
ACCURACY AND RELIABILITY OF MACROMOLECULAR CRYSTAL STRUCTURES

Proceedings of the CCP4 Study Weekend,
26 - 27 January, 1990

Compiled by K. Henrick, D.S. Moss and I.J. Tickle

Science and Engineering Research Council
DARESBURY LABORATORY
Warrington WA4 4AD, U.K.

© SCIENCE AND ENGINEERING RESEARCH COUNCIL 1990

Enquiries about copyright and reproduction should be addressed to:- The Librarian, Daresbury Laboratory, Daresbury, Warrington WA4 4AD.

ISSN 0144-5677

IMPORTANT

The SERC does not accept any responsibility for loss or damage arising from the use of information contained in any of its reports or in any communication about its tests or investigations.

ACCURACY AND RELIABILITY OF MACROMOLECULAR CRYSTAL STRUCTURES

Proceedings of the CCP4 Study Weekend
26 - 27 January, 1990

Compiled by
K. Henrick, Daresbury Laboratory
and
D.S. Moss and I.J. Tickle, Birkbeck College, London

SCIENCE & ENGINEERING RESEARCH COUNCIL
DARESBUURY LABORATORY
1990

PREFACE

At the time that this Study Weekend was held there were 500 coordinate entries in the Brookhaven Protein Data Bank, with more in the pipeline. These have been deposited by researchers employing X-ray, NMR and molecular modelling methods over the last 20 years. These coordinate sets are being used by a much larger community which includes many scientists with no background in structure determination or modelling, and often for purposes which the original depositor most likely never envisaged.

The problem which this meeting attempted to address is that in the PDB entries there is little or no direct indication of either the precision or accuracy of the individual atomic coordinates. Indeed even the depositor usually has only a vague notion of the overall coordinate error, let alone that of individual atoms. Consequently the novice is liable to make use of the information in the Data Bank without questioning its reliability, whilst the more experienced will be forced to make estimates of the errors by various indirect means, for example from the values of the thermal parameters (though frequently not supplied), or by analysis of main and side-chain torsion angles, or by comparison of independently determined structures.

This Study Weekend aimed to make a contribution towards remedying this situation. Time was devoted to an examination of the various sources of error which are propagated into the coordinates: experimental errors, and errors in phase determination, electron density interpretation and structure refinement. Then a session on "Quality Control in Coordinate Databases" aimed to provide an analysis of the errors found in the PDB entries, and to answer the important question of the validity of using such "knowledge-based" information to feed back into new structure determinations. Most of the second day of the meeting was set aside to allow speakers to recount a number of case studies, where the aim was to learn from the mistakes made by even the most experienced!

The meeting was organised and supported by the SERC Collaborative Computational Project in Protein Crystallography (CCP4) at Daresbury Laboratory. We wish to thank the invited speakers for their considerable efforts in making the meeting a success and their cooperation in the preparation of these proceedings. Particular thanks go to David Moss and Ian Tickle for the considerable time and effort they invested in the planning of the meeting.

We thank the Daresbury Laboratory and its Director Professor A.J. Leadbetter, for the provision of organisational help and support, for both the meeting and in the publication of the proceedings. In particular we thank Shirley Lowndes, David Brown and Pauline Shallcross for their great assistance in the planning and organisation of the Study Weekend. In addition the proceedings owe much to the efforts of Geoff Berry and Tracey Booth.

Kim Henrick
July 1990

CONTENTS

	<u>Page</u>
Preface	(iii)
Invited Speakers' Contributions	
Theory of errors J.S. Rollet, University of Oxford	1
Errors in macromolecular crystallographic data collection with area detector instruments P.A. Tucker, EMBL, Heidelberg	11
Notes on the errors of phase determination in the multiple isomorphous replacement method and the molecular replacement method Z. Derewenda, University of York	25
Analysis of errors found in protein structure coordinates in the Brookhaven Data Bank J.M. Thornton, M.W. McArthur, D.K. Smith, S.P. Gardner, E.G. Hutchinson, A.L. Morris and B.L. Sibanda, Birkbeck College, London	39
Bad contacts in protein structures S.A. Islam and M.J.E. Sternberg, Imperial Cancer Research Fund, London, and D.L. Weaver, TUFTS University	53
Structure determination of turkey egg lysozyme S.E.V. Phillips, W.S. Somers, T.N. Bhat and M.R. Parsons University of Leeds	63
The RuBisCO saga H.A. Schreuder, P.M.G. Curmi, D. Cascio and D. Eisenberg University College of Los Angeles	73
Validation of protein structures - a case study: the small subunit of Rubisco S. Knight, I. Andersson and C-I. Brändén Swedish University of Agricultural Sciences, Uppsala	83
Hallmarks of a wrong structure C.D. Stout, Research Institute of Scripps Clinic	91
A tale of four iron-sulfur proteins: sequence errors and other matters E.T. Adman, University of Washington	97
Discussion Contribution	
Uncertainty and bias in difference Fourier maps U. Derewenda, E. Dodson, G. Dodson, D. Hodgkin, and H. Swift, University of York	103
Summary	
Accuracy and reliability of macromolecular crystal structures Closing remarks D. Blow, Imperial College, London	115
List of Delegates	119

Theory of Errors

J.S.Rollett

*Oxford University Computing Laboratory
11 Keble Road, Oxford, OX1 3QD, England.*

1 Introduction

This paper discusses the theory of errors for the method of least squares in general, but will be illustrated by its application to crystal structure refinement.

When a crystal structure has been refined using diffraction data it is usually compared with other similar structures and with molecular structure data from a variety of different physico-chemical methods - spectroscopic, NMR, quantum mechanical calculations and so on. Discrepancies are compared with the error estimates to decide whether the differences represent real effects worth further study or whether they are simply caused by errors of some kind in the analyses.

There are two dangers in such comparisons. One is that we may think that a discrepancy is a real effect when it is caused by errors of analysis. If we pursue the supposed effect we waste our time because we are hunting for something that does not exist. The other danger is that our error estimates are too large. We then ignore a real effect believing it to be due to errors, and we fail to use the structure analysis as fully as we could.

It is clear that we need error estimates which reflect the full errors - systematic as well as random - of our analyses, and we need to be assured that our confidence intervals for parameter values are realistic.

2 Basis of error analysis

First we will see how the least squares process is carried out, then we will consider how the errors of the data produce errors in the answers.

If we minimize a function $M(x)$, we analyse the error of the result by treating the function as a parabola near the minimum. Consider a simple function of one variable to understand the principles. If two points A and B are close together, then Taylor's theorem gives

$$\left(\frac{\partial M}{\partial x}\right)_B = \left(\frac{\partial M}{\partial x}\right)_A + \delta x \left(\frac{\partial^2 M}{\partial x^2}\right)_A + \frac{1}{2}\delta x^2 \left(\frac{\partial^3 M}{\partial x^3}\right)_\xi,$$

where δx is $x_B - x_A$ and ξ is between A and B . The Newton method drops the last term on the right - which is equivalent to assuming that M is parabolic. If we then assume that B is the minimum, we have zero on the left hand side and we get

$$\left(\frac{\partial^2 M}{\partial x^2}\right)_A \delta x = -\left(\frac{\partial M}{\partial x}\right)_A$$

The method converges in one step if M really is parabolic, otherwise we need to iterate, but convergence is fast if the derivatives are continuous and we start near to the minimum.

The generalisation to $M(\mathbf{x})$, a function of an m -vector \mathbf{x} of parameters, is straightforward. We get

$$\sum_{i=1}^m \frac{\partial^2 M}{\partial x_j \partial x_i} \delta x_i = -\frac{\partial M}{\partial x_j}, (j = 1, 2, \dots, m)$$

and the iterative use of this to produce corrections $\delta \mathbf{x} = (\delta x_1, \delta x_2, \dots, \delta x_m)$ converges if the matrix of the second derivatives of M is positive-definite. It may not be so for a least squares problem. If

$$\Delta_i = I_i^o - I_i^c, (i = 1, 2, \dots, n)$$

are discrepancies between observed values I_i^o and calculated values I_i^c which should be equal to them, with $n > m$, then the least squares method minimizes

$$M = \sum_{i=1}^n \omega_i \Delta_i^2,$$

where $\omega_i \geq 0$ are weights. There are several options for the choice of the I_i^o values, including the raw measured intensities of diffraction, the squares of the structure factors and the structure factors themselves. A recent discussion of this and its relationship to error analysis is given by Schwarzenbach *et al* (1989). For the derivatives required by the Newton method, we get

$$\frac{\partial M}{\partial x_j} = \sum_{i=1}^n 2\omega_i \Delta_i \frac{\partial \Delta_i}{\partial x_j},$$

$$\frac{\partial^2 M}{\partial x_j \partial x_l} = \sum_{i=1}^n 2\omega_i \left(\frac{\partial \Delta_i}{\partial x_j} \frac{\partial \Delta_i}{\partial x_l} + \Delta_i \frac{\partial^2 \Delta_i}{\partial x_j \partial x_l} \right).$$

The second term on the right of the last equation can make the matrix indefinite, and so we usually drop it. This gives the Gauss-Newton approximation which is in universal use in structure refinement. For a description of the Newton and Gauss-Newton algorithms, modifications for difficult cases and adaptation to optimization with constraints, see for example Fletcher (1980), (1981). The Gauss-Newton method converges less fast than the Newton method itself, but is less likely to diverge at greater distances from the minimum. It is also cheaper to compute a set of equations without the last term. The matrix of the equations for the Gauss-Newton method is always positive definite unless the parameters that we refine are linearly dependent. That can occur, for example, if we refine separately the position coordinates for all atoms of a structure in the direction of a polar axis.

It is convenient to use a compact notation for the Gauss-Newton equations. Write

$$\Delta^T = (\Delta_1, \Delta_2, \dots, \Delta_n),$$

$$(J)_{ij} = \frac{\partial \Delta_i}{\partial x_j},$$

$$W = \text{diag}(\omega_1, \omega_2, \dots, \omega_n).$$

Then we solve

$$J^T W J \delta \mathbf{x} = -J^T W \Delta$$

so

$$\delta \mathbf{x} = -(J^T W J)^{-1} J^T W \Delta.$$

Now consider errors

If Δ changes by ϵ (vector of data errors)
then $\delta \mathbf{x}$ changes by $\delta \mathbf{q}$ (vector of answer errors)

and

$$\delta \mathbf{q} = -(J^T W J)^{-1} J^T W \epsilon.$$

We form the $m \times m$ matrix

$$\delta \mathbf{q} \delta \mathbf{q}^T = (J^T W J)^{-1} J^T W \epsilon \epsilon^T W J (J^T W J)^{-1}$$

of squares and products of answer errors, and take statistical expectations on both sides. Since everything except $\delta \mathbf{q}$ and ϵ is considered to be accurate, we get

$$E(\delta \mathbf{q} \delta \mathbf{q}^T) = (J^T W J)^{-1} J^T W E(\epsilon \epsilon^T) W J (J^T W J)^{-1}.$$

But $E(\epsilon \epsilon^T) = V$ the variance matrix of the Δ_i so

$$E(\delta \mathbf{q} \delta \mathbf{q}^T) = (J^T W J)^{-1} J^T W V W J (J^T W J)^{-1} \quad (1)$$

gives the variance matrix of the x_j . Ideally we take $W = V^{-1}$, so $VW = I$, and most of the expression cancels, leaving us with

$$E(\delta \mathbf{q} \delta \mathbf{q}^T) = (J^T W J)^{-1}. \quad (2)$$

This says that the variance matrix of the answers is just the inverse of the normal matrix. Note however the requirement that $W = V^{-1}$, which is, in general, difficult to satisfy. The major difficulty is to find V .

3 The variances of the residuals

The difficulty of deciding on the variance matrix, V , of the residuals between the observed and calculated quantities is that there are many sources of error.

The following checklist shows this:

- Random fluctuations in quantum counts
- Experimental accidents (sticking shutters, crystals falling off, or moving on mounts ...)
- Error in correcting for background radiation
- Error in correcting thermal diffuse scattering
- Effects of multiple reflections
- Error in absorption correction
- Error in extinction correction
- Effects of damage to crystal by beam

- Fluctuating power in beam
- Failure of counting train to record all quanta
- Incomplete allowance for anisotropic vibration
- Incomplete allowance for anharmonicity
- Error in correcting for substitution of one atomic species by another
- Error in allowing for disorder
- Effects of incorrect symmetry assignment
- Error in allowing for anomalous dispersion
- Deviations from isolated atom approximation

Some of these types of error are hard to correct - for example extinction and thermal diffuse scatter. Some may require many parameters - for example deviations from harmonic vibration requiring third order terms in the vibration model. There may be insufficient data to maintain an excess of number of data over number of parameters if we try to elaborate our structure model to make proper analysis of the effects which are present. This is particularly true in case of analysis of structures containing large molecules where vibration amplitudes and disorder effects cut off the data at relatively low order.

It is clear that where a discrepancy is due to a measurement error, we should do all that is possible to remove that error. Equally where a discrepancy is due to an effect not taken into account in the structure model used for the I^c we should, if possible, extend the model to allow for it. In practice we may not succeed with either type of remedy. We may lack knowledge, eg. of the detailed shape of the crystal and properties of its surroundings. We may be short of diffraction data. We may even fail to discover just what is causing the discrepancy that we see. On occasion, there are errors which have no large effect except on a small minority of the observations. This situation can be handled by the robust-resistant refinement technique discussed by Prince and Nicholson (1983). More generally the effects of errors are spread over the whole of the data, and no realistic assessment of error is possible unless they are considered in the analysis.

For such reasons we have to find ways of assessing the errors of the results in spite of the presence of systematic errors, that is to say, errors which are not reduced by repeating our measurements and which may correlate the residuals. It is not sufficient to estimate these entirely by considering the residuals between I^o and I^c , because there can be compensating errors.

Consider the relationship between absorption errors and errors in isotropic vibration parameters and scale factor. If a crystal absorbs radiation then the measured I^o will be reduced. The factor by which it is reduced will tend to be greater for forward scattering than for back scattering. This is because forward scattered rays must typically pass through the whole thickness of the specimen whereas much of the back scatter is from parts of it near to the rear surface. We can get a very similar effect on the I^c by lowering the overall scale factor and at the same time reducing the isotropic part of the vibration parameters. Hence two large errors affecting the Δ_i produce residuals smaller than either of the errors.

Note that there is an ambiguity between errors in the measured data and errors in the model used to produce the I_i^c . We can think of absorption as causing an error in I_i^o , or we can regard it as something we have failed to allow for in our calculation of I_i^c . The latter is the preferable viewpoint. If we put parameters for absorption into I_i^c as well as the scalefactor and the vibration parameters then we find that these various parameters are highly correlated. As a result there are large elements in the inverse matrix $(J^T W J)^{-1}$, and the variances allow for the correlation. In this sense, inclusion in the model is an alternative to the strategy of correcting the errors. This alternative is not available if we are short of data to define the extra parameters required.

In many cases it is possible to estimate the likely size of errors of each of the various types from first principles, and so to make estimates of their contributions to the elements of V , the variance matrix for the Δ_i . So long as the different types of error can be considered to be independent, their variance contributions simply add together. If we assess the errors in this way we can look at the estimated mean values of Δ_i^2 and $\Delta_i \Delta_j$ for classes of diffraction orders and compare with the actual values after refinement. Three things can happen

(a) There is no significant difference for any class of reasonably large size. Then we have no evidence of any mis-estimation of error and can use the estimated V matrix with confidence.

(b) The estimated values are smaller than the actual values. This is clear evidence that some type of error has been under-estimated or omitted. We should try to identify this type of error and allow for it, using the pattern of discrepancy as a guide. Even if we fail to identify the error we can increase the elements of V concerned to allow for 'some error or errors unknown'.

(c) The estimated values are larger than the actual ones. This is not clear

evidence that the errors have been over-estimated. It may be that they compensate for each other's effects on the residuals. In this case we should leave the elements of V unchanged unless we can find a blunder in the estimates.

If we make the best estimate that we can of V , in this way, then we can decide whether V is such that $VW = I$ or $VW = kI$ where k is a scalar constant. If so we are entitled to use equation(2), suitably scaled for k . If not then we should go back to equation(1). The writer does not know of any case in which a structure analyst has used equation(1), but the size and power of available computers are now such that it does appear to be a feasible option.

4 Linearity

Equations (1) and (2) are based on linear theory. We linearize the least-squares equations about our current position in parameter space and use the linear equations to get an adjusted position. At convergence we are linearizing about the converged position and assuming accuracy of the elements of J . That is to say, we assume that the derivatives of the Δ_i with respect to the x_j remain constant over the region in parameter space measured by the potential errors. For most structures this will be a good approximation and our error estimates will be valid. There are exceptions. Notably, a pseudo-symmetric structure will have a 'mirror image' structure at a nearby point in parameter space which gives just as good a value of the minimization function (ignoring any breakdown of Friedel's law) as the structure we have. We then get very low curvature of M between the two points and larger curvature in other directions away from our solution. In that case linear theory is not valid and we need to fall back on direct comparison between M values at our solution and at hypothetical alternative positions. For an example of the difficulties caused by such situations, see Haasnoot *et al* (1972) and Gopinathan *et al* (1974).

This problem is not likely to affect the majority of analyses, but we need to be on our guard to recognize situations in which the powerful methods that we can usually use will give misleading indications.

5 Variances, standard deviations and confidence intervals

The theory which has been presented so far gives us variances and covariances. The variance of x_i is the estimated value of $(x_i - x_i^t)^2$, where x_i^t is the (unknowable) true value of x_i . The covariance of x_i and x_j is the estimated value of $(x_i - x_i^t)(x_j - x_j^t)$. Variances are convenient because we can manipulate them. Commonly our least-squares parameter set will include atomic positional coordinates and we will wish to estimate the error of a bond length. Suppose (for simplicity) that the bond lies parallel to the x axis, then if the atomic coordinates are x_i and x_j the length is $x_i - x_j$. Writing σ^2 for variance and *cov* for covariance we have

$$\sigma^2(x_i - x_j) = \sigma^2(x_i) - 2cov(x_i, x_j) + \sigma^2(x_j).$$

Once we have found the variance of a derived quantity, such as a bond length, we can take its square root, which is the standard deviation σ .

We are not usually interested in the standard deviation of a quantity for its own sake (except in so far as publication rules require us to produce it). The purpose of finding standard deviations is to decide whether the quantity could deviate from its value determined sufficiently to agree with some alternative value, or whether the probability of so large a deviation is too small to be accepted.

A common way to deal with such a question is to define a 'confidence interval'. We say that we have determined the value as x and the probability that $|x^t - x| > k\sigma$, where σ is the standard deviation of x , is some small fraction ϵ . values of 0.05, 0.01 and 0.001 for ϵ are widely used and define 5, 1 and 0.1 per cent 'significance levels' respectively.

We analyse this by considering the probability density function $p(x)$ for x . For simplicity we suppose that the distribution of x has mean zero. We take the probability that $a \leq x \leq b$ to be

$$P(a \leq x \leq b) = \int_a^b p(x) dx$$

The multiple k of σ giving $y = x - x^t$ that we can have before

$$P(|y| \geq k\sigma) = \int_{-\infty}^{-k\sigma} p(y) dy + \int_{k\sigma}^{\infty} p(y) dy$$

falls to, say, 0.05 depends on the nature of $p(x)$. If we know that $p(x)$ is that for the normal distribution this multiple is small. Otherwise it may be much larger, but we can still bound it. We will consider in the next section whether we are entitled to use the normal distribution.

6 Normal and non-normal distributions

There is a 'central-limit theorem' which says that if an error is the sum of a large number of different independent errors, each of which has a finite variance of similar size to the others, then the total error is approximately normally distributed.

The use of the normal distribution has been justified for structural parameters on the grounds that the error in such a parameter is the sum of many similar errors, each due to the error of one I_i^o . Since the I_i^o are regarded as independent observations, the conditions of the central limit theorem apply.

This is satisfactory if the errors are solely due to random errors of quantum counting, but that is, unfortunately, seldom true. Contributions from errors such as absorption and from errors caused by an inadequate structure model give potentially large correlations between the Δ_i . Then it is unsafe to assume normal distribution of a structure parameter. We still have a finite variance for it from equation (1). For any distribution of finite, known, variance and zero mean we can show that $P(|x| \geq k\sigma) \leq 1/k^2$. This gives much wider confidence limits than the normal distribution provides, but it may be prudent to use the wider intervals. This is especially so if the penalty for underestimating the interval is serious.

Table 1 indicates how the difference between the two situations affects our ability to exclude large deviations.

7 Conclusions

This paper has outlined means by which confidence intervals for structural parameters can be assessed, even if it is infeasible or uneconomic to correct systematic errors in the structure analysis. The methods suggested are not yet embodied in standard structure-analysis packages, but there seems to be no reason why this cannot be done. If it is not possible to remove substantial

$P(x \geq k\sigma)$	k_1	k_2
0.05	1.96	4.47
0.01	2.58	10.00
0.001	3.29	31.62

Table 1: k_1 is the multiple of σ for which the probability that the error can exceed $k_1\sigma$ is equal to $P(|x| \geq k\sigma)$ if the distribution is normal. k_2 is the corresponding multiple for which the probability does not exceed $P(|x| \geq k\sigma)$ if the distribution has finite variance.

systematic errors then the confidence intervals may be large in relation to standard deviations, especially for low percentage significance levels.

8 References

Fletcher,R. Practical Methods of Optimization, Vol 1, Unconstrained Optimization, (1980), Wiley, Chichester.

Fletcher,R. Practical methods of Optimization, Vol 2, Constrained Optimization, (1981), Wiley, Chichester.

Gopinathan,M.S., Whitehead,M.A., Coulson,C.A. Carruthers,J.R. and Rollett,J.S. Acta. Crystallogr.**B30**(1974)731-737.

Haasnoot,J.G. Verschoor,G.C., Romers,C. and Groeneveld,W.L. Acta. Crystallogr.**B28**(1972)2070-2073.

Prince,E. and Nicholson,W.L. Acta. Crystallogr.**A39**(1983)407-410.

Schwarzenbach *et al* [11 authors] Acta. Crystallogr.**A45**(1989)63-75.

Errors in Macromolecular Crystallographic Data Collection with Area Detector Based Instruments

P.A.Tucker

Biological Structures Programme
EMBL, Meyerhofstrasse 1
D6900 Heidelberg, F.R.G.

Introduction

The purpose of this article is to discuss errors that may arise in the measurement of diffracted beam intensities, in particular when area detector based instruments are employed for the measurements. In order to illustrate some of the points I will use the example of data measured from the same crystal of tetragonal hen eggwhite lysozyme on a CAD4 diffractometer¹ (C), a DIP100 image plate device² (D), a FAST area detector³ (F) and an X100A area detector⁴ (X). These data collections are summarized in Table 1. On the area detector based instruments the data collection is by a frame based rotation method. The (C) data set, although measured to a lower resolution, is, for reasons specified below, likely to contain few(er) systematic errors and is therefore used as a reference data set in this article. Finally included in Table 1 are R_{ref} from refinement⁵ against coordinates extracted from the Brookhaven Data Base. Given that the data were collected relatively quickly, the refinement is satisfactory in each case and the differences in final R_{ref} are marginal. Trends in final R_{ref} with resolution are shown in Table 2.

Systematic Errors

Systematic errors (when the mean value of a population of measurements deviates significantly from the assumed true value) will be subclassified according to the primary source of the error. In general we must either avoid these errors by good experimental technique or be able to independently measure the error and apply a correction to the data.

