK.AC.OX.BIOP ! Taylor, Garry

TICKLE@UK.AC.BBK.CR ! Tickle, Ian

WATSONHC@UK.AC.BRISTOL.BSA ! Watson, Herman

BI1JW@UK.AC.SHEF.IBM ! White, Jan

DISTRIBUTION OF THE LEEDS STRUCTURE PREDICTION SUITE

There have been some discussions recently about program distribution.

To clarify the situation with respect to prediction programs which have been developed at Leeds (and are distributed from Leeds rather than from Daresbury) Professor Tony North has written the following.

The Leeds Protein Structure Prediction Suite is supplied free of charge to members of CCP4 (in return for a blank magnetic tape). Overseas academics are asked to pay 50 pounds to cover the cost of magnetic tape, airmail postage, etc. Commercial groups are asked to pay a licence fee.