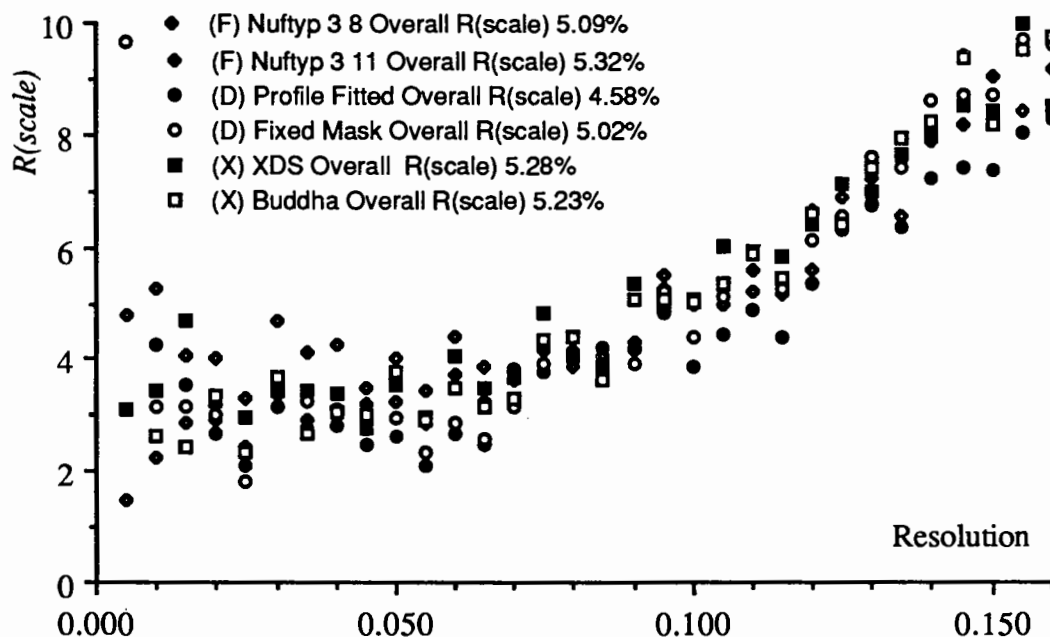


Fig.1 Variation of R_{scale} as a function of resolution for different data sets and processing methods. Resolution here and throughout is given as $4\sin^2\theta/\lambda^2$.

Table 1 Data Collections from a crystal (0.7x0.8x0.6mm) of Hen Eggwhite Lysozyme

	(C)	(D)	(F)	(X)
Instrument:				
X-ray Source: $\lambda=1.5419\text{\AA}$				
Generator Type	FR586 Sealed Tube	MX18 Rotating Anode	GX21 Rotating Anode	GX18 Rotating Anode
Focus	Fine	300 μ	300 μ	100 μ
Power	42kV 30mA	40kV 90mA	40kV 70mA	30kV 30mA
Beam Size at Crystal	1.0mm	0.6mm	0.6mm	0.6mm
Monochromation	β -filter	Graphite Crystal	Graphite Crystal	Graphite Crystal
Collection:				
Dates	30/11-5/12/89	12-14/12/89	27-29/11/89	8-9/12/89
Exposure Time	5100 mins	1500 mins	865 mins	725 mins
Data Collection Rate	1.7 reflins/min	27 reflins/min	45 reflins/min	49 reflins/min
Crystal-detector dist.	400mm	100mm	65mm	90mm
Detector swing angle	As required	0.0°	25.0°	22.0°
Frame size and time	---	1.0° for 1000s	0.1° for 30s	0.1° for 30s
Maximum resolution	2.5 \AA	2.04 \AA	2.0 \AA	1.90 \AA
Passes	1	1	2	1
Datum		ω 180 κ 0 ϕ 0	ω -122.97 ϕ 0	ϕ 90
Rotation axis/range		ϕ 0 to 90°	κ -134.74 ϕ 57.03	ω 0 to 90°
Av. symmetry equivs.	2.0	5.4	4.0	5.0
Radiation damage	3 standards used			
Absorption	Ψ scans, average transmission 86.5%.			
Processing:				
Observations	8828	Profile Fitted	NUFTYP 3 8	XDS
% obs. with $I < 3\sigma(I)$ #	3.8	40750	33248	45778
% obs. rejected	0.022	5.7	10.6	10.9
Unique reflections	4929	0.081	0.313	7.333
R_{merge} to θ_{max} .	-	7507	7899	9085
R_{merge} to 2.5 \AA	2.9%	5.7%	6.4%	8.1%
Completeness to 2.5 \AA	100%	-	-	-
to 2.1 \AA	-	95%	87%	97%
R_{ref} to 2.5 \AA	17.2%	17.1%	-	17.4%
R_{ref} to 2.0 \AA	-	19.0%	-	20.3%
# note that $\sigma(I)$ is unreliable for analog devices such as (D) and (F) because the gain must be measured or estimated				

Failing this, and as a general observation, repeated measurement of a reflection and its symmetry equivalents, preferably under different recording conditions, is a surprisingly good way of eliminating many systematic errors. The R_{scale} (as a function of resolution) after scaling of the three data sets (in each case processed in more than one way) against the reference, (C), data set is shown in Fig. 1. It will be evident that, after averaging, R_{scale} is very similar (Fig. 1), despite the larger differences in R_{merge} between the (D), (F) and (X) data sets (Table 1).

Table 2. R_{ref} as a function of resolution for (D), (X) and (F) data sets (see Table 1)

Resolution Range(Å)	20-5.52	5.52-3.76	3.76-3.17	3.17-2.84	2.84-2.61	2.61-2.44	2.44-2.31	2.31-2.20	2.20-2.11	2.11-2.03
$R_{ref}(\%)$ (D)	20	14	17	19	19	22	22	24	26	33
$R_{ref}(\%)$ (F)	20	14	17	20	19	22	23	24	25	29
$R_{ref}(\%)$ (X)	22	16	18	19	19	22	22	23	25	29

1) Xray Source

1.1) Fluctuation in incident beam intensity.

At both neutron and synchrotron facilities exposures are routinely measured relative to incident beam counts rather than time. In general this is not done on rotating anode sources. On such sources there is, however, a finite probability of a flashover. The result is either that the generator stays off and the data collection stops (as would be the case for the generator on which (D) sits) or ramps up again. For the (F) system when operating

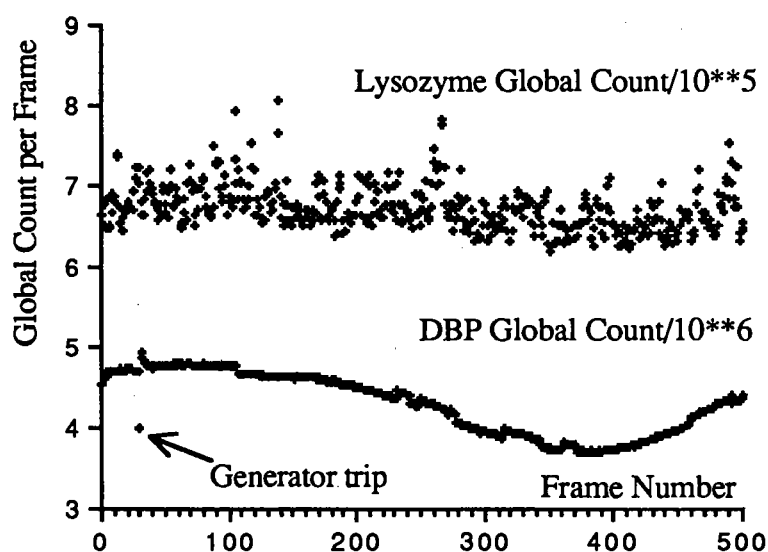


Fig.2 Variation of Global Count per Frame on (X). The top curve illustrates a case where strong Bragg scattering causes moderate frame to frame variations. The bottom curve for a more weakly diffracting crystal, shows a generator trip.

with a GX21 rotating anode generator the hardware-software interface currently allows a generator trip to be detected and the frame currently being collected to be remeasured. On the (X) instrument no such facility is available. For (X) the result, provided that the number of oscillations per frame collection period is large, is that a frame with lower scale factor than its neighbours is recorded. This scale factor can be estimated from the global Xray background, which is a slowly varying function of rotation angle, provided that the global count is not dominated by the Bragg peaks. Unless special precautions are taken to minimise the Xray background (see below) this is the case. Two illustrative examples are shown in Fig.2. With the generator running normally the

short time scale fluctuations of the Xray output (typically $< \pm 1\%$ on our generators) are averaged out over the frame collection period. Longer term drift is usually taken care of by the internal scaling of a data set (Section 4.2). The stability of the sealed tube generator is in all respects superior to the rotating anode generators and therefore these errors are smaller for the (C) data set.

1.2) Beam inhomogeneity

The beam on all our area detector systems is to some extent inhomogeneous, largely due to the pyrolytic graphite monochromators we use. The errors introduced are in effect similar to, but smaller than, those due to changes in illuminated volume (Section 2.3). This error is, to first order, absent in the (C) data set.

2) Sample

2.1) Radiation damage

Undoubtedly it is best to avoid radiation damage. This can be done by using shorter wavelengths or by having the sample held at low (*e.g.* liquid nitrogen) temperatures during data collection. We also find that a 'clean' beam often reduces radiation damage somewhat, which is why we use a monochromator rather than β filter. Standards are not efficiently measured on an area detector and therefore we rely on measurement of symmetry equivalent reflections to correct for radiation damage. As a rule of thumb we avoid using data beyond the point where the average scale factor has fallen to 60% of its initial value. This is not because data are uncorrectable beyond this point but simply because we know that damage does not give rise to the same decay rate for each reflection. In fact, for some reflections, intensities may initially increase with radiation damage, because the intensity depends not only on the long range order but also on the detailed nature of the radical reactions that take place in the crystal.

On our sources the use of lower wavelengths is limited by the availability of targets and by the detection response. (D) could be used with Mo radiation but (X) could not and (F) could only be used efficiently with a thicker phosphor layer (but this would worsen the point spread function (PSF)). Higher energies are also advantageous because absorption (for biological molecules) is reduced substantially. Even so there is an overwhelming disadvantage because the flux in the α_1/α_2 doublet, for a Mo target at the same generator power, is estimated to be a factor of about 4 down on that for a Cu target.

The low temperature techniques⁶ hold most promise. The reduction in radiation damage is a general, well established, feature. Thermal diffuse scattering (TDS) is also reduced. Unfortunately, unlike the case for small molecule crystallography, it is not possible to lower the temperature slowly from ambient unless cryoprotectants can be soaked into the crystal⁷. Slow cooling of a crystal grown from salt tends to shatter the crystal. In consequence a shock freezing technique is required. This usually has the undesirable effect of increasing the mosaicity of the crystal and thus reducing the effective resolution of the data obtainable. Putting it another way it reduces the signal to noise ratio at a given resolution by "smearing" the peak. We currently do not try to measure data routinely at low temperatures unless the crystal lasts less than about 8hrs on our Xray sources. Maybe when we have more experience our attitude will change. Indeed, there are cases where data collection at low temperature increases the resolution attainable⁸.

2.2) *Crystal slippage*

This is a very common problem, especially working with crystals grown from PEG. Obviously avoidance is the best policy. In general this will be aided by roughening the capillary surface, by, for example, coating the inside of the capillary with a thin layer of denatured egg albumin. Often motors (especially stepper motors as found on (D) and (X)) accentuate the problem as does excessive generator vibration (due to either anode drive or, more usually, vibration from the backing pump). The software running on (X) and (F) will tolerate some degree of slippage if it is slow however the radius of convergence for recovery, is in our experience, less than 0.3° . Larger discontinuous changes can be recovered after the event if frames have been stored and can be reprocessed. The larger frame size for (D) means that, for this instrument, the recovery after the event requires resources that are not available with a PC as frame collector.

2.3) *Absorption and illuminated volume effects*

These effects are sometimes not distinguished, both are a function of rotation angle but absorption is also a function of scattering angle. It is instructive to calculate the relative effects for a platy crystal of dimensions $0.4 \times 0.4 \times 0.1 \text{ mm}^3$ illuminated by a uniform beam of 0.3mm diameter. In extreme orientations the variation in illuminated volume gives rise to a 58.9% decrease in intensity, however for the same orientations true absorption for a 6.3 \AA reflection (by the crystal only) results in a relative 37% increase in intensity. The illuminated volume effect is dominant but can work in the opposite sense to the true absorption. By definition the illuminated volume effect is eliminated by bathing the crystal in the beam. When this is done accurate methods⁹ are available for calculating absorption corrections for a crystal with known dimensions and morphology, and indeed these methods may become more popular in protein crystallography when cryogenic techniques are more standard. The observation that protein crystals often have ill defined morphologies is a poor excuse because it is always possible to approximate a set of bounding planes. A calculated correction¹⁰ has also been coded for the case of a protein crystal in a capillary with mother liquor. Alternatively we can try to measure the absorption correction. This was done using ψ scans at χ near 90° ¹¹ for the lysozyme data set measured on (C) and is valid because the crystal was bathed in the beam and all reflections were measured in the same plane. The method does not, of course correct for the scattering angle dependence. Such a method or extensions thereof¹² are inefficiently measured on an area detector, and in any event require goniometric capabilities not available on (X) and (D). Where, as for (F), full sample orientation is possible, an absorption correction can be computed from measurements of transmission of the main beam¹³. A beam much smaller than the crystal is used and the main beam intensity as a function of two mutually perpendicular axes (each normal to the beam) is recorded. The method is only applicable where the sample is bathed in the beam during data collection. More often than not we have variations in the illuminated volume with rotation angle, in which case we prefer to use the semi-empirical corrections discussed in Section 4.2. It is clear from the size of this effect that time spent checking the crystal centring is time well spent.

2.4) *Thermal Diffuse scattering*

The effect will clearly vary from crystal to crystal. TDS features that do not lie under Bragg peaks can give rise to errors in the background estimate when a run averaged global background (rather than a background based on pixels adjacent in X, Y and ϕ) is used. However the major TDS contribution is under Bragg peaks and can be very large¹⁴. Fortunately, as a percentage of the Bragg intensity, it is strongly resolution dependent, being smaller at lower resolution. The major result will be artificially low thermal parameters in the refined model.

3) Detector

3.1) Deadtime

One form of deadtime is the overhead for the frame collector of getting the frame to the frame processor. For (D) the instrument deadtime is the plate erase time plus the plate readout time plus the transfer time to the frame processor. In total this is 5 mins, contrasted with around 10s for (X) or (F). Therefore on (D) a frame time of less than 5mins is inefficient because the deadtime would be more than 50% of the total experimental time. This restricts us to the use of larger frame widths and the result is in principle a poorer signal to noise ratio than for (F) or (X). The effect will depend on the absolute count in the peak and background but the general pattern follows the curve in Fig.3. For a typical

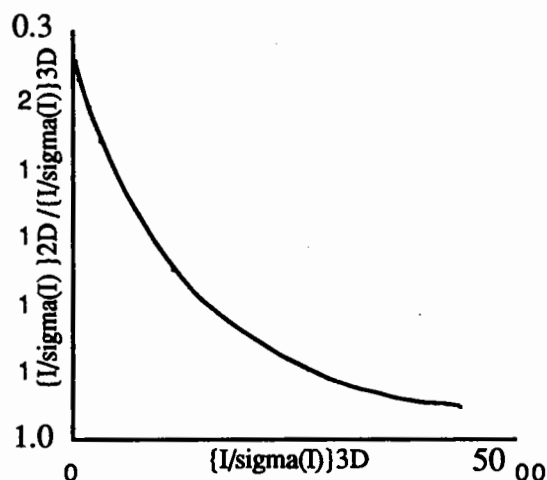


Fig.3 Better signal to noise ratio for 3D integration relative to 2D integration

case and assuming a frame size of around three times the rocking curve width the ratio $(I/\sigma(I))_{2D} / (I/\sigma(I))_{3D}$ in Fig.3 rises from around 0.3 when $(I/\sigma(I))_{3D} = 3$ to approaching 1.0 when $(I/\sigma(I))_{3D}$ integration is above 50. Despite this the results in Table 1 show no evidence of worse data quality for (D), probably because of the averaging over symmetry equivalents and because, for lysozyme to 2Å, the data are fairly strong anyway.

On (X), and indeed any single photon counter, there is a deadtime associated with each encoded event which means that the recorded event rate increases less rapidly than the true global count rate. Depending on the encoding method used the recorded count may actually decrease above a certain global count rate. In the case of (X), as in the lysozyme example, we may need to reduce the generator power to keep the global rate below 30kHz.

3.2) Linearity of response

Test measurements have never indicated any non-linearity below the local count rate limit of a particular area detector instrument. These are 300 counts/pixel/second for (F) and 225 counts/pixel/second at a 30kHz global rate for (X)¹⁵. For (D), which has two photomultipliers, one taking 1% of the photostimulated luminescence, the problem is more complex. Results seem to confirm that the setting of the relative gains gives an overall linear response to a precision better than crystallographic measurements allow us to check.

There is a possible problem with the reference (C) data set. If weak intensities are ignored we observe a general trend that the diffractometer data is underestimated at high intensity (see Figs. 8 and 9 below). This, being a consistent trend against all area detector data sets, tends to suggest that the diffractometer may have a measurable non-linear response but we have not checked this.

3.3) Uniformity of response with position

This is probably the most serious systematic error, specific to area detectors, that is introduced into the data. A calibration method for (F) has been described¹⁶ and one can verify that this is a good first order approximation over most of the active area of the detector. The variation in response over most of the active area of the detector is around $\pm 12\%$. However, because the form and size of the point spread function are not taken into account

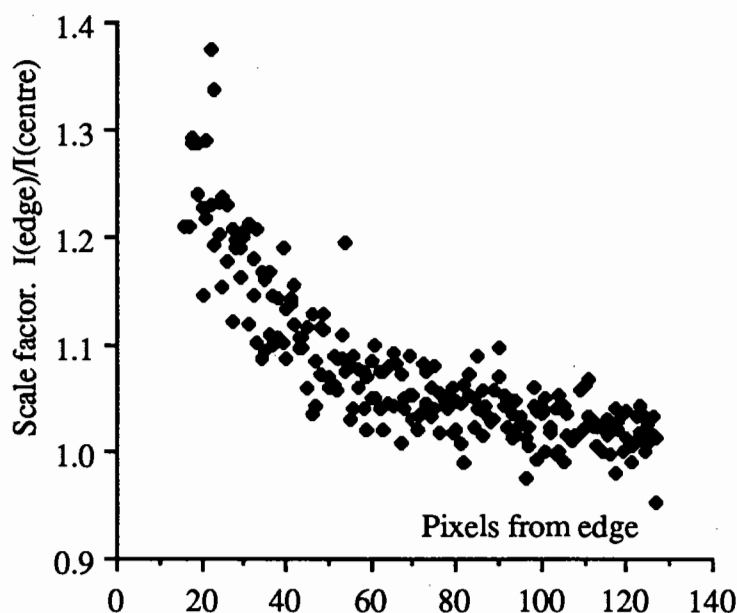


Fig.4 Scale factors for equivalent reflections as a function of position from the edge of (F)

when calculating the expected response to a flood field¹⁶, this response is overestimated at the detector edges and hence the correction factor is overestimated in these regions. The extent of the overestimation is shown in Fig.4 which derives from crystallographic experiments performed at different detector swing angles. We have applied this secondary correction and the consequent improvement in agreement with the reference data set is shown in Table 3. As expected the overall effect is small but the biggest differences are at high and low resolution because these resolution ranges contain most of the reflections close to the detector edges. For (X) there is no available software for calibration of response. The problem is a difficult one because it is not possible to distinguish differences in pixel size from variations in efficiency by means of a single flood field experiment. More thorough treatments to obtain a uniform response have been developed¹⁷. One can try to verify crystallographically how good the uniformity of response is and the results of such an experiment are shown in Fig.5.

Table 3. The effects of correcting for variations of response with position on the detector surface for (X) and (F).

	(F) (see Fig.4)		(X) (see Fig.5)	
	Before	After	Before	After
No. of observations	31811	31811	38969	38969
No. of rejects	113	112	34	36
R_{merge}	6.42	6.26	8.37	8.30
R_{merge} (1000-10Å)	4.77	3.94		
R_{merge} (2.09-2.04Å)	12.16	10.78		
R_{merge} (1st pass)	5.49	5.49	4.67	4.42
R_{merge} (2nd pass)	4.39	4.39	5.88	5.91

The effects of applying this correction are shown in Table 3. Again the effect is small but real. For (D) we have not yet performed the same sort of crystallographic experiment. Measurement of a given source for a given time at different points on the detector face has simply shown that variations of response with time are larger than

with position. The instrument is calibrated by attaching a ^{55}Fe source to the read head and scanning it across the plate several times before reading out the accumulated counts. This is used to apply a radial correction factor to the data. The (D) data set is scaled as a function of resolution in Fig.6 against the (C) data set. This clearly indicates that the radial correction is inadequate.

					0.3	0.4	0.4	0.6	0.7	1.4	0.0							
					0.6	0.6	0.6	1.0	1.0	2.2	2.2	1.9	1.9	1.4				
					0.7	0.9	0.9	1.0	1.4	1.4	1.8	2.4	2.0	2.2	2.1	1.9		
					1.6	0.9	1.1	1.0	1.0	1.7	2.3	2.1	2.2	1.7	2.1	1.6	1.6	0.5
0.3	1.4	0.9	0.9	0.9	0.8	1.9	0.8	1.6	0.8	0.7	0.8	0.6	1.0	0.6	0.6	0.0	0.0	
0.7	1.1	0.8	0.7	1.4	0.7	-0.1	0.7	0.4	0.9	0.3	-0.6	0.0	-0.4	-1.2	-1.2			
0.6	1.0	1.0	0.6	0.4	0.6	-0.3	-0.7	-0.8	-1.3	-1.2	-2.6	-3.1	-2.2	-2.0	-0.6			
-0.2	0.0	-0.1	-0.3	-0.4	-0.6	-0.8	-1.8	-2.7	-2.9	-3.0	-4.0	-3.6	-3.5	-2.4	-0.3			
0.2	-0.3	0.7	-1.0	-0.7	-0.3	-1.1	-1.4	-1.4	-1.5	-2.7	-3.1	-3.0	-2.7	-1.9	-0.1			
0.2	1.4	0.5	0.4	0.7	0.8	-0.3	-0.3	-0.7	-1.7	-1.5	-2.1	-2.2	-2.5	-1.7	-0.1			
1.2	2.8	4.1	2.9	1.6	2.5	0.7	-0.3	0.0	-0.7	-0.9	-1.7	-1.2	-0.5	-0.3	0.2			
0.0	4.5	3.9	2.8	2.5	1.7	0.1	0.0	0.1	-0.6	-0.8	-0.7	-0.6	-0.4	0.0	0.1			
	1.5	2.8	2.1	0.9	0.2	-0.1	-0.4	-0.5	-0.8	-0.9	-0.8	-0.5	-0.3	0.0				
		0.9	1.4	0.5	0.2	0.1	0.0	-0.2	-0.5	-0.5	-0.4	-0.4	-0.2					
			0.1	0.4	0.1	0.0	-0.3	-0.3	-0.4	-0.4	-0.3	-0.1						
					0.0	0.0	0.0	0.0	0.0	0.0								

Fig.5 Crystallographically measured response variations over the surface of (X). Values are the % deviation from an arbitrary reference with the detector area described as a 16x16 array.

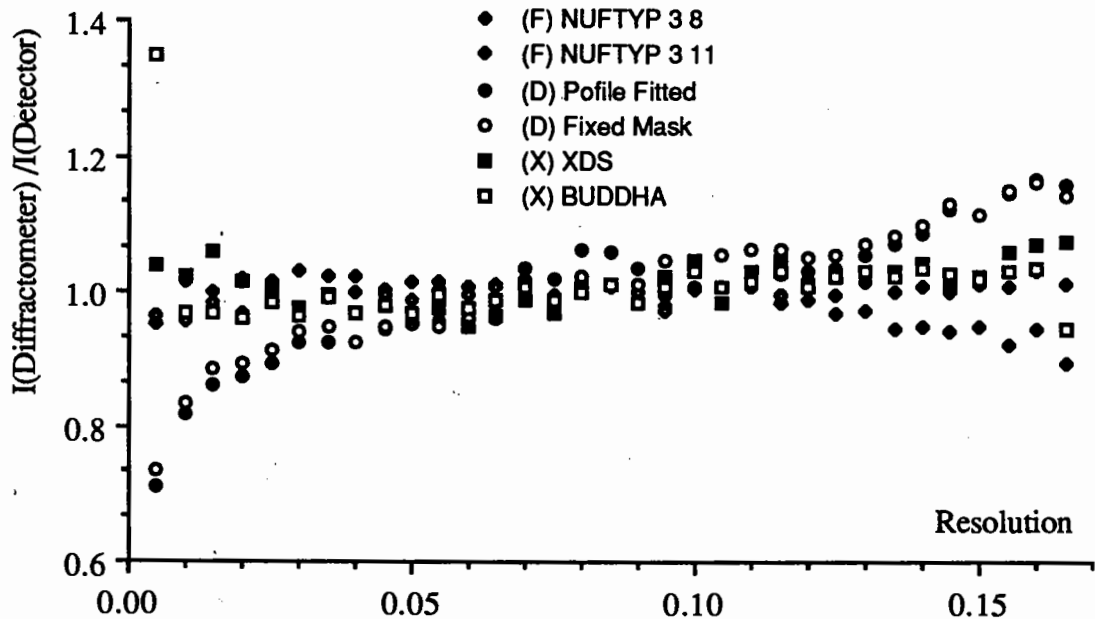


Fig.6 Scaling of (C) data set against area detector data sets as a function of resolution.

3.4) Uniformity of response with time

For (D) the latent image on the plate decays with time (an example is shown in Fig. 7) and we cannot be absolutely certain that it does not decay during the readout period. The plate is read from outside to centre (high to low resolution) but scaling against the reference (C) data set shows the high resolution data to be relatively underestimated. Consequently this error cannot result from decay of the latent image during readout. However decay of the latent image during the longer time period calibration experiment would result in an overcorrection being applied towards low resolution. We are currently working on a better calibration method.

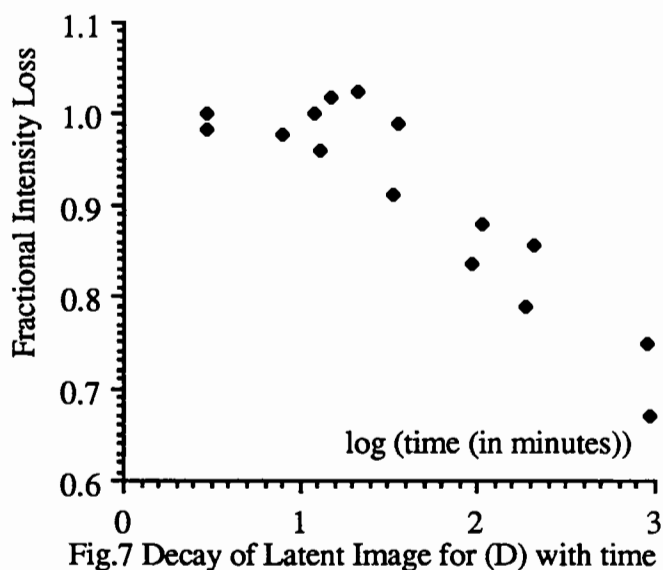


Fig.7 Decay of Latent Image for (D) with time

4) Software

4.1) Algorithms to minimise random errors

It is well known that profile fitting¹⁸, a method of intensity calculation designed to minimise random errors, introduces systematic errors into the data, especially for strong (sharp) reflections. Additional systematic errors can arise from methods of background estimation or from integration mask positioning. Therefore it is of interest to compare various software treatments in the sense of how well they scale against our reference (C) data set. For (X) data processed with the XDS¹⁹ and BUDDHA²⁰ packages are scaled against the (C) data set as a function of intensity in Fig 8. (F) data were processed using the program MADNES²¹ with two available NUFTYPs and are also compared in Fig.8. Finally included on the plot are (D) data integrated either with a fixed mask or by use of a profile fitting algorithm²². The problems are all at low intensity. There is no obvious systematic error in the scaling of strong intensities in the profile analysis¹⁹ used either on (X) data with XDS or (F) data with the PROCOR postprocessing program. More surprisingly no error is evident with the (D) data either; perhaps because the profile fitting method used attempts a form of interpolation when fitting the profile. The BUDDHA (Version 20-05) package tends to overestimate weak reflections because the integration box is allowed to move and there is a tendency to hunt for noise. NUFTYP 3 8 also tends to overestimate weak intensities but in this case it is because the criteria for choosing pixels to be used for the background estimation are biased against high background pixels. The background is therefore underestimated and the intensity overestimated.

There is also a major effect in the (C) data processing. The ordinate analysis²³ used in processing the (C) data with the SDP package²⁴ will have the effect of overestimating the weak data and could be better done^{25,26}. The result, shown in Fig.9 is that when scaling the (C) against the area detector data sets the diffractometer data is overestimated at low intensity. A second reason for this effect is discussed in Section 4.3 below.

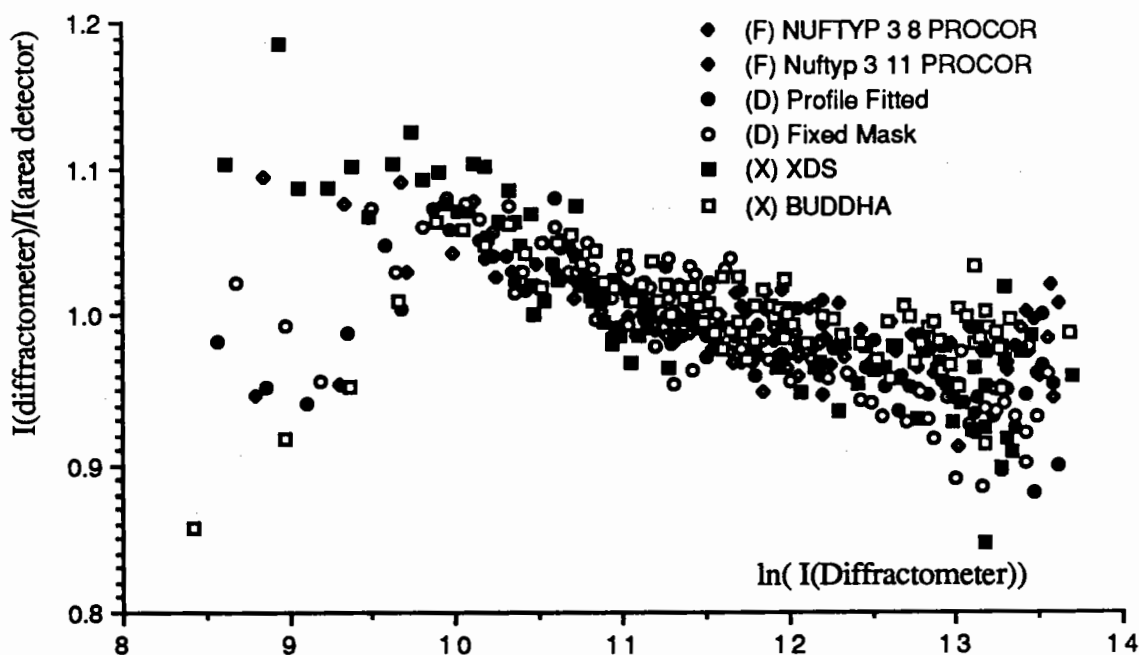


Fig.8 Scaling of (C) data set against area detector data sets as a function of intensity

4.2) Internal scaling

The most common method²⁷ of internally scaling a data set is to divide it up into a set of batches for each of which a scale factor and temperature factor may be calculated by a least-squares method using symmetry equivalent reflections. The corrections thus made are assumed to be for illuminated volume, radiation damage and absorption. How poorly they actually correct for these effects is illustrated by the (X) and (F) data sets. Examination of Table 3 shows the internal consistency for each pass separately to be far better than the internal consistency of the combined data.

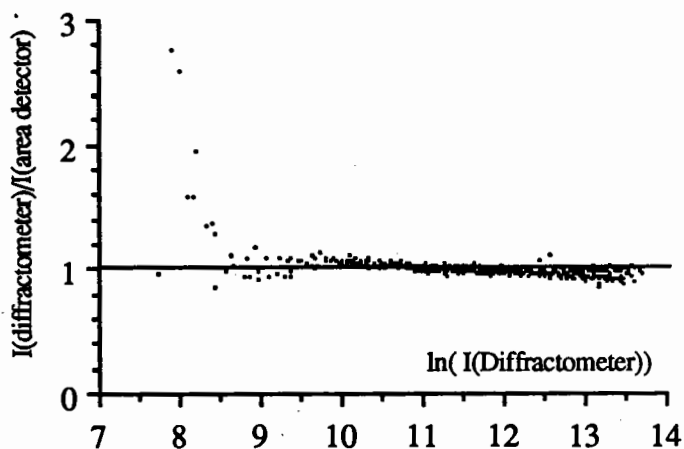


Fig.9 Scaling of diffractometer data against area detector data over the whole intensity range

Table 4. Comparison of different internal scaling methods for (X) and (F) data

	(X)			(F)		
	Batch	Continuous	Function of	Batch	Continuous	Function of
Parameters	Scaling	$\omega^{\#}$	Beam u, v, w	Scaling	$\omega^{\#}$	Beam u, v, w
Parameters	126	(162)	6	56	(135)	6
Observations	42371	42371	42371	31191	31181	31181
Rejects	34	0	219	113	113	119
R_{merge}	8.37	8.15	7.29	6.42	6.28	6.03

also with a detector position dependence, but each pass treated separately. See Ref. 19

The scaling can also be treated as a continuous function of rotation angle with some additional function of position on the detector surface added. Such a treatment is used by XDS and PROCOR¹⁹ and the improvement, indicated in Table 4, is relatively small but real. Scaling against a reference data set is also common practice^{28,19} especially for derivative data sets. We do not use these methods, largely because of the danger of perpetuating any systematic error in the data set against which the scaling has been done. More interesting methods are described by Katayama²⁹ and Takusagawa³⁰. We have not tried the former but for the latter the very simple functionality (equivalent to an ellipsoidal absorption surface) in terms of the direction cosines of the incident and diffracted beams in a coordinate system tied to the crystal, has a very real effect when applied to the (X) and (F) data sets. Better internal consistency is achieved with fewer parameters (Table 4). Clearly an extension of the method to a more complex absorption surface holds promise. R_{scale} before and after the application of corrections mentioned for (X) and (F) is shown, together with the expected curve based on Poisson statistics, in Figs. 10 and 11.

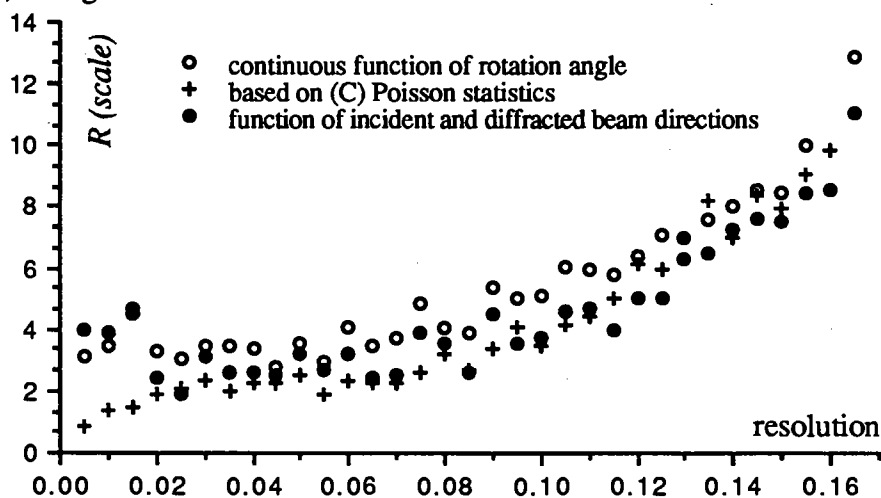


Fig. 10 . R_{scale} as a function of resolution for (X) using different scaling methods. Also shown is the expected R_{scale} if the only errors are counting errors in the (C) data set.

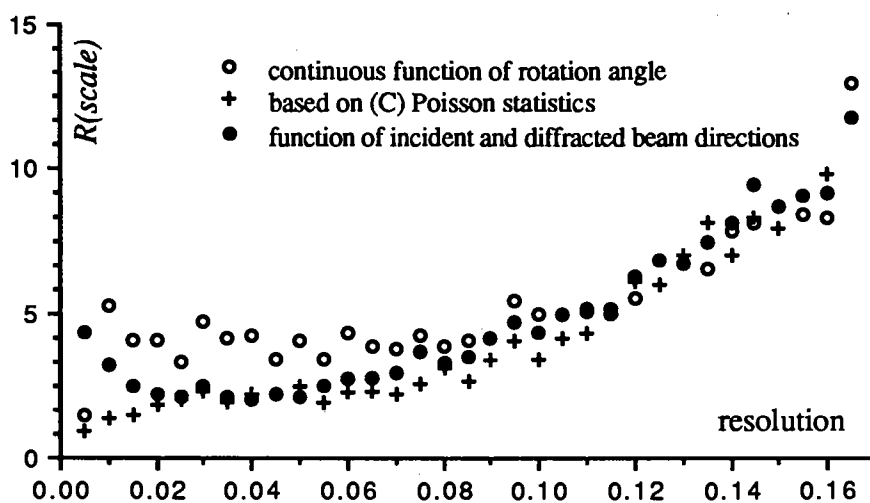


Fig.11 R_{scale} as a function of resolution for (F) using different scaling methods. Also shown is the expected R_{scale} if the only errors are counting errors in the (C) data set.

4.3) Treatment of negative intensities

The effects under this head are widespread and arise from a counterproductive desire to work with F rather than F^2 . As previously discussed (Fig. 9), when data from (X), (F) and (D) are scaled against the reference data set (C), the diffractometer data is grossly overestimated at the lowest intensities. One reason for this has already been discussed above. There is however a second reason in that the (C) data set, despite the longer collection time and lower resolution, has (in general) the lower $I/\sigma(I)$. After averaging within a data set there are still negative and zero intensities which are currently rejected in the software we use. The result will be (see Fig. 9) that the broader distribution (in this case (C)) has the scale shifted to higher values at low intensity. There is at least one effective method of avoiding this bias³¹ and its use is highly recommended (and incorporated in the CCP4 program TRUNCATE).

5) Completeness of data

The importance of data completeness is not often appreciated. A recent report concerning the effect on the success of a rotation function search³² is unlikely to be unique and, although a rotation/translation search may be especially susceptible to the absence of regions of reciprocal space, one might also expect map quality and the ease of map fitting to be affected. For (F) one can set a crystallographic axis along the rotation axis and use classical³³ methods to determine a data collection strategy. The programs RSPACE³⁴ and LATTICEPATCH³⁵ that run on graphics devices allow one to explore visually the best strategy for data collection. Programs giving a simple guide to strategy are also available for (F) and (X). In general though there is a lack of appropriate software for determining data collection strategy on area detector based instruments. Different goals may be required in addition to achieving data completeness; for example we might wish to minimise the data collection time (if, for example, the crystal suffers rapid radiation damage) or we might wish to optimise the number and relationship of repeated measurements to allow the best scaling and elimination of systematic errors (our preferred strategy).

6) Unit cell dimension errors

Accurate determination of unit cell dimensions is also essential for accurate structure determination. A 15° precession photo at 100mm will, with care, give an accuracy in cell parameters of around $\pm 0.1\%$. It is difficult to do much better than this on any of (C), (D), (F) or (X). The lysozyme cell dimensions determined on the four instruments are given, together with the literature values, in Table 5. These are reasonably consistent, however, we have found (at least with the XENGEN³⁶ and XDS software) on the (X) instrument that the determined cell parameters are unreliable. This is not simply due to correlation between cell parameters and crystal to detector distance, but rather results from errors in the spatial distortion calibration being carried, via the refinement against observed spot positions, into the cell parameters.

Table 5. Cell dimensions for the lysozyme crystal as determined on different instruments.

	(C)	(D)	(F)	(X)		Literature ³⁷
a (Å)	79.23	79.12	79.19	XDS 79.24	BUDDHA 79.21	79.1
c (Å)	38.09	38.01	37.99	38.02	38.03	37.9

Random Errors

1).Signal to noise ratio

1.1) Background level and estimation

Clearly reduction of background is of immense importance. With an "as delivered" setup on (X) some 30% of the background originates from air scatter of the main beam. Reducing the pathlength of the main beam in air is worthwhile and simple to do by bringing the collimator exit aperture and the beam stop as close as is feasible to the crystal. Matching the collimator size to the crystal size is also advantageous (but has deliberately not been done for the lysozyme data used in this article). At large crystal to detector distances a He flight path is worthwhile since it reduces absorption of the scattered beams. As a rule of thumb 1% increase in signal per centimeter of flight path can be gained. Fairly obviously it is preferable to mount the crystal as dry as can be tolerated. These observations apply equally well to (D) and (F), the latter being the only instrument we have not yet applied the suggested modifications to (although they have been applied by others³⁸).

In the case of (X) where we occasionally have the situation where we turn down the generator power to keep the global count rate below 30kHz there is good reason to think that the lost flux would be better employed in the improvement of beam properties ,for example to use mirrors rather than a monochromator to obtain a more parallel beam. We have not yet made any comparison of data sets collected with different incident beam optics.

1.2) Choice of frame size

We have found that ,for both (X) and (F), the choice of 0.2° as opposed to 0.1° frame size has little effect on data quality. This is perhaps not too surprising because the "effective mosaicity" is usually determined by the ,larger in our case, collimator cross fire. However we usually use a 0.1° frame size for (F) because these frames also have an appreciable DC level and we wish to avoid wrap around for the stronger reflections.

Acknowledgments

I thank my colleagues at EMBL for their tolerance of my use of instrument time. My special thanks to Dietrich Suck and Armin Lahm for helpful discussions.

References

$$R_{scale} = \frac{\sum_{hkl} \left| |F^{hkl}| - |F_{reference}^{hkl}| \right|}{\sum_{hkl} |F_{reference}^{hkl}|} \quad R_{ref} = \frac{\sum_{hkl} \left| |F_{obs}^{hkl}| - |F_{calc}^{hkl}| \right|}{\sum_{hkl} |F_{calc}^{hkl}|}$$
$$R_{merge} = \frac{\sum_{hkl} \sum_n |F_{hkl,n}^2 - \langle F_{hkl}^2 \rangle|}{\sum_{hkl} n \langle F_{hkl}^2 \rangle}$$

- 1 Manufacturer:- BV Enraf-Nonius,Delft,The Netherlands: J.Schagen, L.Straver, F.van Meurs and G. Williams (1988),CAD4 Operators Guide.
- 2 Manufacturer:- MAC Science Co.Ltd.,Tokyo,Japan: I. Tanaka, M.Yao, M.Suzuki, K.Hikichi, T.Matsumoto, M.Kozasa and C. Katayama (1990) *J.Appl.Cryst.* (in press).
- 3 Manufacturer:- BV Enraf-Nonius,Delft,The Netherlands: U.W.Arndt and D.J.Thomas (1982) *Nucl. Instr. Meth.*, 201, 21-25; U.W.Arndt and G.A.in't Veld (1988), *Adv. Electron Phys.*, 74 ,285.

- 4 Manufacturer:- Siemens International, Madison,U.S.A.: R.M.Durbin, R.Burns, J.Moulai, P.Metcalf, D.Freymann, M.Blum, J.E.Anderson, S.C.Harrison and D.C.Wiley (1987), *Science*, 232, 1127-1132.
- 5 Refinement using TNT (D.E.Tronrud, L.F.Ten Eyck and B.W.Matthews (1987) *Acta Cryst.*, A43, 489-501). Coordinates taken blindly from entry 6LYZ of Brookhaven Data Base (F.C.Bernstein, T.F.Koetzle, G.J.B.Williams, E.F.Meyer Jr., M.D.Brice, J.R.Rodgers, O.Kennard, T.Shimanouchi, and M.Tasumi (1977) *J.Mol.Biol.*, 112, 535-542) omitting C_γO_{δ1},O_{δ2} of Asp101 and O91 which had bad contacts. In all cases refinement terminated with rms errors less than 0.02Å in bond length, 4.0° in bond angle, 0.02Å in non-planarity of planar groups and 0.05Å in non-bonded contacts.
- 6 H.Hope (1988) *Acta Cryst.*, B44, 22-26.
- 7 G.Petsko (1975) *J.Mol.Biol.* 96, 381-392.
- 8 M.M.Teeter and H.Hope (1986) *Computer Simulation of Chemical and Biomolecular Systems*. Ann. of New York Acad. Science ,482,163-165.
- 9 J.de Meulenaer and H.Tompa (1965) *Acta Cryst.*, 19, 1014-1018.
- 10 G.T.DeTitta (1985) *J.Appl. Cryst.*, 18, 75-79.
- 11 A.C.T.North, D.C.Phillips and F.S.Mathews (1968) *Acta Cryst.* A24, 351-359.
- 12 G.Kopfmann and R.Huber (1968), *Acta Cryst.*, A24, 348-351; R.Huber and G.Kopfmann (1969) *Acta Cryst.*, A25, 143-152.
- 13 MADNES Users Guide(1989) J.W.Pflugrath and A.Messerschmidt. B.V.Enraf-Nonius
- 14 see for example P.A.Kroon and A.Vos(1979) *Acta Cryst.* A35, 675-684.
- 15 P.A.Tucker. *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography*. Daresbury Laboratory (1989) No.24 p47.
- 16 D.J.Thomas(1989) *Proc. Roy. Soc.. Lond.* A425, 129-167 ; D.J.Thomas (1990) *Proc. Roy. Soc. Lond.* A428, 181-214.
- 17 K.C.Holmes, personal communication.
- 18 M.G.Rossmann (1979) *J.Appl.Cryst.* ,12, 225-238.
- 19 W.Kabsch (1988) *J.Appl.Cryst.*, 21, 916-924.
- 20 M.Blum,P.Metcalf,S.C.Harrison and D.C.Wiley (1987) *J.Appl.Cryst.*, 20, 235-242.
- 21 A.Messerschmidt and J.W.Pflugrath (1987) *J.Appl.Cryst.*, 20, 306-315.
- 22 P.A.Tucker (1989) Program for DIP100 Data Postprocessing.
- 23 M.S.Lehmann and F.K.Larsen (1974) *Acta Cryst*, A30, 580-584.
- 24 SDP Users Guide (1985) B.A.Frenz and Associates Inc., College Station, Texas 77840 and Enraf Nonius Delft.
- 25 I.J.Tickle (1975) *Acta Cryst.* B31, 329-331.
- 26 S.Oatley and S.French (1982) *Acta Cryst.*, A38, 537-549.
- 27 C.G.Fox and K.C.Holmes (1966) *Acta Cryst.*, 20, 886-891.
- 28 D.Stuart and N.Walker (1979) *Acta Cryst.*, A35,925-933 ; C.E.Schutt and P.R.Evans (1985) *Acta Cryst.*, A41, 568-570.
- 29 C.Katayama (1986) *Acta Cryst.*, A42 19-23.
- 30 F.Takusagawa (1987) *J.Appl.Cryst.* ,20, 243-245.
- 31 S.French and K.Wilson (1978) *Acta Cryst.*, A34, 517-525.
- 32 J.Kallen and R.Pauptit. *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography*. Daresbury Laboratory (1989) No.24 p63.
- 33 S.V.Nikonov and Y.N.Chirgadze (1985) *Acta Cryst.*, A41, 195-203.
- 34 M.R.Harris,M.Fitzgibbon and F.Hage(1988). *A.C.A. Abstracts.Philadelphia Meeting* p51.
- 35 A.L.Klinger and R.H.Kretsinger(1988) *A.C.A. Abstracts.Philadelphia Meeting* p80.
- 36 A.J.Howard, G.L.Gilligand, B.C.Finzel, J.L.Poulos, D.H.Ohlendorf and F.R.Salemne (1987), *J. Appl. Cryst.*, 20, 383- 387.
- 37 C.C.F.Blake, L.N.Johnson, G.A.Mair, A.C.T.North, D.C.Phillips,and V.R.Sarma (1967) *Proc. Roy..Soc. Lond., Ser B* ,167,365-377 ; M.Levitt (1974) *J.Mol.Biol.* , 82, 393-420; R.Diamond (1974) *J.Mol.Biol*, 82, 371-391.
- 38 U.Arndt, private communication.

*Notes on the errors of phase determination in the Multiple
Isomorphous Replacement Method and the Molecular
Replacement Method.*

by
Zygmunt Derewenda

Department of Chemistry, University of York, Heslington,
York, YO1 5DD U.K.

The success of any X-ray structural work depends critically on the determination of the initial set of phases, which are then used (via the usual route of a Fourier synthesis) to establish a starting set of atomic coordinates. The coordinates are almost routinely refined by one of numerous least-squares methods, unless a simple difference Fourier map is sufficient to obtain useful information. If a new, hitherto unknown structure is studied, the isomorphous replacement method in any of its existing variations (multiple or single replacement, with or without the use of the anomalous scattering contribution) continues to be the most successful tool. However, the ever increasing library of structures determined and refined to a near atomic resolution allows one to make extensive use of the molecular replacement method (MR), whereby the phase angle estimates are calculated using an approximate atomic model placed in the unit cell with various rotation and translation function programs. The treatment of errors in the Isomorphous Replacement method had been dealt with at the outset (see Blow & Crick (1959); a detailed account can also be found in Blundell & Johnson(1976). This was possible, because the general statistical theories dealing with random errors are applicable. The MR method, on the other hand, invariably produces systematic errors with random component limited only to the errors in the measured intensities. This phenomenon makes any error analysis very difficult. In this short paper, I will discuss some experiences with the use of both methods. As an illustration I will use some case studies selected from the spectrum of projects completed, or currently under investigation, in our group.

The cases:

Aspergillus niger acid α -amylase: this is a single polypeptide chain (471 residues) enzyme of a molecular weight of ca. 45kD; it forms good quality crystals, space group C222₁ with axial lengths a=80.1, b=98.3, c=138.0A; the solvent content is less than 50%. It has been solved by combined Isomorphous and Molecular Replacement (Boel et al., 1990, Brady et al., 1990b) and is particularly suitable for a comparison of the two methods.

Mucor miehei triacylglyceride lipase: a single chain (269 residues) enzyme (mw ca. 27kD); good quality crystals exhibiting orthorhombic symmetry (space group P2₁2₁2₁, a=71.6, b=75.0, c=55.0A); the solvent content is also within the expected range of ca. 50%. The structure was solved by Isomorphous Replacement, with the aid of molecular dynamics refinement of a partial structure

(Brady et al., 1990a). The knowledge of the overall fold from the studies of a related lipase (see below) was of help.

***Humicola* sp. lipase:** a related enzyme of almost identical molecular weight, but of limited sequence homology to the *M.miehei* lipase. Crystals are hexagonal ($P6_1$, $a=b=143.0$, $c=81.0\text{\AA}$) with two molecules in the asymmetric unit, and very fragile owing to high solvent content (80%). So far they have resisted high resolution data collection attempts using a synchrotron source due to their extreme radiation damage. They survive well on a conventional source, allowing one to collect 3.2 - 3.5 \AA resolution data from one crystal. The structure has now been solved (unpublished data).

Low pH porcine insulin: this is a dimeric form of insulin. It forms very good quality orthorhombic crystals (space group $P2_12_12_1$, $a=58.0$, $b=51.5$, $c=38.0\text{\AA}$) diffracting to high resolution. The structure was solved by Molecular Replacement using the known 2Zn insulin structure, albeit with considerable difficulties (Derewenda, U. 1990), some of which are documented in this paper.

For additional crystallographic details regarding data collection and the handling of heavy atom derivatives the reader is advised to refer to the original papers.

***The distribution of error in the phase angles
determined by Multiple Isomorphous Replacement.***

The theory of phase determination via Isomorphous Replacement has been widely documented, and there is no need to duplicate it in this short paper. The notation used here follows that originally introduced by Blow and Crick (1959), and expanded by Blundell and Johnson (1976).

When a structure is solved and refined it is instructive to look back at the original phasing experiment to establish the nature of the errors. Hitherto this was not common practice, partly due to computational costs (this is no longer a real obstacle), but largely to one's reluctance to pursue a seemingly futile goal of tracing back the steps which have already led to a much more interesting structural problem. It is therefore an accepted practice to list the mean figure of merit $\langle m \rangle$, and various R factors relating to the accuracy of heavy atom refinement as a measure of the accuracy of the original phase determination. In this paper I will show analyses based upon a direct comparison of the various phase sets with the final calculated phases based on refined structures. Although there are a number of possible representations, I will focus mainly on the global distribution of phase error. Figure [1] illustrates a typical phase error

distribution for a protein solved by the MIR method. The graphs show the percentage of acentric data with a given phase error, using 10° intervals. The distribution generally follows the Gaussian distribution of random errors, with one exception: it is slightly bimodal, with two equivalent peaks app. at -15° and $+15^\circ$. This can probably be explained by the use of the best phase rather than the most probable phase in all the calculations; a histogram

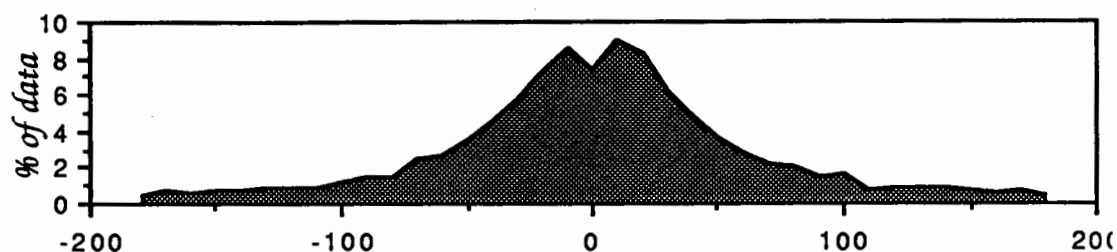


Fig. 1 A phase error distribution diagram (% of acentric terms vs. the phase error against the final phases calculated from atomic coordinates) for mucolipase.

using the most probable phases (not shown here) is unimodal with the maximum centered on 0° . It is worth noting, that the analyses shown here are based only on fully phased reflections, i.e. reflections for which intensities from all available derivatives in addition to the native crystal have been measured.

While the distribution of the phasing errors may not be surprising *per se*, it would be of a great advantage to predict it as accurately as possible in order to apply an appropriate weighting scheme. The latter is commonly based upon the figure of merit m , which in theory is a cosine of the phase error $\Delta\alpha$. In reality, it is well established, that $\langle m \rangle$ reflects primarily the sharpness of the probability distribution and may be misleading. However, the use of m leads in principle to the best Fourier which is calculated using $m |F_{\text{obs}}| \exp i\alpha_{\text{mir}}$ as coefficients. This reduces the contribution of the error vector F_e , which for simplicity we will define here as $|F| \exp i\alpha_{\text{mir}} - F_p$, where F_p is the true protein structure factor, while $|F_e| = 2\sin(\Delta\alpha/2)$. In this notation F_{best} (or $m |F_{\text{obs}}| \exp i\alpha_{\text{mir}}$) is the sum of $m F_p$ and $m F_e$. The electron density map generated using $m F_p$ as Fourier coefficients will (in theory) generate the true structure with peak heights of $\langle m \rangle^2$ relative to the true structure, while $m F_e$ will give rise to (hopefully random) noise. Fig. [2] illustrates the distribution of $\langle m \rangle$ vs. phasing error. It is clear, that while in general the distribution follows that of the phasing error, the poorly phased reflections are grossly

overestimated with well phased reflections underestimated by low $\langle m \rangle$. It is known, that the values of m are dependent upon the estimates of the standard deviations σF . Fig. [3] shows two plots of $\langle m \rangle$, one for the strong ($F > 100 \sigma F$), and one for the weak ($F < 100 \sigma F$) reflections. It shows that the estimation of m is intensity

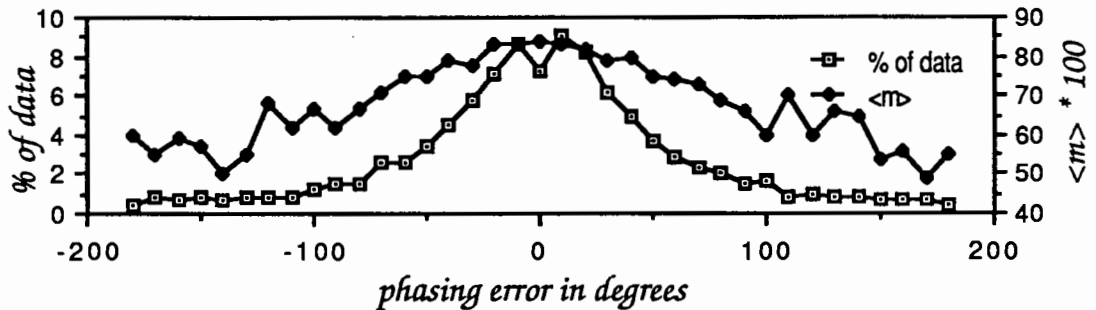


Fig. 2 The distribution of the mean figure of merit $\langle m \rangle$ against the actual MIR phase error; the calculation refers to all fully phased acentric data of mucolipase.

dependent, leading to higher values of m for stronger terms (it is noteworthy, that a plot of $\langle |F| \rangle$ vs. phasing error - not shown here - shows no dependence of $\Delta\alpha$ on $|F|$, rather contrary to a commonly held view). The explanation is likely to be in the estimates of σF , which are used to determine the values of E 's.

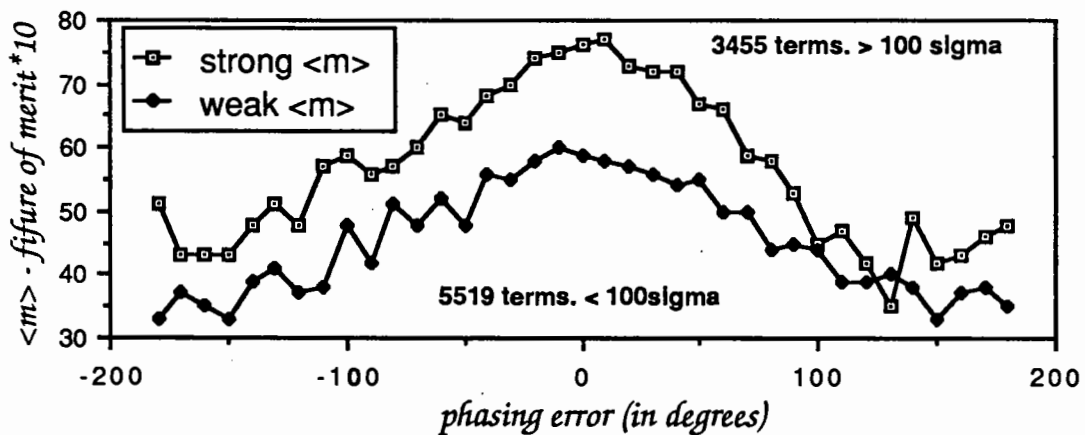


Fig. 3. The analysis of mean figure of merit ($\langle m \rangle$) values for strong and weak reflections separately, against the actual phasing error; the data shown refers to the acid amylase fully phased acentric data.

Various data collection programs developed for a spectrum of instruments (area detectors, imaging plate, film, diffractometer) include different algorithms for the evaluation of standard

deviations, with some requiring empirically determined machine constants. A detailed analysis of standard deviations is routinely carried out at the end of the program AGROVATA (CCP4 suite), and it allows to correct the standard deviations using real error distribution to calculate the appropriate scaling factor k . With data collected using the imaging plate system at the EMBL synchrotron station in Hamburg) we found it necessary to inflate the standard deviations by a factor (k) of 1.4. At the same time, no variation of k vs. $\langle I \rangle$ can be seen. A similar analysis of data collected using the Xentronics/Siemens area detector shows a linear dependence of k on $\langle I \rangle$ with k ranging from 1.0 for the weak reflections to app. 4.0 for the strongest terms.

In their original paper Blow and Crick (1959) point out that complicated weighting schemes usually have little impact on the appearance of the Fourier. Unfortunately the judgement is highly subjective and rests entirely upon one's ability to interpret the troughs and ridges of the final electron density map. Fig. [4] illustrates this point vividly. While there appears to be little difference between the experimentally weighted Fourier and the ideally weighted one (see Fig. [4] for full details), the break in main chain connectivity replaced by an erroneous connection between the sidechain of Trp and the mainchain of Tyr may make all the difference during the initial tracing of the structure.

Single Isomorphous Replacement.

During our investigation of the α -amylase structure it became apparent that the Pb derivative was by far superior (with a phasing power of 6 in the medium resolution range) owing to its almost perfect isomorphism. This effect is probably largely due to the fact that Pb replaced another metal (Ca) in metal binding pockets, while Hg was bound to a semi buried Cys residue, what could have caused structural perturbations. I have used the Pb derivative to calculate SIR (single isomorphous replacement) phases and compare them with the final calculated ones. Fig. [5] illustrates this comparison and also shows the extent of the superiority of the MIR phasing. As one would expect, the greatest improvement is achieved for reflections with an absolute error of approximately 90° , since the additional derivatives resolve the bimodality of phase probability.

MIR and SIR phase modification by solvent flattening.

Density modification methods have been in use for some time. B.C. Wang (1986) described the way to resolve phase ambiguity in the SIR method using a back-transformed solvent-flattened electron density map to calculate a set of phases, subsequently combined with the experimental ones. This technique has since gained many

best Fourier - values of m obtained in the normal way



best Fourier with computed ideal values of m

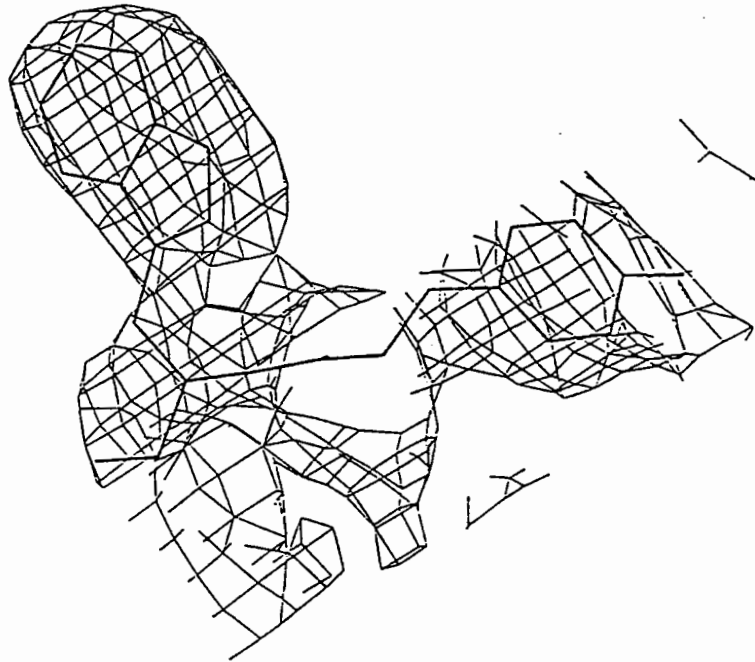


Fig. 4. Two MIR electron density maps calculated using different weights: a) experimentally determined figures of merit (all reflections used); b) theoretically correct weights calculated for each reflection as $\text{acos}(\Delta\alpha)$, where the phase error has been evaluated by comparison to the final α_{calc} . Reflections with phase error greater than 90° have been assigned a figure of merit equal to zero.

followers, and was applied in both SIR and MIR cases to enhance the accuracy of the phases. While it is recognized that the technique works best with crystals containing high percentage of solvent, the

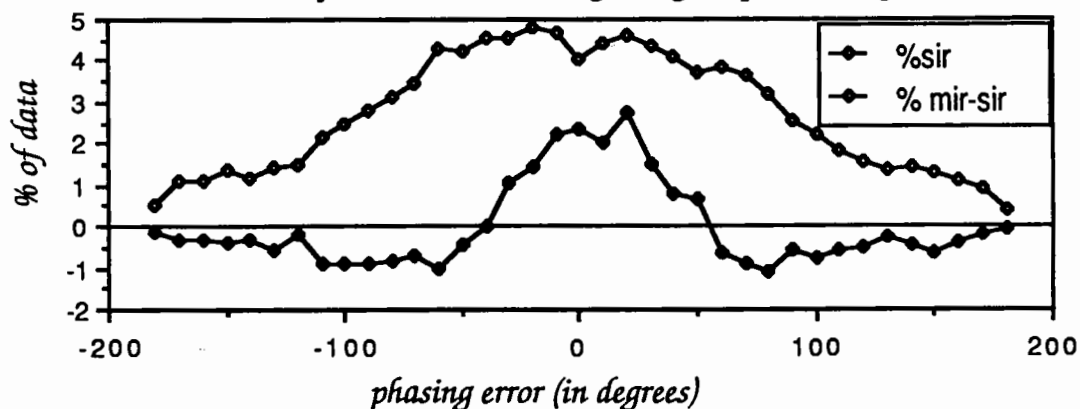


Fig. 5. The comparison of the phase error distribution in data phased by MIR and SIR methods; the top curve shows the percentage of data phased by SIR vs. the error, the bottom curve gives the increase of the percentage of data for each error range of the MIR over SIR phasing.

allegedly successful applications over a wide range of cases have been reported. In this paper I would like to discuss briefly the results of phase modification in the context of the phase error distribution. In each case the same procedure of 4 cycles of solvent flattening, followed by phase combination, was applied.

a) α -amylase:

The amylase crystals contain less than 50% of solvent. Our initial investigations, plagued by difficulties with heavy atom derivatives at a low pH (3-3.5), focused on a single Hg derivative. Solvent flattened electron density SIR maps allowed us to identify the correct hand, and was also used subsequently with the MIR phases, when they became available. While a clear contrast between the protein regions and solvent cavities was immediately visible, the improvement in the interpretability of the of electron density inside the molecule was less apparent. Fig. [6] illustrates the marginal change in phase error distribution achieved by solvent flattening and subsequent phase combination. The associated absolute mean change in phase angles was just under 20° , while the combined figure of merit (only fully phased acentric data has been used throughout these calculations) increased to 0.75. The only interesting feature of this comparison (denoted by the arrow) is a gradient which shows a small shift of the main MIR peak, originally centered around 15° (see Fig [7] shows the results of the application of solvent flattening to the SIR phases

calculated for the Pb derivative. The phase ambiguity has

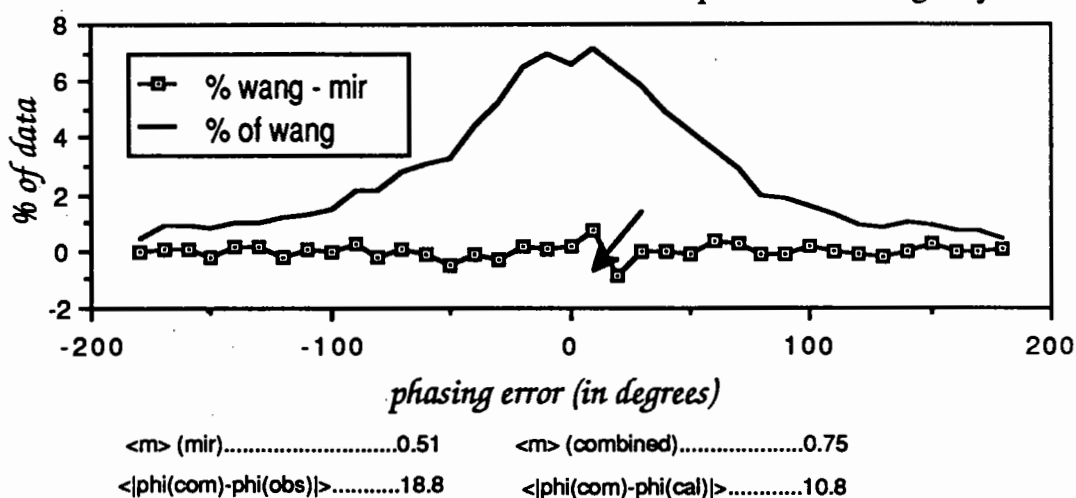


Fig. 6. The effect of solvent flattening on the acid amylase phasing accuracy; the graph shows the percentage of data against the actual phase error, and the increase in correctly phased reflections over pure MIR calculation. The improvement small and virtually limited to a small subset of reflections (indicated by the arrow) whose slight error was subsequently traced back to heavy atom scaling inaccuracy. The changes in the values of figures of merit and shifts in the phase angles are also given.

been successfully resolved in a high percentage of data and the resultant phase error distribution is similar to that of the MIR phases.

b) *Humicola sp. lipase*:

As already mentioned the *Humicola* lipase presented severe experimental difficulties largely owing to the extremely high solvent content (80%). The only advantage of this situation was that it could have been exploited by the solvent procedure. The quality of the MIR phasing was relatively poor. Only some 55% of data were phased with an error between -50° and $+50^\circ$. Fig. [8] show the result of the successful application of solvent flattening. The resultant electron density map, additionally helped by averaging around a non-crystallographic two-fold symmetry axis, was fully interpretable. Note, that the combined figure of merit increased to 0.75 (same value as for amylase), while the mean shift in phase angles was 21.9° (only insignificantly higher than for amylase). It must be concluded, that these usually reported indicators are insufficient to assess the real improvement in phase quality.

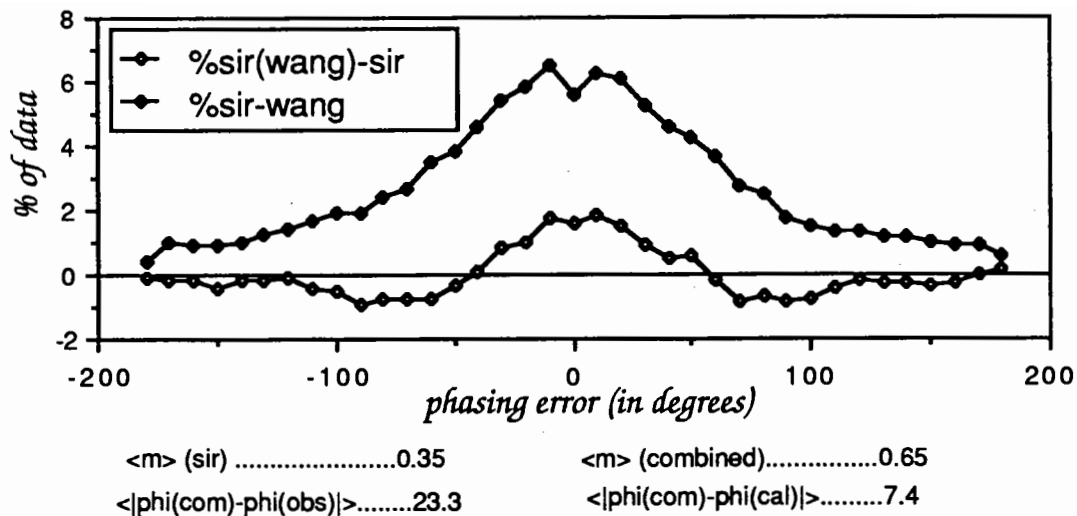


Fig. 7. The effect of solvent flattening on the accuracy of acid amylase SIR (Pb) phasing (see text for further details); the graphs shows both the distribution of the phase error achieved via this method, and the statistical improvement (the increase in the percentage of data) of the quality of phasing. The usual statistics (figure of merit, shifts of phase angles) are also listed.

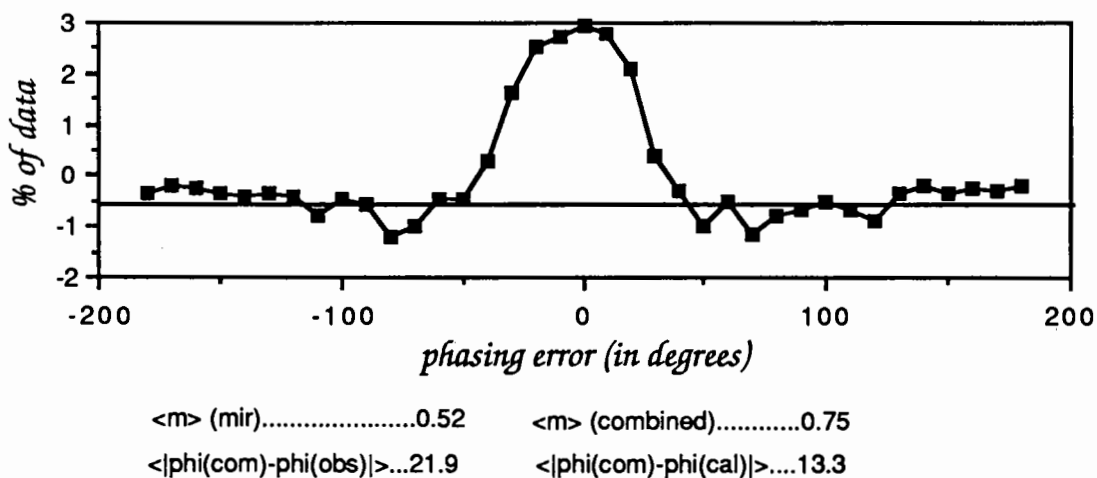


Fig. 8. The improvement of phasing accuracy in the case of the *Humicola* lipase, following solvent flattening. Only the increase in the percentage of well phased reflections is shown (% of data after solvent flattening minus % of data before, against the actual error in phase, calculated vs. the value from the final set of coordinates). See text for other details.

fig. 6). This asymmetric distribution of phase error was traced back to a slight scaling error in one of the derivatives. Solvent flattening corrected this error.

The Molecular Replacement Method and the associated error in phase angles.

a) α -amylase:

During our investigations of the amylase crystal structure we found that the atomic coordinates of the TAKA amylase (entry 2TAA in the Protein Data Bank) could be used successfully for the Molecular Replacement calculations. The validity of this model was checked by difference Fourier maps between the heavy atom derivatives and the native data, phased on the Molecular Replacement solution. The positions of the metal ions were clearly revealed by this procedure. However, the TAKA model was too inaccurate to provide a good starting point for refinement. Fig. [9] shows the distribution of the phase error. The contribution of the incorrectly phased reflections is additionally enhanced by the lack of an adequate weighting scheme. As a result, the error vectors exceed those in the MIR phasing considerably. In addition, the strong internal bias towards the model structure largely invalidates the MR phases. It did occur to us, however, that the MR solution could be used in principle in a manner similar to solvent flattening, to resolve the phase ambiguity in the SIR case. We have therefore carried out phase combination using the MR and Pb SIR phases. The

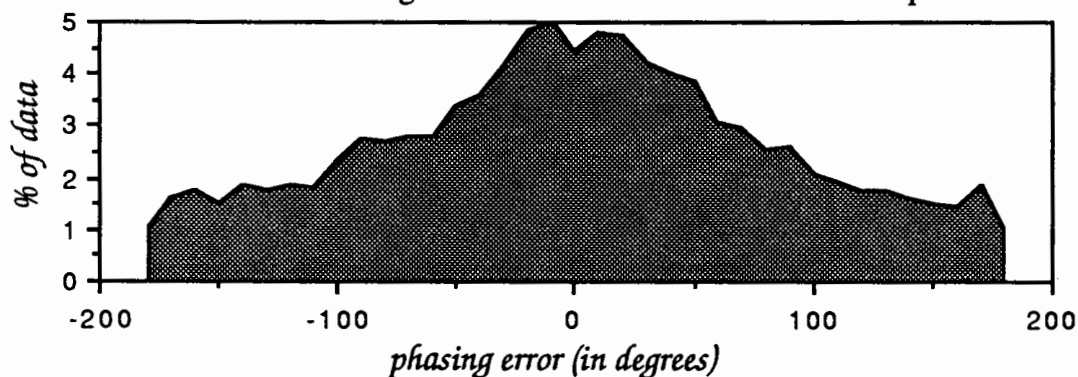
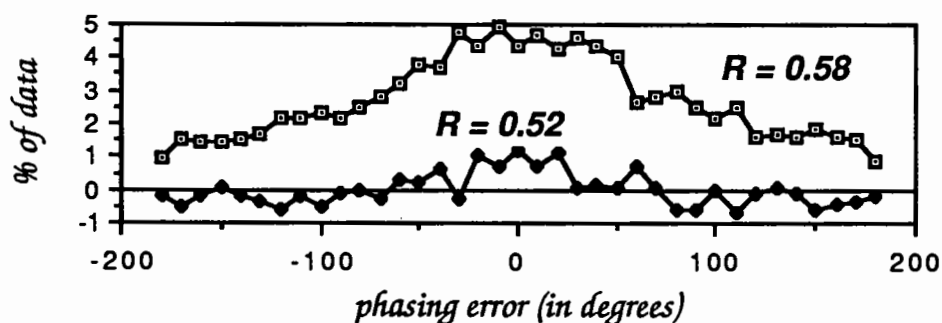


Fig. 9. The distribution of phase error in acentric acid amylase data phased on the Molecular Replacement solution using the 2TAA coordinate set from the Protein Data Bank.

was found to be strongly resolution-dependent. The combination works best at low resolution, outperforming solvent flattening, while beyond 4-5Å it actually deteriorates the quality of the SIR phases.

b) low-pH insulin:

Molecular Replacement has been used extensively in York to solve a number of different insulin crystal forms. Our investigations of the orthorhombic form were particularly difficult, since the R-factor search failed to give an unambiguous solution. A number of possible models (different only with regard to their translational parameters) were studied, with some refined to an R factor of as low as 0.24, before being abandoned (Derewenda, 1990). Fig. [10] illustrates the phase error distribution for the correct MR solution, and the improvement in phasing accuracy achieved by the use of the least-squares rigid body refinement protocol (Derewenda, 1989). It can easily be seen, that while the initial phase error distribution is by far inferior to the MIR cases studied, the use of the least-squares protocol introduces significant change. It is also instructive to analyse what exactly happens to the phase error distribution upon the least squares refinement of an incorrect solution, in comparison to a correct one. Fig. [11] shows the phase improvement obtained



Centric data:

in phase.....614 (+82)
out of phase494 (-82)

Fig. 10. The distribution of errors in phase angles (acentric data) obtained from a correct Molecular Replacement solution. The bottom curve shows the improvement in phasing following the application of the least-squares rigid-body refinement protocol, suggested by Derewenda (1989). The statistics for centric reflections are given separately, together with the details of improvement following the refinement.

after several cycles of restrained least-squares refinement, and contrasts this result with a totally random distribution of phase error for a "refined" incorrect solution. The relevant R factors are also shown.

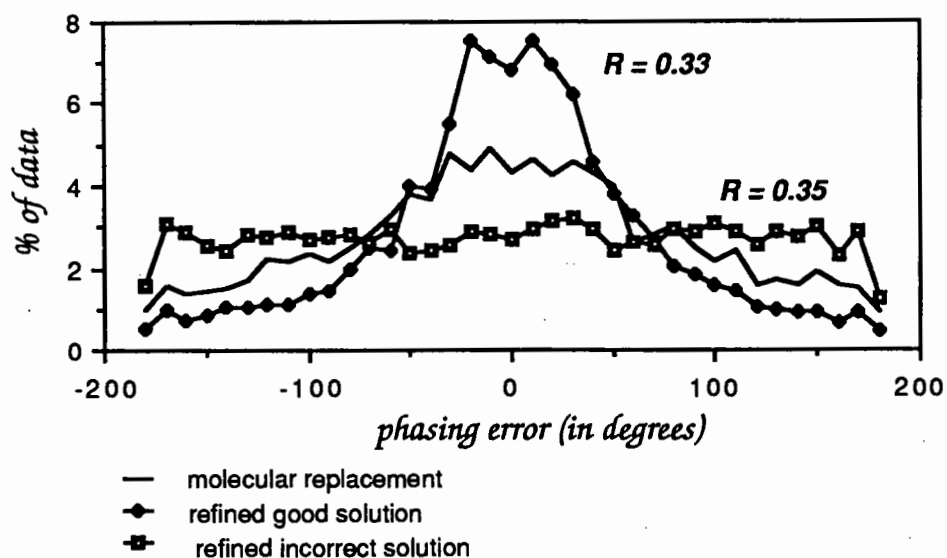


Fig. 11. The improvement in phase accuracy obtained by restrained least-squares of the correct solution, and the random distribution of phase error for the "refined" incorrect one. Note, that the R factors are comparable, and the quality of the resulting electron density map is similar (though one displays only the "ghost" molecule).

Conclusions:

The analyses shown here illustrate the difference in the accuracy of phase angle estimates obtained by various techniques. It seems appropriate to draw certain general conclusions.

1. The Molecular Replacement method returns phases which in most cases are inferior to those calculated from an MIR experiment. However, if the MR solution is correct, it will allow one to calculate a more interpretable electron density map. Any rigid-body refinement of the parameters of the model molecule is strongly recommended, as it substantially improves the quality of phasing. Utmost care should be taken in the assessment of the MR solution. We have no objective criteria, which could differentiate between a correct and an incorrect solution. It is also possible to perform miracles with subsequent refinement; we have had cases in our laboratory when a wrong solution was brought down to an R value of ca. 0.20.

2. The MIR method has been shown allow an accurate calculation of phase angles, with an expected distribution of error. Solvent flattening was shown to work well improving the accuracy of phasing only if the solvent content was considerably above average.

3. More work should be carried out on the proper weighting of terms in the calculation of the MIR electron density; the commonly used figures of merit do not reflect the true distribution of error.

Acknowledgements:

The projects referred to in this paper are financed by a grant from NOVO-Nordisk Industri A/S, Copenhagen; structural studies in York are supported by a consolidated grant from the Science and Engineering Research Council, U.K.

References:

- Blundell and Johnson* (1976) *Protein Crystallography*, Academic Press (London)
- Blow, D.M., Crick, F.H.C.* 12, (1959) 794
- Boel, E., Brady, L., Brzozowski, A.M., Derewenda, Z., Dodson, G.G., Jensen, V.J., Petersen, S.B., Thim, L., Woldike, H.F.* (1990) *Biochemistry* - in press
- Brady, L., Brzozowski, A.M., Derewenda, Z.S., Dodson, E., Dodson, G., Tolley, S., Turkenburg, J.P., Christiansen, L., Huge-Jensen, B., Norskov, L., Thim, L., Menge, U.* 343 (1990a) 767
- Brady, L., Brzozowski, A.M., Derewenda, Z.S., Dodson, E.J., Dodson, G.G.* (1990b) *Acta Crystallogr.* - submitted
- Derewenda, U.* (1990) York University, Thesis.
- Derewenda, Z.* A45 (1989) *Acta Crystallogr.*, 227
- Wang, B.C.* 115, (1985) *Meth. Enzym.* 90

Analysis of Errors found in Protein Structure

Coordinates in the Brookhaven Data Bank

**Janet M Thornton, Malcolm W McArthur, David K Smith,
Stephen P Gardner, E Gail Hutchinson, A Louise Morris, Bancinyane L. Sibanda**

**Laboratory of Molecular Biology
Crystallography Department
Birkbeck College
Malet Street
London WC1E 7HX**

1. Introduction

There are now over 400 coordinate sets in the Brookhaven Databank⁽¹⁾ and it is becoming impossible to 'know' each structure individually and remember details such as resolution, refinement, 'poorly resolved' regions of the structure etc. To date the resolution of a structure has been widely used as a guide to its accuracy, so that for example, in considering detailed residue interactions only structures resolved to 2.0Å or better are used. However, using the data immediately highlights some coordinate sets with unacceptably bad close contacts or 'forbidden' dihedral angles. Consequently it has become increasingly important to develop alternative measures of accuracy or 'tests' which can be used to check structures automatically. However, the 'accuracy' within any structure is variable, with the 'core' usually being well resolved with good electron density, whilst many of the loops have very weak density and are often a best 'guess' rather than a definitive interpretation. Therefore there are two important problems to be considered:

- (1) Given poor density, what information derived from our current knowledge-base of protein structures can be used to improve our 'best guess'?
- (2) Given the coordinates, which methods can be used to assess accuracy? These methods may be useful for identifying 'incorrectly interpreted' structures as refinements proceed or 'get stuck'.

At Birkbeck College, in collaboration with Leeds University, we have been establishing a database of protein structures^(2,3). This has required programs to be written which can handle any Brookhaven file and identify minor errors - such as labelling etc. In this contribution we present some of our findings, calculated using the new database STEP⁽⁴⁾, which provide a basis for a comparison of structures and assessment of the accuracy of deposited and future coordinate sets. We also describe briefly our classification of β -hairpin loops which may assist in the interpretation of weak density in surface loops.

2. Labelling Errors

Many Brookhaven files include labelling errors or inconsistencies which need to be identified and flagged.

2.1 Sequences

The SEQRES records and the atom records sometimes disagree. We found 11 examples including missing residues, transposed sequences and incorrect atom records. We found a further 6 examples where the chain length was incorrect in SEQRES. In addition there are a surprising number of discrepancies between the sequence in the Brookhaven file and the sequence stored in one of the sequence databanks. Arthur Lesk and colleagues explored the latter⁽⁵⁾ and found only 31% of sequences in Brookhaven were identical in the NBRF-PIR sequence databank. Most

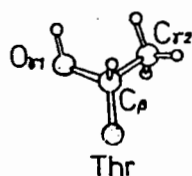
of these discrepancies arise from differences in the state of maturation of the proteins but some are more significant.

2.2 Atom records

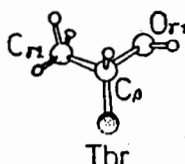
Amino acid labelling is clearly defined by the IUPAC-IUB Commission⁽⁶⁾, but is not always followed in Brookhaven files. In particular labelling of arginine, aspartic and glutamic acids and the aromatics, phenylalanine and tyrosine, is often incorrect. For arginine the labelling error is significant since the guanidine group is planar and it is always possible to distinguish NH1 from NH2 (see Fig 1). This inconsistency, found in 13% of arginines in the database, becomes apparent when arginines are superposed. For the other side-chains the labelling depends on identifying which atom defines the smallest value of the relevant angle, eg in tyrosine, CD1 is defined as the atom which gives the lowest value of X_2 . Consequently for these residues the labelling actually changes as the conformation of the side chain changes and can therefore be considered rather arbitrary. Almost 50% of Phe, Tyr, Asp and Glu residues are 'incorrectly' labelled. However, to be consistent we feel this labelling should be checked before deposition with Brookhaven. We have relabelled all the atoms consistently in the database.

Fig. 1 Atom Labelling

Threonine

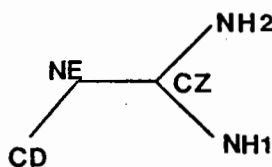


correct

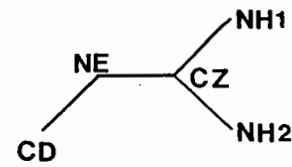


incorrect
(20 examples found)

Arginine



correct



wrong

In threonine and isoleucine the chirality of the sidechain must be correct. (See Fig 1). On checking we found 20 examples of threonines with incorrect chirality out of a total of 6,374 (0.3%). All the isoleucines had the correct chirality but 2 were incorrectly labelled. In threonine relabelling, interchanges the carbon and oxygen and so will affect local hydrogen-bonding.

Since the electron density cannot distinguish oxygen, nitrogen and carbon another potential error occurs in asparagine, glutamine and histidine, where the orientation of the side-chain is mainly decided by the non-covalent contacts formed. Checking through the coordinates reveals several examples where two apparently 'bad' contacts could become two good hydrogen bonds if the atoms were relabelled (equivalent to

the side-chain being rotated through 180°). For example, glutamine 49 in wheat germ agglutinin looks a good candidate. Again such a check is a useful way to highlight possible errors.

3. Indicators of Local Errors

The ultimate assessment of a correct interpretation of the electron density must be found by calculating the agreement between the model and the density. The real-space R-factor plotted by residue is often a good guide to local accuracy.⁽⁷⁾ However, Table 1 lists some other useful indicators which can be derived from Brookhaven data files. Below we describe how we have used some of these parameters to investigate the 'quality' of the structure.

Table 1 INDICATORS OF LOCAL ERRORS

<u>Electron Density Fitting:</u>	<u>Contacts:</u>
Local R-factor	Van der Waals interactions
B-values	Hydrogen Bonds
<u>Basic Geometry:</u>	<u>Chirality:</u>
Bond lengths and Angles	L-Amino Acids
Planarity: rings	Right-handed α -helices and twist in β -strands
peptide group	Preferred packing angles between strands/helices
arginine guanidium group	Right-handed $\beta\alpha\beta$, $\beta\chi\beta$ units
<u>Dihedral angles:</u>	Type I, II β -turns in short loop β -hairpins
ϕ, ψ (glycine, proline, all other aa)	Handedness of motifs eg. 4 α -bundle
ω (cis and trans peptides)	
x_1 trans, g^+ , g^-	
x_3 disulphide	

4. Bond Lengths and Angles

In a 'good' structure internal bond lengths and angles should conform to known stereochemistry and 'regularisation' of structures aims to improve this. We have not yet completed the study of all bond lengths and angles and their variation, but consideration of disulphide geometry illustrates the problem. The distribution in S-S bond lengths is shown in Fig 2. The average sulphur separation equals $2.04 \pm 0.16 \text{ \AA}$. However, the databank includes some disulphides with very unusual S-S separations which suggest that the detailed geometry and packing in the vicinity of these disulphides may be wrong. Inspection of individual proteins shows that many S-S separations have been restrained to 2.0 \AA during refinement. Consequently this parameter is not always a good guide for 'accuracy'.

Fig. 2 Distribution of Sulphur-Sulphur Separations in Disulphide Bridges

SSDIST	TOTAL	HISTOGRAM
1.5	1	*
1.6	2	*
1.7	8	*
1.8	11	*
1.9	40	**
2.0	412	*****
2.1	199	*****
2.2	15	*
2.3	6	*
2.4	1	*
2.6	1	*
2.7	2	*
2.8	1	*
2.9	2	*
3.0	6	*
3.1	2	*
3.2	1	*
3.3	1	*

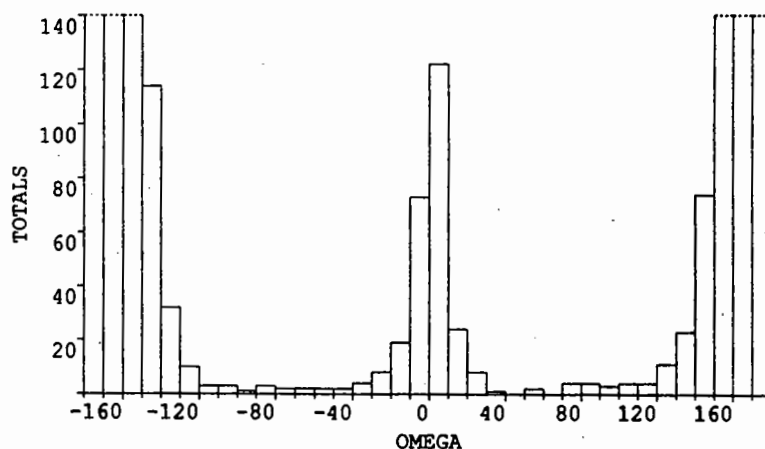
5. Dihedral Angles

In contrast most dihedral angles are not tightly restrained during refinement and so provide a good assessment of local 'accuracy'. The 'useful' dihedral angles for this purpose are ω , ϕ , ψ , χ_1 's and χ_3 (disulphide).

5.1 Omega Angle

The ω angle values were extracted from the database and average values and standard deviations calculated for proteins at a given resolution. The distribution of ω angles is shown in Fig 3. Although the vast majority cluster within $\pm 30^\circ$ of the perfect cis or trans conformers, there are a number of 'rogue' values almost at 90° . Some proteins show very low standard deviations of ω angles reflecting restraints applied during refinement. Consequently, there is no correlation between the standard deviation and resolution of the protein.

Fig.3 Distribution of ω Angles.






5.8% prolines exhibit cis peptides before the proline ring compared to only 0.06% of other residues. The occurrence of cis peptides is influenced by the amino acid type before proline, so that for example 18% of Tyr-Pro sequences in the databank have a cis peptide compared to only 1% of Asp-Pro dipeptides⁽⁸⁾. Inspection of the database reveals only 45 non-identical non-proline cis peptides ($\omega < 90^\circ$) in 15 proteins. A single protein (1HDS) accounts for 25 of these occurrences, and 1CY3 for another 4. Excluding these two proteins and restricting attention to resolutions better than 2.5Å and $|\omega| < 30$ leaves only 8 examples in almost 90,000 residues. Whether this reflects reality or merely a reluctance to interpret density as such cis-peptides is debatable. Interpretation depends critically on identifying the density for the peptide oxygen and the local backbone atoms.⁽⁹⁾ It requires very good density to be sufficiently confident to suggest such an 'unusual' conformation.

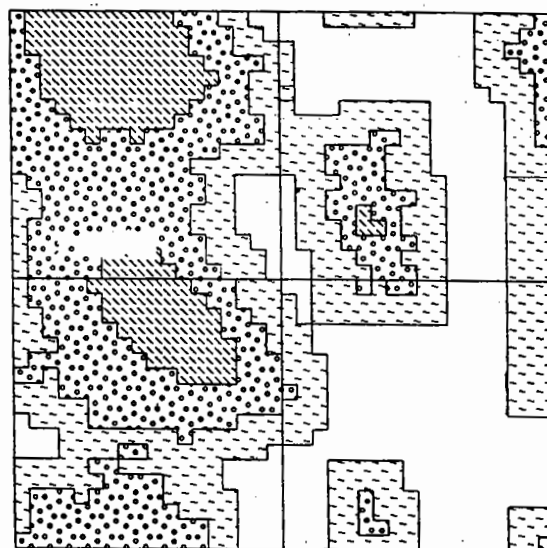
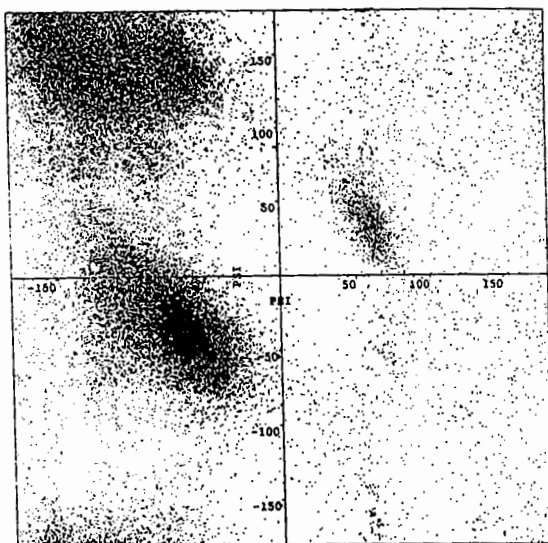
5.2 ϕ, ψ Angles

Ramachandran⁽¹⁰⁾ showed that ϕ, ψ space for a dipeptide is very restricted for all residues except glycine due to steric clashes. It has often been observed that as the refinement of a protein proceeds the ϕ, ψ angles migrate into 'allowed' conformations. Therefore the distribution of ϕ, ψ angles can provide a guide to the accuracy of a structure. We empirically defined three sets of 'allowed' ϕ, ψ values with increasing tolerance (see Fig 4), using observed ϕ, ψ distributions for 310 proteins extracted from the database. [Glycine and proline residues were excluded from this study.] The core region included all $10^\circ \times 10^\circ$ pixels with more than 100 entries. The next level (allowed) was extended to include all areas with 8 or more entries. The third level was defined by extending out by 20° all round the allowed region. Whilst this is somewhat arbitrary, given the different shapes of the ϕ, ψ energy surface, it provides a good simple definition.

Fig. 4.

a) Observed ϕ, ψ Distribution for 310 proteins, excluding gly and pro.

b) Three sets of ϕ, ψ :  'core',  'allowed' and  'generous'.

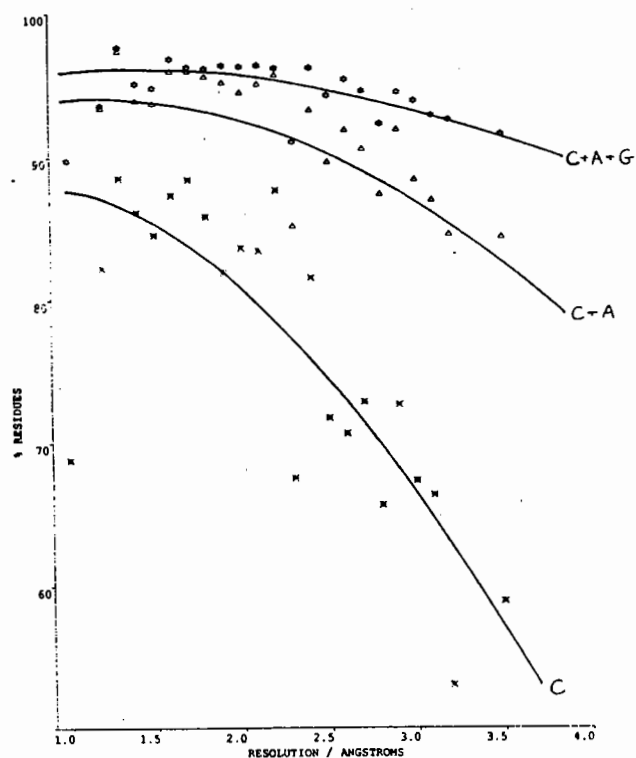
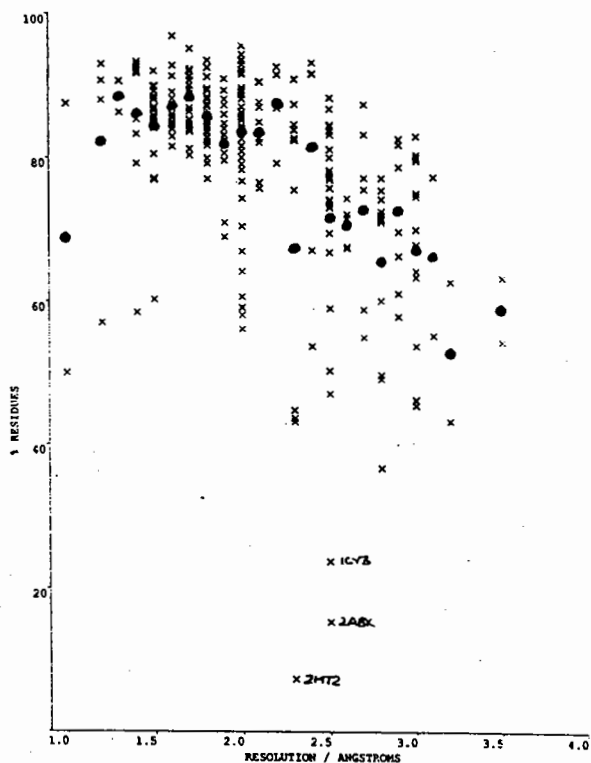


We then calculated for each protein the percentage of residues which lie inside these regions. The results are shown in Fig 5, plotting each protein against its resolution and calculating the average value for each resolution.

Fig. 5

a) % Residues in core region for all proteins.

b) % Residues in each of 3 regions averaged for a given resolution.



x Individual protein
● Averaged by resolution

* Core
△ Core and Allowed
◇ Core + Allowed + Generous

Four facts clearly emerge:

a) There is a correlation between resolution and the tightness of the ϕ, ψ distribution. On average at higher resolutions fewer residues lie outside the 'allowed' regions.

b) For a given resolution range the proteins show very different 'accuracies'. Some proteins have good ϕ, ψ distributions even at relatively low resolution, whilst others show remarkably bad distributions even though the data go to high resolution. It is possible to define 'outliers', and these proteins often have low R-factors and incomplete refinement. Some of these incorrect ϕ, ψ values almost certainly correspond to peptides which have flipped over by 180° . This conformational change has relatively little effect on the direction of the chain. One method to identify such flips may be to use the database of structures and, given C_α coordinates of a model,

search for matching structures using the approach suggested by Jones and Thirup⁽¹¹⁾. If all such extracted structures have allowed ϕ, ψ values this may indicate that the model should be changed. This was apparent when we attempted to model the structure of flavodoxin from C_{α} coordinates⁽¹¹⁾. Our model of the backbone, derived from database-extracted peptides, disagreed with the structure in four places; three corresponded to disallowed ϕ, ψ values and an apparent flip of the peptide. This suggests that the electron density in this region of the structure should be reinvestigated.

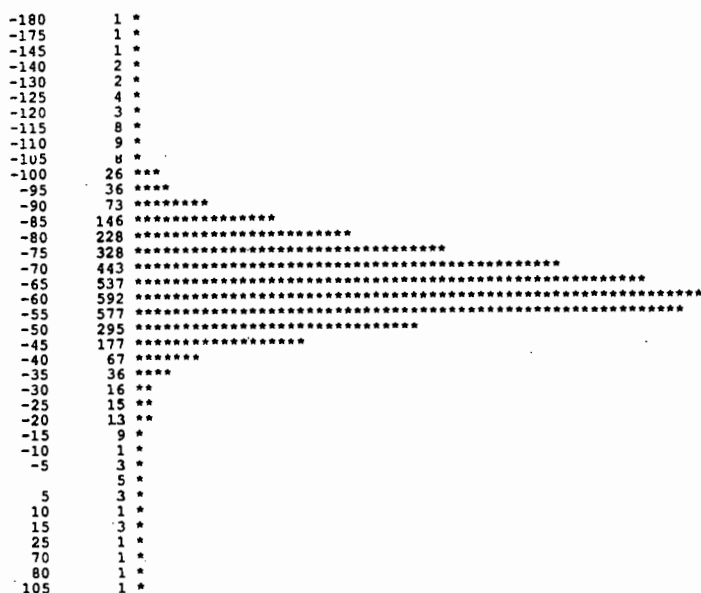
c) At high resolution there remain a few residues which adopt 'disallowed' conformations, even using the most generous definition of 'allowed' ϕ, ψ .

d) Below 2Å the curves flatten out which suggests that, as expected, almost all the backbone conformation is relatively well defined by this resolution. (but see 5.4).

5.3 Proline ϕ Angles

The proline ring restricts its ϕ angle to an average value of $-63^{\circ} \pm 15^{\circ}$. The ring itself is usually slightly puckered and may or may not be restrained during refinement. Again we used the database to explore the distribution of proline ϕ 's, as shown in Fig 6, which illustrates that some prolines are distorted beyond the limits of acceptability.

Fig. 6 Distribution of Proline ϕ Angles.

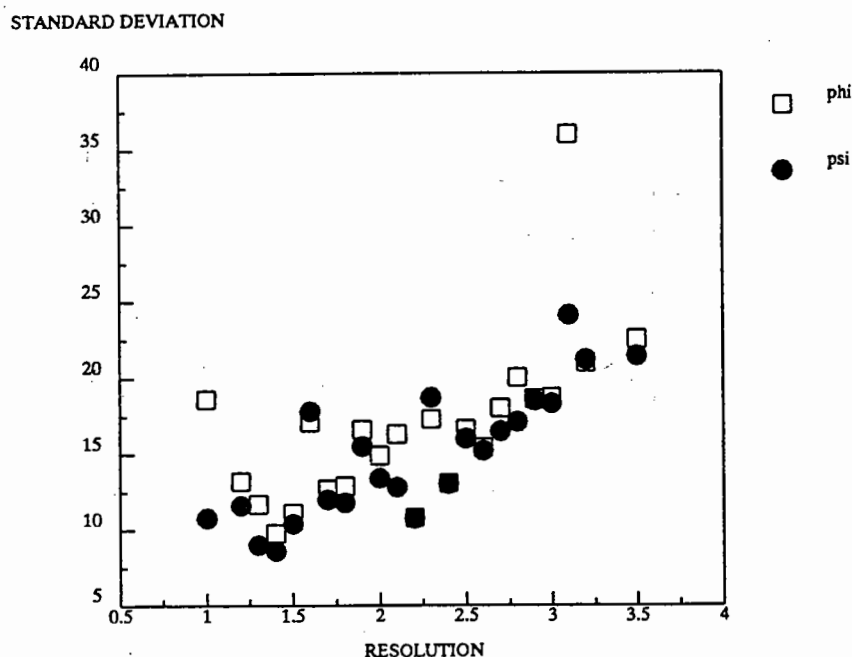


5.4 Helical Structures

Further consideration was given to the backbone dihedral angles in secondary structures. ϕ, ψ values of helical residues are tightly clustered and as refinement

progresses the standard deviations of such angles should decrease as helix geometry improves. We used the database to calculate averages and standard deviations for ϕ and ψ in helices, defined by the Kabsch-Sander algorithm¹⁸. The plot of average values and of standard deviations against resolution shown in Fig 7, reveals a striking correlation with resolution. The standard deviations appear to decrease even below 2Å suggesting continued improvement.

Fig.7 STANDARD DEVIATIONS OF PHI & PSI ANGLES VERSUS RESOLUTION WITHIN HELICES



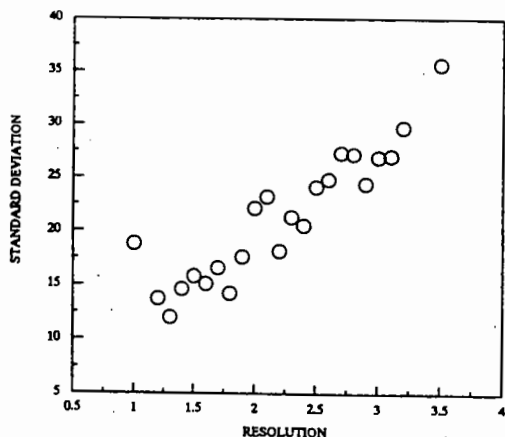
The same plot for β -strand residues shows much less correlation, almost certainly because ϕ, ψ values in strands are not tightly clustered since the strands have different twists and distortions.

5.5 χ_1 Angles

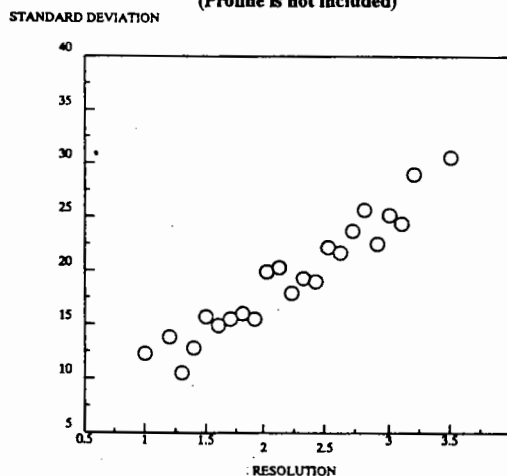
χ_1 angles have been shown to adopt one of three preferred conformers $g^+(+60^\circ)$, $t(+180^\circ)$ or $g^-(-60^\circ)$ ⁽¹³⁾. As refinement proceeds the χ_1 distribution becomes more tightly clustered into these three idealised energy wells^(14,15). This is illustrated by Fig 8 which shows that as the resolution improves the standard deviation for χ_1 angles reduces dramatically. The plot shows no flattening off below 2Å, unlike the coarse ϕ, ψ classification. Some side chains are difficult to locate even at 2Å and significant improvements can be made at higher resolution. Again the standard deviation for χ_1 in each rotamer can provide a guide to the 'global' accuracy of the structure, whilst very unusual χ_1 's may signify a local 'error'. Secondary structure preferences could also be taken into account⁽¹⁵⁾.

Fig. 8

a. CORRELATION OF GAUCHE-MINUS CH1 WITH RESOLUTION
(Proline is not included)



b. CORRELATION OF GAUCHE-PLUS CH1 WITH RESOLUTION
(Proline is not included)



5.6 Disulphide χ_3 angle

It is known that the χ_3 angle $C_\beta-S-S-C_\beta$ in a disulphide bridge adopts either a right or left handed conformation^(16,17). The distribution of angles plotted in Fig 9 shows strong clustering around the right handed ($\chi_3 = 93.9^\circ \pm 10.1^\circ$) and left-handed ($\chi_3 = -87.1^\circ \pm 8.6^\circ$) conformers. The distribution is not quite symmetrical and there are many disulphides which have been built into very distorted conformations.

Fig.9 Distribution of χ_3 Angles for Disulphides.

ANGLE	TOTAL	HISTOGRAM
-180	1	*
-160	6	**
-140	5	*
-120	17	**
-100	93	*****
-80	181	*****
-60	18	**
-40	9	*
-20	2	*
0	1	*
20	1	*
40	3	*
60	27	***
80	96	*****
100	216	*****
120	25	***
140	3	*
160	6	**
180	1	*

6. Chirality

One consequence of the restriction to L amino acids in proteins is that many structural features have a preferred chirality, summarised in Table 1. In assessing coordinates the chirality of such features should be checked and if incorrect may suggest either an unusual structure which deserves investigation or an incorrect interpretation. To date we have only addressed the first simplest feature, L and D amino acids, by calculating the virtual dihedral angle $C_\alpha-N-C-C_\beta$. For 93,000 residues this angle has an average value of $33.81^\circ \pm 4.17^\circ$, as appropriate for L amino acids. There are 29 occurrences of values $<0^\circ$, which can be considered D amino acids, although the distribution shows distortions occur in both directions. We are currently considering the chirality of secondary, supersecondary and tertiary structures.

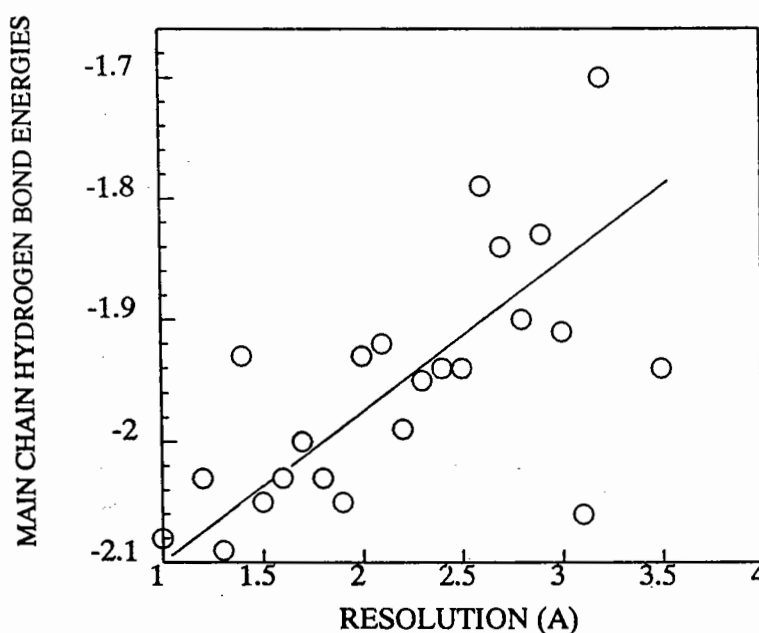
7. Non-covalent Interactions

These perhaps provide one of the best guides to structural accuracy since they are rarely constrained during refinement.

7.1 Hydrogen Bonds

Main chain hydrogen bonds are numerous in all proteins, mainly in the α -helix, β -sheet and β -turn secondary structures. We would expect that as refinement proceeds the geometry of the hydrogen bonds will improve with a corresponding reduction in energy. Fig 10 shows the variation of average H-bond energy, calculated using the Kabsch-Sander method¹⁸, for residues defined to a given resolution. It shows a definite correlation, confirming the improvement in hydrogen bonding geometry as the resolution decreases. The scatter at low and high resolution reflects the smaller number of proteins (and residues) at the extremes of the resolution range. Given the gross simplification of considering all hydrogen bonds the correlation is surprisingly good.

Fig. 10 Hydrogen Bond Main Chain Energies versus Resolution.



7.2 Non-covalent interactions

Close atom contacts which violate the minimal accepted Van der Waals separation usually signal poor density or interpretation. Islam et al (in this volume) find a strong correlation between resolution, R-factor and the numbers of bad contacts.

8. Loop Conformations

As discussed in the introduction, the electron density for the loop regions of a protein is often poor. For these loops crystallographers either provide best-guess coordinates or omit these atoms altogether. Our studies of loops¹⁹ in supersecondary motifs have allowed a classification of structural families which can provide guidelines for modelling these loops when the density is poor.

The β -hairpins are perhaps the best example²⁰. Here the constraints imposed by the β -ladder restrict the number of possible conformers of short loops between strands. A classification given in ref 20 can be used to improve the 'best-guess'. Indeed we have found that loops with the appropriate length and sequence often change from the 'irregular non-specific structures' to become one of the known family conformers as refinement proceeds. (This parallels the observed improvement in ϕ, ψ and X values). This is illustrated in Fig 11 which shows loops in endothiapepsin and penicillopepsin whose conformers have 'standardised'. Thus we suggest that if the density of a loop is too poor to interpret 'ab initio', and the loop length and sequence (especially critical glycines which can adopt an α_L conformation) correspond to a 'family' structure, this structure should be used as the first guess. The method of Jones and Thirup¹¹ combined with cognizance of amino acid sequence will effectively perform this function. There are clearly 'dangers' associated with this knowledge-based approach, and the crucial test is to use the electron density and local R-factor as the final adjudicator. If it cannot distinguish between a very unusual structure and a very common one, choosing the common structure will most often give the 'correct' answer. The use of the knowledge-base to help interpret electron density is already widespread and this extension will almost certainly become part of all programs to model into density.

9. Conclusions

The results of these analyses can be used in two important ways:-

- (i) To help interpret electron-density by locating 'incorrect' regions in any map. Probably the best approach is to highlight areas (using colour coding) with poor dihedral angle values, poor hydrogen bonding and bad Van der Waals contacts. The electron density should be inspected by eye to see if alternative interpretations are possible.
- (ii) To assess the 'accuracy' of any given structure, using different criteria. The parameters which are not constrained during refinement and show a clear correlation with resolution, would seem to be best suited for this purpose. In particular ϕ, ψ distributions, X_1 angle distributions and H-bond energies. Non-covalent contacts also important (see Islam et al, in this volume).

The analyses we have presented here are preliminary and this important area requires further study. In particular we need to explore the correlation between different parameters, the correlation of 'errors' with R-factor and refinement method, and the level of 'acceptable' errors. For example, do residues with bad ϕ, ψ values, also have bad X angles and bad non-covalent contacts? This work is in progress.

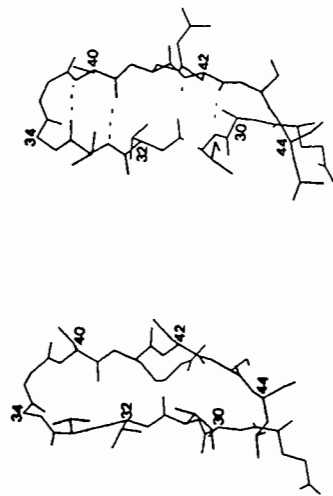
Fig. 11

LOOP CONFORMATION

As refinement proceeds loop conformations sometimes become 'regular'.

- eg. (1) Strept. Griseus Proteinase A., Brayer et. al. (1978)
 J. Mol. Biol. 124 243

Loop in β -hairpin residues 29-44



2.8 Å resolution
 6:6 β -hairpin

1.5 Å resolution
 2:2 II' β -hairpin

Thr - Thr - GLY - Gly - Ser : Lys

Sequence and ϕ, ψ values (deg.) at 2.8 and 1.5 Å resolution for β -hairpin 29-44 in *S. griseus* proteinase A

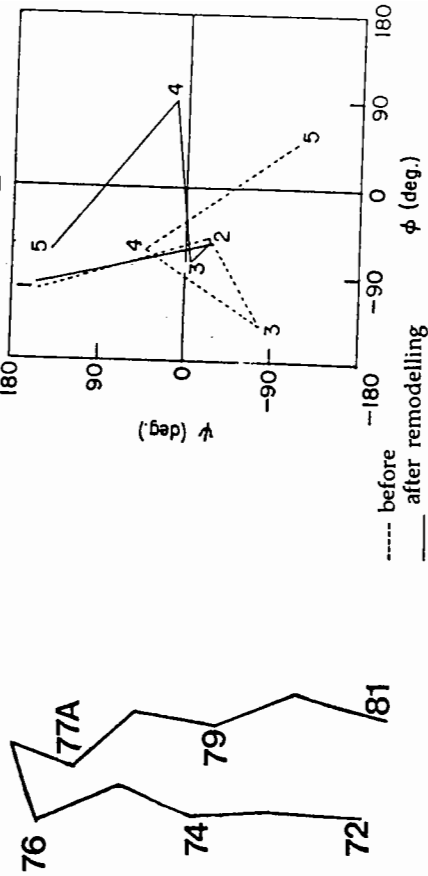
Residue no.	Sequence	1.8 Å 2.8 Å		2.8 Å 1.5 Å	
		ϕ	ψ	ϕ	ψ
29	Glu				
30	Ala				
31	Ile				
32	Thr	-141	-143	-136	160
33	Thr	165	-109	-147	121
34	Gly	-81	-119	53	-130
39	Gly	-95	26	-88	-1
40	Ser	-117	152	-89	148
41	Lys	-141	128	-134	135
42	Cys				
43	Ser				
44	Leu				

(2) Endothiapepsin

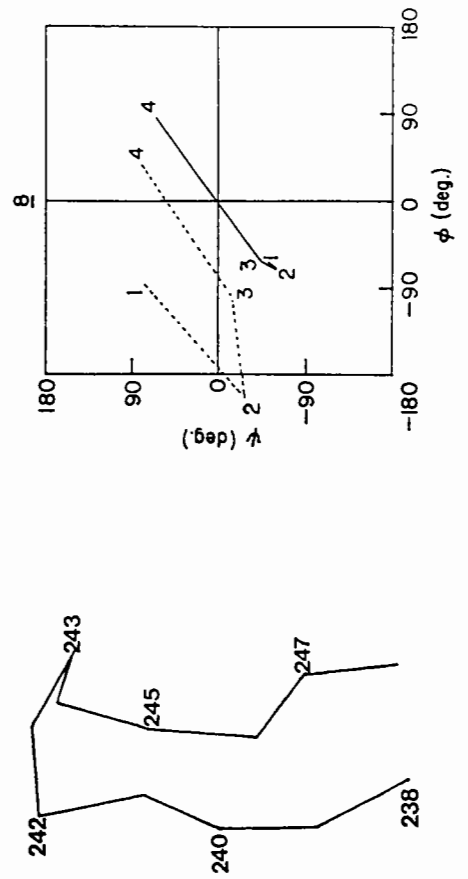
Blundell et.al. (1990) J. Mol. Biol. 211 919-941

Hairpin 3 : residues 75-78

Tyr - Gly - Asp - GLY - Ser



Hairpin 8 : residues 241 - 244
 Ser - Ser - Val - GLY 3:5



References

1. Bernstein, F.C. et al. *J.Mol. Biol.* 122 (1977) 535.
2. Akrigg, D. et al. *Nature*, 335 (1988) 745.
3. Islam, S.I. & Sternberg, M.J.E. *Prot. Eng.* 2 (1989) 431.
4. Smith, D.A. et al (1990) in preparation.
5. Lesk, A.M. et al. *Prot. Sequences & Data Analysis*, 2 (1989) 295.
6. IUPAC-IUB Commission, *J. Mol. Biol.* 52 (1970) 1.
7. Branden, C-I. & Jones, A. *Nature*, 343 (1990) 687.
8. McArthur, M.W. & Thornton, J.M. (1990) in preparation.
9. McPhalen, C. & James, M. *Biochem.* 27 (1988) 6582.
10. Ramachandran G.N. et. al. *J. Mol. Biol.* (1963) 7 95.
11. Jones, T.A. & Thirup, S. *EMBO J.* 5 (1986) 819.
12. Reid, L.S. & Thornton, J.M. *Proteins*, 5 (1989) 170.
13. Janin, J. et al. *J. Mol. Biol.* 125 (1978) 357.
14. Ponder, J.W. & Richards, F.M. *J. Mol. Biol.* 193 (1987) 775.
15. McGregor, M. et al. *J. Mol. Biol.* 198 (1987) 195.
16. Thornton, J.M. *J. Mol. Biol.* 151 (1981) 261.
17. Richardson, J. *Adv. Protein. Chem.* 34 (1981) 167.
18. Kabsch W. & Sander C. *Biopolymers* 22 (1983) 2577.
19. Thornton J.M. et. al. *Bioessays* 8 (1988) 63.
20. Sibanda B.L. et. al. *J. Mol. Biol.* 206 (1989) 759.

BAD CONTACTS IN PROTEIN STRUCTURES

Suhail A. Islam, Michael J.E. Sternberg

Biomolecular Modelling Laboratory
Imperial Cancer Research Fund
P.O. Box 123
Lincoln's Inn Fields
London W.C.2 3PX, UK

David L. Weaver

Department of Physics
TUFTS University
Medford
Massachusetts 02155
USA

SUMMARY

It has become common practice amongst workers in protein modelling, protein databases and general analysis of protein structures to use the resolutions and R-factors of protein structures as the basis for the critical selection of working sets of data. These criteria are taken as indicators of accuracy and reliability. One may ask: is the extent of observable errors occurring within protein structures related to their resolutions and R-factors? In this paper it is shown that the extent of bad contacts (violation of van der Waal radii), assumed to be indicators of errors occurring within data, are correlated with the resolutions of protein structures. However errors within individual proteins at a given resolution can show large deviations. It is also shown that there are "accurate and reliable" structures (based on the resolution and R-factor criteria) which contain larger observable errors than "less accurate" structures. Correlation of errors with the R-factors is less obvious. This suggests that we (as non-crystallographers) need a better understanding and more reliable definitions relating to the errors occurring within protein x-ray crystallographic structures.

INTRODUCTION

The Brookhaven protein databank¹ (PDB) contains data for approximately 350 x-ray crystallographic protein structures (there also a few neutron diffraction crystal structures). Many of these protein data sets have identical sequences and sometimes vary only in the bound ligands. Critical selection of data e.g. selection of the "best" protein from a set, is a very common practice. As with all crystallographically related data, these structures contain systematic and random errors. Our basic problem lies in the fact that accurate errors in the atomic positions of the atoms are not usually available or cannot be obtained at all from the crystallographic data. Errors in the atomic positions would allow us to calculate errors in geometrical measurements e.g. distances, angles, and particularly errors in the proteins **internal coordinates**² i.e. **bond lengths, angles and torsion angles**. Estimates of errors in torsion angles^{2,3} would be of the most practical importance.

It has become common practice in many areas of research (eg. modelling by homology, construction of databases, general conformational analysis) to use the protein **resolution** as a criterion for selecting a working set of structures. In addition some workers also take into consideration the structural refinement by considering the **R-factor** of the data set. In practice proteins with high resolution (< 2.1 Å) and low R-factors (< 25%) seem to form, at least qualitatively, the definition of "accurate or reliable structures". Although we do not know the precise errors in atomic positions we can perhaps obtain estimates by measuring quantities which may reflect these errors. Two such quantities, which are of practical importance, are close contacts between atoms and deviations from standard geometry of the internal coordinates. Bond lengths and bond angles tend to be generally well behaved and remain within ranges observed from small molecule x-ray crystallography. Torsion angles show the most variation, but since these quantities are also by far the most flexible, it is difficult to distinguish between genuine deviations and errors. Errors in the internal coordinates are under investigation by Thornton et al.⁴. Thornton et al have noted that a number of well refined high resolution proteins structures contain some bad torsion angles (namely, the phi and psi angles are found to occur in disallowed regions). It would be of interest to see if there are well refined high resolution proteins which contain relatively large errors.

In this work the following questions will be addressed:

- a) **What is the extent of bad contacts (violation of van der Waal radii) in protein structures and can we quantify this for comparative purposes ?**
- b) **Is the extent of bad contact related to the protein resolution and R-factor ?**

We need a suitable method for measuring the extent of bad contacts occurring within protein structures so that different structures can be compared using a standard scale. One may consider using intramolecular

energies as a measure but this would not be practically possible, since this quantity is too sensitive (i.e. a single bad contact can alter the energy by many orders of magnitude) especially for use in non-related structures. The dependence of experimental error upon resolution in crystallographic data has been noted by Hubbard and Blundell⁵, by the study of rms differences between proteins structures having the same sequences but having been independently refined.

METHOD

Our method essentially involves looking at all non-bonded atom-atom distances within each protein with the following exceptions:

- a) distances within residues and between adjacent residues ,
- b) between potential hydrogen bonding groups and
- c) between disulphide bridge forming residues.

All atom-atom distances calculated were compared with the sum of the van der Waal radii of the individual atoms. The van der Waal radii used were taken from Kollman et al⁶. United atom radii were used (i.e. hydrogen atoms implicitly included). It should be noted that calculations were also performed using the CHARMM⁷ parameters to ascertain possible dependence of results on data used (results not presented in this paper). However, no significant differences were found in the results between Kollman et al⁶ and the CHARMM⁷ data. Each atom-atom distance calculated was checked to see if it was shorter than the sum of the van der Waal radii of the individual atoms. If the distance was shorter than the sum of the van der Waal radii then it was noted. We define the "proteins bad length" , **L**, by :

$$L = \sum_N (R_{vi} + R_{vj}) - R_{ij} \quad - \quad 1$$

where **R_{ij}** is the distance between atoms i and j,
R_{vi} is the van der Waal radii of atom i,
R_{vj} is the van der Waal radii of atom j and
N total number of atom-atom contacts for which
R_{ij} < (R_{vi} + R_{vj}).

We can further define, for a given protein containing **NRES** residues:

The average error per residue,	Er	= L/NRES	-	2
The average error per contact,	Ec	= L/N	-	3
The average error per contact atom	Ea	= L/(2N)	-	4

Er, Ec and Ea all have the dimension of length. **Ec** represents the errors in the distance between two atoms and **Ea** may be seen as the error in the position of an atom i.e. distance from "true" position". To obtain an estimate in the errors in the coordinates (i.e. x,y and z) of an individual position we note that :

$$Ea = (\Delta x^2 + \Delta y^2 + \Delta z^2)^{1/2} \quad - \quad 5$$

Where $\Delta x, \Delta y$ and Δz are shifts along the x,y and z axis from a "true position".

If we assume further that $\Delta x = \Delta y = \Delta z$, then the mean error in an atomic coordinate, E_x is given by:

$$E_x = E_a/\sqrt{3} \quad - \quad 6$$

E_x can be used to estimate errors in geometrical parameters and we will use this quantity to estimate errors in torsion angles.

313 proteins were selected from the March 1989 issue of the PDB. All model built proteins and proteins containing backbone atoms only were excluded. Of the 313 protein files, only 209 provided R-factors whereas the remaining 104 files did not provide this information. The 104 protein files which did not state R-factors included protein structures which were clearly noted as not having been refined with the rest simply not stating the R-factors although some form of refinement had been indicated.

RESULTS AND DISCUSSION

Figure 1 shows the variation of mean error per residue (equation 2 above) as a function of resolution (data shown in fig 1 is not presented in this paper). Two points are apparent from this graph. Firstly, one notes that overall, the error per residue increases with decreasing resolution. Secondly, for all given resolutions there is a large amount of scatter and in fact some high resolution structures have greater error/residue than many of the low resolution proteins. Since figure 1 includes all proteins (including non-refined), one is tempted to ask whether or not the outliers are non-refined or have high R-factors. Data for some of the proteins is given below :

Protein	Resolution	R-factor (%)
1NXB	1.38	24.0
2SNS	1.50	NOT GIVEN
1HDS	1.98	NOT GIVEN
2GN5	2.30	21.7
3PGK	2.50	NOT REFINED
3PGM	2.80	29.0
1PYP	3.00	Diamond S program

It can be seen from the above sample data set that some of the proteins structures are of high resolution and have been refined to a low R-factor. Detailed results will be presented elsewhere⁸.

To simplify the results shown in figure 1 the mean error per residue was averaged for all proteins at each given resolution. The results are shown in figure 2 and the data is presented in Table 1. The results have been further divided into those for proteins with R-factors given and for those with no R-factors specified (figure 2 and Table 1). Least squares fits were performed assuming exponential and linear relationships (only for proteins with R-factors given). The average error per residue (E_r) can be related to the resolution (R)(figure 2) by assuming an exponential relationship (cc is the correlation coefficient) :

$$E_r = 0.031 * 10^{(0.488 * R)} \quad (cc=0.952) \quad - \quad 7$$

Assuming a linear relationship (line not shown in figure 2):

$$E_r = -0.625 + 0.507 R \quad (cc=0.843) \quad - \quad 8$$

The least squares fit for equation 7 is very good for the proteins with R-factors given. It can be seen from figure 2 that proteins which do not have R-factors given generally have a higher error per residue than the proteins with R-factors given.

Figure 3 (data in Table 1) shows the variation of "mean error in atomic position" (E_a) (equation 4 above) as a function of resolution. As with figure 2 the results have been divided for proteins with and without R-factors (R) given. The relationship between E_a and R (figure 3) is:

$$E_a = 0.202 * 10^{(0.067 * R)} \quad (cc=0.851) \quad - \quad 9$$

and assuming a linear relationship (line not shown in fig 3):

$$E_a = 0.185 + 0.046 R \quad (cc=0.819) \quad - \quad 10$$

The results shown in figure 3 compare well with those of Hubbard and Blundell⁵ (figure 4 from their work). Since the range of "atomic" errors extends to approximately 0.35 Å (least squares fit, figure 3), the error in distance between two atoms would be approximately 0.7 Å.

Using equation 6 and the results shown in equations 9 and 10, we can estimate the mean error in the atomic coordinates, E_x (assuming exponential and linear relationships respectively) as :

$$\begin{array}{l} E_x = 0.117 * 10^{(0.067 * R)} \quad - \quad 11 \\ \text{or } E_x = 0.107 + 0.046 R \quad - \quad 12 \end{array}$$

For a range of resolution values of 1.0 to 3.5 the estimate of errors in atomic coordinates is 0.135 to 0.200 (equation 11) and 0.153 to 0.268 Å (equation 12). To understand what this would mean in terms of errors in torsion angles, figure 4 shows relationship between the error in a given torsion angle and the error in the atomic coordinates (the error in torsion angle was calculated using reference 3). From figure 3 it can be seen that an approximate relationship between the error in torsion angle (E_t) and error in atomic coordinate (E_x) is given by:

$$E_t = 115.6 * E_x \quad - \quad 13$$

Using equation 13 (we can substitute equations 11 or 12 into equation 13) the errors in torsion angles for resolutions in the range of 1.0 to 3.5 as being approximately 17° to 30°. However since the error in some high resolution proteins is high, it is perhaps not surprising that torsion angles can be found in disallowed regions.

The mean error per residue divided by the resolution⁹ is shown as a function of the R-factor in figure 5. Although a cluster is found for R-factors in the range of 15-20 %, some proteins with low R-factors show a large deviation. The least squares fit to the data (figure 5) has a correlation coefficient of only 0.310, so a dependence of error on resolution is negligible.

CONCLUSIONS

The quantitative relationships derived in this work show a correlation between errors and resolution but a dependence on the the R-factors is more difficult to ascertain. These correlations must only be seen as approximate guides, since individual proteins can show large deviations. If the errors measured in this work are truly representative of the systematic and random errors occurring within protein structures, then one needs to be able to explain why a number of "well resolved and refined" proteins contain relatively large errors. It would be of interest to look at refinement methods and procedures in detail. Further work is in progress⁸.

The purpose of this work has not been in any way to criticise the work of protein crystallographers , but merely to show that the definitions used by a vast majority of people (mostly non-crystallographers) as indicators of reliability and accuracy are inadequate or misunderstood.

ACKNOWLEDGEMENT

We would like to thank Drs Ian Tickle and Paul Freemont for useful discussions and comments.

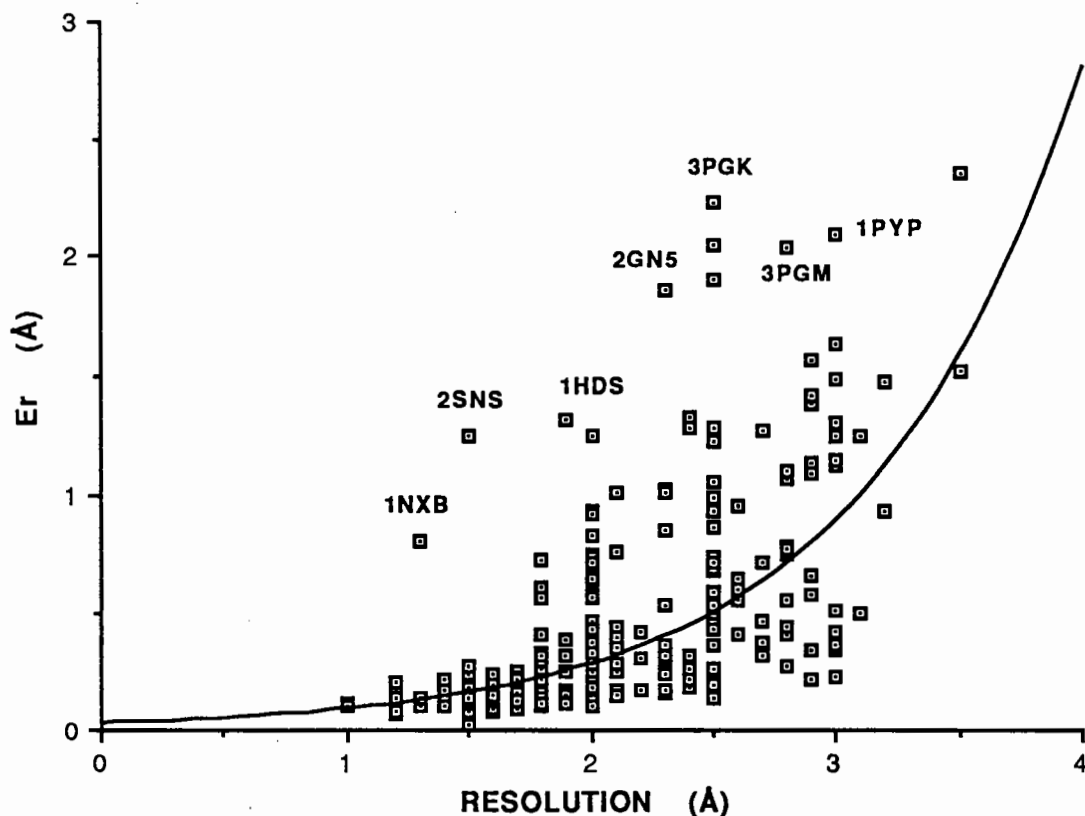


Fig 1 The mean error per residue (equation 2 in text) vs. resolution (all 313 proteins have been included).

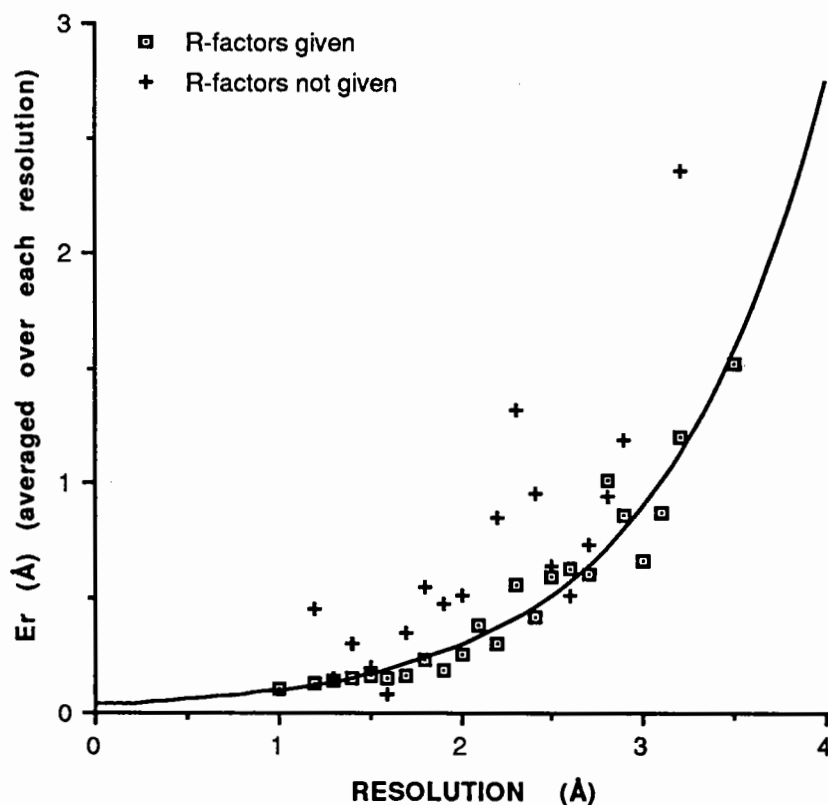


FIG 2 The mean error per residue at a given resolution vs. resolution. Equation of the line (for proteins with R-factors only) is $y = 0.031 \cdot 10^{(0.488 \cdot x)}$. Data for graph is shown in Table 1.

TABLE 1 (a) Errors for proteins with R-factors given. Data has been averaged for all proteins at a given resolution. Er is the mean error per residue and Ea mean error per contact atom.

Resolution (Å)	No. of proteins	Er (Å)	Ea (Å)
1.0	2	0.110	0.242
1.2	5	0.130	0.253
1.3	1	0.140	0.266
1.4	3	0.148	0.260
1.5	11	0.164	0.257
1.6	18	0.154	0.253
1.7	35	0.169	0.252
1.8	28	0.239	0.266
1.9	14	0.186	0.260
2.0	21	0.257	0.264
2.1	6	0.381	0.289
2.2	3	0.301	0.268
2.3	11	0.563	0.285
2.4	6	0.420	0.275
2.5	18	0.598	0.304
2.6	4	0.629	0.303
2.7	4	0.611	0.308
2.8	3	1.010	0.321
2.9	2	0.860	0.298
3.0	9	0.662	0.305
3.1	2	0.873	0.325
3.2	2	1.201	0.345
3.5	1	1.521	0.390

TABLE 1 (b) Errors for proteins with no R-factors given. Data has been averaged for all proteins at a given resolution

Resolution (Å)	No. of proteins	Er (Å)	Ea (Å)
1.2	2	0.455	0.299
1.3	5	0.150	0.255
1.4	7	0.307	0.271
1.5	2	0.203	0.263
1.6	1	0.087	0.243
1.7	3	0.348	0.265
1.8	3	0.551	0.286
1.9	37	0.474	0.297
2.0	3	0.509	0.300
2.2	1	0.849	0.333
2.3	1	1.321	0.365
2.4	13	0.952	0.331
2.5	1	0.642	0.327
2.6	2	0.510	0.302
2.7	9	0.740	0.317
2.8	7	0.948	0.339
2.9	6	1.186	0.339
3.2	1	2.356	0.423

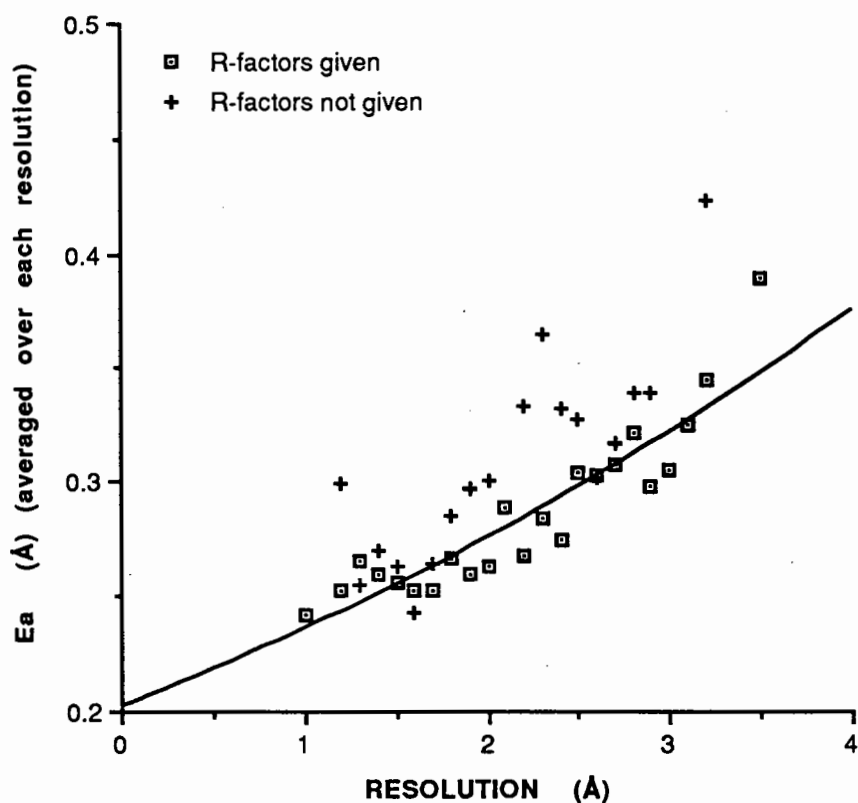


FIG 3 The average error per contact atom (equation 4) vs. resolution for the 209 proteins with R-factors given. The equation of the line (for proteins with R-factors only) is $y = 0.202 \cdot 10^{(.067 \cdot x)}$. Data for graph is shown in Table 1.

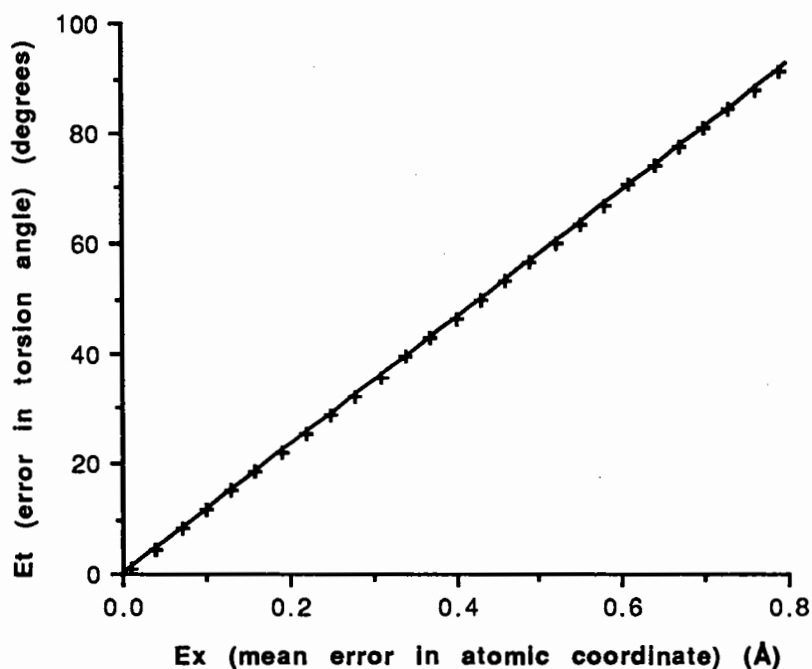


Fig 4 Error in a torsion angle of -60° (defined by the psi angle of a protein back bone with standard geometry) as a function of the mean error in atom coordinates i.e. each x,y and z of the 4 atoms defining the torsion angle has been assigned the same error (E_x). The least squares fit is $E_t = 115.6 \cdot E_x$ (correlation coefficient = 1.000).

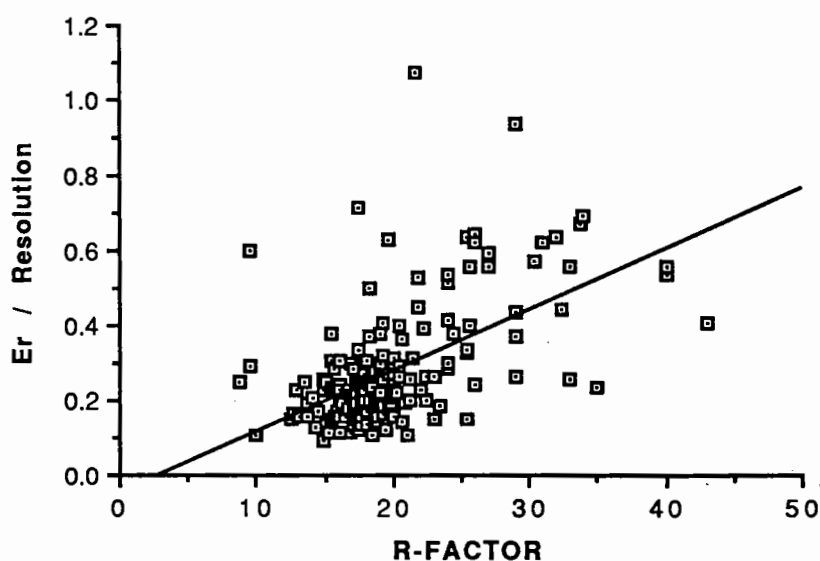


Fig 5 The average error per residue divided by the resolution vs. R-factor (given as %). Equation of line is $y = -0.05 + 0.02x$ (correlation coefficient = 0.310).

REFERENCES

1. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rogers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M.
J. MOL. BIOL. 144 PP.441-484 (1977).
2. Dunitz, J. D. Chapter 5, "X-Ray Analysis and the Structure of Organic Molecules". Cornell University Press. (1979).
3. Stanford, R.H., Wasser, J.
ACTA CRYST. A28 pp. 213-215 (1972).
4. Thornton, J. Analysis presented at meeting and in proceedings of the meeting.
5. Hubbard, T.J.P., Blundell T.L.
PROTEIN ENGINEERING vol 1 no. 3 pp.153-171 (1987).
6. Welner, S.J., Kollman, P.A.
J. AM. CHEM. SOC. 106, p 765 (1984).
7. Brooks, B.R., Brucoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M.
J. COMP. CHEM. vol 4, no. 2, pp187-217 (1983).
8. Islam, S.A., Sternberg, M.J.E., Weaver, D.A. Manuscript In preparation.
9. Dr. Ian J. Tickle (Birkbeck college). Private communication. Since the R-factor is a dimensionless quantity, dividing the error/residue by resolution makes the y-axis also dimensionless.

STRUCTURE DETERMINATION OF TURKEY EGG LYSOZYME

S.E.V. Phillips, W.S. Somers, T.N. Bhat¹ and M.R. Parsons²
Astbury Department of Biophysics, University of Leeds,
Leeds LS2 9JT, U.K.

Introduction

The re-evaluation of the crystal structure of turkey egg lysozyme (TEL) was undertaken at Leeds in the spring of 1987 (Parsons, 1988). The impetus was provided by the observation that the monoclonal anti-hen egg lysozyme antibody D1.3 shows exquisite fine specificity for various closely related avian lysozymes, such that single amino acid differences in the lysozyme antigen cause large changes in affinity. The determination of the crystal structure of the Fab fragment of D1.3 in complex with hen lysozyme (HEL) (Amit et al, 1986) revealed a large area of interaction between the two molecules, and a topographic antigenic site consisting of residues 18 to 27 and 116 to 129 of HEL. D1.3 binds hen lysozyme with an affinity constant of 10^8 M^{-1} , and also that of bobwhite quail, which differs in four amino acid residues, with similar affinity. Other closely related avian lysozymes such as California quail (4 sequence differences from HEL), pheasant (3 differences) and TEL (7 differences) do not show measurable binding to D1.3. Analysis of the sequences shows that replacement of Gln121 of HEL by His is sufficient to abolish antibody binding, consistent with the crystal structure of the complex where Gln121 is lodged in a tight pocket on the antibody surface in the middle of the combining site. Due to the similarity of Gln and His side chains, the explanation of the large affinity change was not immediately obvious, although careful energy calculations have shown that it could be accounted for (Haneef, 1990). However, in order to be certain the affinity difference is solely due to interaction of the different side chains with the antibody, it was necessary to show that the Gln->His121 replacement did not change the structure of the free antigen, either in the main chain conformation or by His forming a salt bridge to the neighbouring Asp119. Since TEL has His121, with the other six sequence differences from HEL lying outside the antigenic site, and its structure had been solved and deposited in the Brookhaven Protein Data Bank, it was chosen for study.

A redetermination of the turkey lysozyme structure was also undertaken independently by L. Howell and G. Petsko at MIT using Laue techniques as a prelude to kinetic studies of the catalytic mechanism.

Original Structure Determination

The original structure determination of TEL was carried out by Bott and Sarma (1976) using molecular replacement. The structure of HEL, both intact and with all side chain atoms beyond C β removed, was used in rotation and

¹Immunologie Structurale, Institut Pasteur, 28 rue du Dr. Roux, 75724 Paris Cedex 15, France.

²Present address: Department of Molecular Biophysics and Biochemistry, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA

translation searches carried out with data to 6Å resolution. The translation searches, using the strongest terms only, showed many equivalent R factor minima, so packing considerations were used to select the best solution. The model chosen gave R=47% at 6Å, with four intermolecular contacts of less than 5Å between α -carbons. This model was used to phase an electron density map at 5Å resolution, which showed some, but not all, of the expected disulphide bridges. The resolution was extended to 2.8Å (Sarma and Bott, 1977), and the model rebuilt to give R=45% for data 10-2.8Å. It was used to phase an isomorphous difference map for a platinum derivative, and peaks were found for sites in reasonable locations relative to the protein molecule. Substrate binding studies were also carried out. The C α coordinates of the structure were deposited in the Brookhaven Protein Data Bank as entry 1LZ2.

Correcting the Published Structure

High resolution refined coordinates were not available from the original authors, so the structure refinement was undertaken independently using the Data Bank co-ordinates as a starting point.

New crystals were grown under similar conditions to those originally reported, but using a double dialysis technique, yielding hexagonal rods up to 1.2 mm long. Crystal data are shown in Table 1.

Table 1

Crystal Data

Crystallization conditions: 100mM ammonium acetate pH 7.0
18% w/v NaCl
10 mg ml⁻¹ protein (Sigma)

Space group P6₁22 with a=b=71.0 c=84.9 Å, z=12.

Native data: Rotation films collected at Daresbury SRS for 30° rotation about c, merged to 6Å CAD4 diffractometer data. 6414 unique reflections to 2.2Å with R_{merge}=0.075.

Refinement: CCP4 FFT Hendrickson-Konnert
R=19.3% for all data 10-2.2Å, RMS bond dev=0.028Å
994 protein and 110 solvent atoms

The space group and unit cell parameters were confirmed as those previously observed, and a preliminary 3Å data set collected on a CAD4 diffractometer. A model of TEL was built from a 1.6Å refined structure of HEL (originally obtained from Peter Artymiuk), replacing or truncating the seven side chains that differed between the two. The crystallographic R factor was 60% for this model, and the 2F_o-F_c map was of suspiciously poor quality, although main-chain density was reasonably well connected, and side chains were often visible. A high resolution data set was collected on film at Station 9.6 at the Daresbury SRS, and processed using the Imperial College version of MOSFLM, to the maximum observed resolution (2.2Å) with a merging R of 7.5% (Table 1). With the new data R remained at 60% to 2.8Å, and could not be reduced below 50% by rigid body minimization. Concerned about the

correctness of the model, we decided to redetermine the structure independently.

To avoid possible problems of false solutions for the molecular replacement in a high symmetry space group that might have misled the previous workers, we chose to use heavy atom derivatives at low resolution to calculate experimental phases. 6\AA data sets were collected on a CAD4 diffractometer, for a native crystal, and two derivatives, K_2PtCl_4 and $\text{K}_3\text{UO}_2\text{F}_6$ respectively. The difference Patterson maps were very complex, due to the existence of multiple sites and the high symmetry, and proved impossible to solve. The anomalous data were of especially high quality for the uranyl derivative, and were used to calculate F_{HLE} coefficients, which were then used in the SHELX direct methods program (Sheldrick, 1976,1986). The best E-map clearly showed two uranyl sites as the only strong peaks. F_{HLE} refinement of these sites gave $R=27\%$, and SIRAS phases were calculated with a mean figure of merit $\langle m \rangle = 0.78$. Uranyl phased difference maps for the platinum derivative showed three sites, which refined to $R=53\%$ with F_{HLE} . The space group was confirmed as $P6_122$, and the enantiomorph determined. MIRAS phases gave $\langle m \rangle = 0.83$, and the native map was of exceptional quality, showing clear rods of density corresponding to α -helices. Using FRODO (Jones, 1985) the $\text{C}\alpha$ backbone of HEL was fitted to the map as a rigid body. The first attempt to do this resulted in a reasonable fit, with most of the backbone in density, which proved to be equivalent to the original TEL crystal structure. A second attempt resulted in a better fit, where the entire backbone was in good density, with the molecule rotated by about 180° relative to the first solution. This proved to be correct. The crystal packing diagrams for these two models are shown in Figure 1. In the correct model (a) a β -hairpin loop can be seen extending from the molecule away from the unique 6_1 axis, whereas in the original solution (b) it points inwards towards the axis.

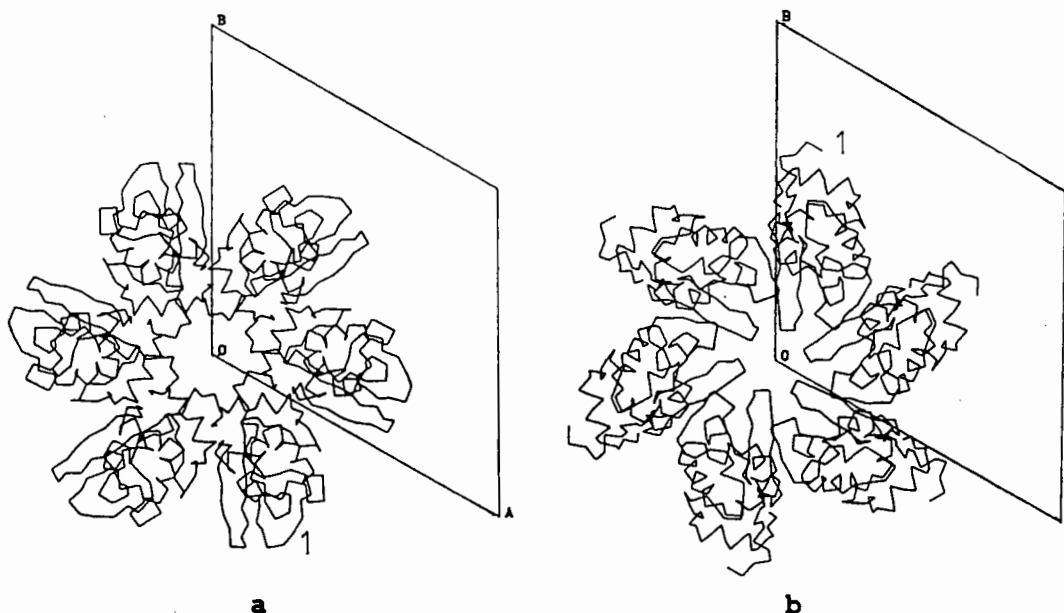


Figure 1: Crystal packing diagrams for correct (a) and incorrect (b) structures, viewed down the 6_1 axis. Only six of the twelve molecules per unit cell are shown, the other six being related to these by two-fold axes perpendicular to the 6_1 axis.

A 3\AA $2F_o - F_c$ map was calculated for the new model, and was seen to be of better quality than the map calculated with the previous one. Following minor rebuilding, the structure was subjected to FFT Hendrickson-Konnert (H-K) refinement at 3\AA , and the resolution extended to 2.2\AA with cycles of refinement and model building. The final R factor was 19% for all data from $10\text{-}2.2\text{\AA}$, with good stereochemistry. The co-ordinates have been deposited in the Data Bank with code 2LZ2.

Comparison of Correct and Incorrect Structures

The relationship between the correct and incorrect structures is interesting in two ways, both related to the approximate 180° rotation between them. Firstly, Figure 2 shows a comparison of the $C\alpha$ backbones, and it can be seen that some of the helical regions approximately coincide with each other, or lie on top of β -sheet. The overall effect of the rotation is to exchange the two domains which lie to either side of the active site cleft. The observation that both models make a reasonable fit to an experimentally phased map immediately suggests that molecular replacement might also be ambiguous. This could be thought of as corresponding to an ambiguity in the rotation function solution.

Secondly, the space group has two-fold axes perpendicular to the six-fold, which relate the six molecules missing from Figure 1 to the ones shown. These two-folds approximately coincide with the rotation axis between the two structures in Figure 2, so the incorrect molecule is in an orientation which would be approximately correct if it occupied a different position in the unit cell. The actual error in orientation would then be 17° , with an additional error in the translation function solution. In this case the molecular transform of the incorrectly oriented molecule is similar to that of a correctly oriented one elsewhere in the cell. If the diffraction pattern is considered as resulting from two contributions, scattering from the individual molecules and scattering resulting from the arrangement of the molecules in the cell, the former contribution would be similar for both models. This would lead to ambiguity in molecular replacement and, indeed, refinement calculations, since the transform of the unit cell of the incorrect solution would be partially correct, and could lead to unexpectedly low R factors.

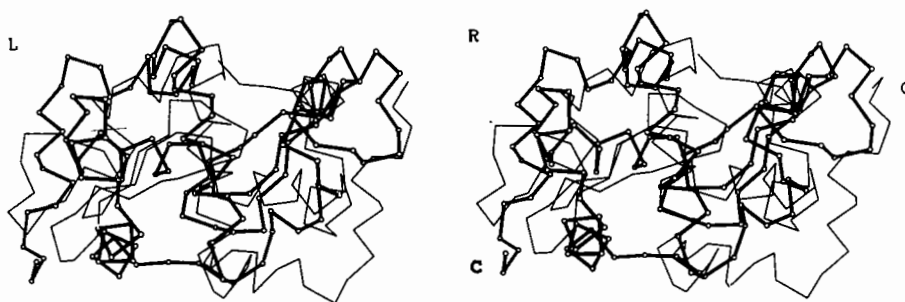


Figure 2: Stereo diagram of the α -carbon backbones of the correct (ball and stick) and incorrect (thin lines) starting structures. In the correct structure the active site cleft is uppermost, while the incorrect structure is rotated by approximately 180° about an axis approximately perpendicular to the paper.

We therefore asked ourselves whether the incorrect model could be refined

to a low R factor at medium resolution while retaining reasonable stereochemistry, and what indicators could be used to distinguish the correct solution. From the experience of the original workers it seemed that neither inspection of the initial map (without the benefit of modern computer graphics), nor phasing heavy atom derivative difference maps and checking the resulting sites for consistency could be absolutely relied upon.

The correct and incorrect structures were refined using XPLOR (Brünger et al, 1987), starting from our 6Å map fitted model, and the Data Bank derived model respectively. The results are shown in Table 2. The correct starting model refines well at 3Å, rapidly dropping to R=21% after 4 cycles of simulated annealing, with tightly restrained stereochemistry (model III). The incorrect model also refines well at 3.5Å, reaching R=23% on the first cycle, and falling to 21% after 4 more cycles, but with poorer stereochemistry (although no worse than is often tolerated at this stage). Extending the resolution to 3Å, and restraining the stereochemistry more strongly raised R to 31% (model II), but the R factor seems an insensitive measure of the correctness, bearing in mind that one does not normally know how well the correct solution would refine if only one had found it.

Table 2

Comparison of refinement of correct and incorrect structures

<u>Model</u>	<u>Bond Devⁿ</u>	<u>Resolution(Å)</u>	<u>R(%)</u>
Incorrect models:			
Sarma and Bott (1977), as reported in paper - unrefined model	-	2.8	45
Model (I) built from fit of hen lysozyme to Protein Data Bank CO, new SRS data	-	3.0	53
As above after 1 pass XPLOR 7-3.5Å	.036	3.5	23
" " " 5 passes " "	.036	3.5	21
Final XPLOR model (II) 7-3Å	.025	3.0	31
Same model, CCP4 SF calculation	"	3.0	37
Correct models:			
Initial fit of hen based model to 6Å MIR	-	3.0	49
As above with rebuild and 5 cycles H-K (on xyz)	.040	3.0	33
Final model (IV)	.028	2.2	19
Initial model after 4 passes XPLOR (III)	.013	3.0	21
Same model, CCP4 SF calculation	"	3.0	25

Figure 3 shows the electron density maps for the two models at various stages. (b) and (c) show the correct and incorrect starting maps. although (b) is better, the difference is not as great as one might naïvely expect. This is, of course, due to the feedback inherent in maps with phases calculated from a model. In fact it is well known that a map calculated with model phases, but with structure amplitudes all set equal, or even given random values, still gives a good image of the starting model (see Ramachandran and

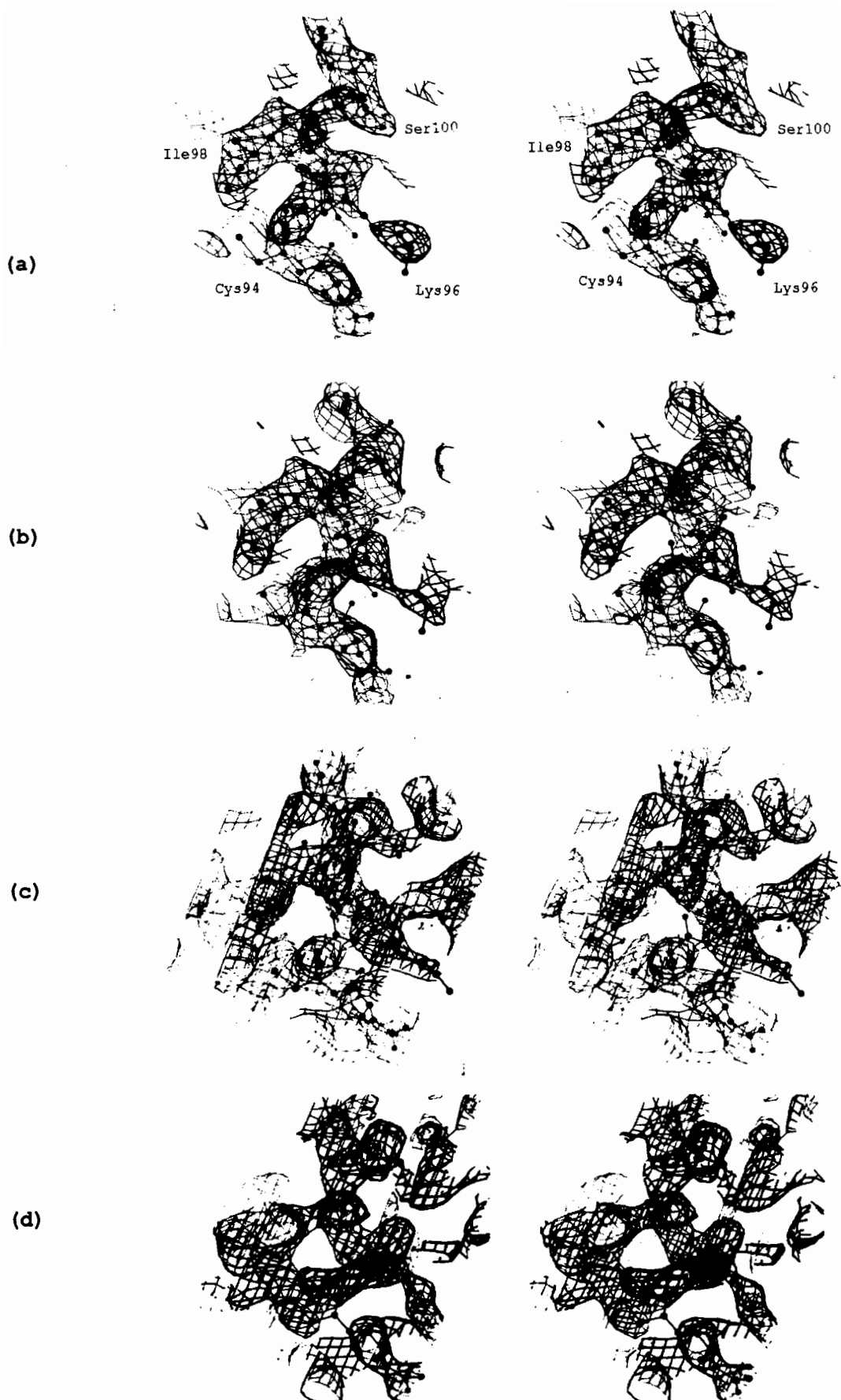


Figure 3: Stereo diagrams of $(2F_o - F_c)O_{calc}$ maps at 3Å resolution for residues 93-101. A high contour level is shown for clarity, at the expense of loss of some continuity. Residues 93-98 form the end of a helix, with a disulphide bridge branching off at Cys94.

(a) Final model (IV) in final map. (b) Final model (IV) in map phased from initial model (I) in correct orientation. (c) Wrong initial model in its own map. (d) XPLOR refined wrong model (II) in its own map.

Srinivasan, 1970). The map calculated from the final refined model (a) is necessarily better, but the incorrect XPLOR refined map in (d) is of good quality for a partially refined structure. The latter map shows good side chain density, for instance Lys 96 at the lower right, and even has clear indications of the main chain carbonyl groups. What gives it away, however, is the geometry. The bond lengths and angles are reasonable, of course, since they have been restrained, but the hydrogen bonding no longer looks good, and the helix is breaking up. The torsion angles for main and side chains do not fit with expected values from known highly refined protein structures. This can be detected by making Ramachandran plots for the main chain, and comparing side chains to a conformational data base such as that published by Ponder and Richards (1987). Figure 4 shows Ramachandran plots for various models. That for the starting model is very good, since it is based on a 1.6Å refined HEL

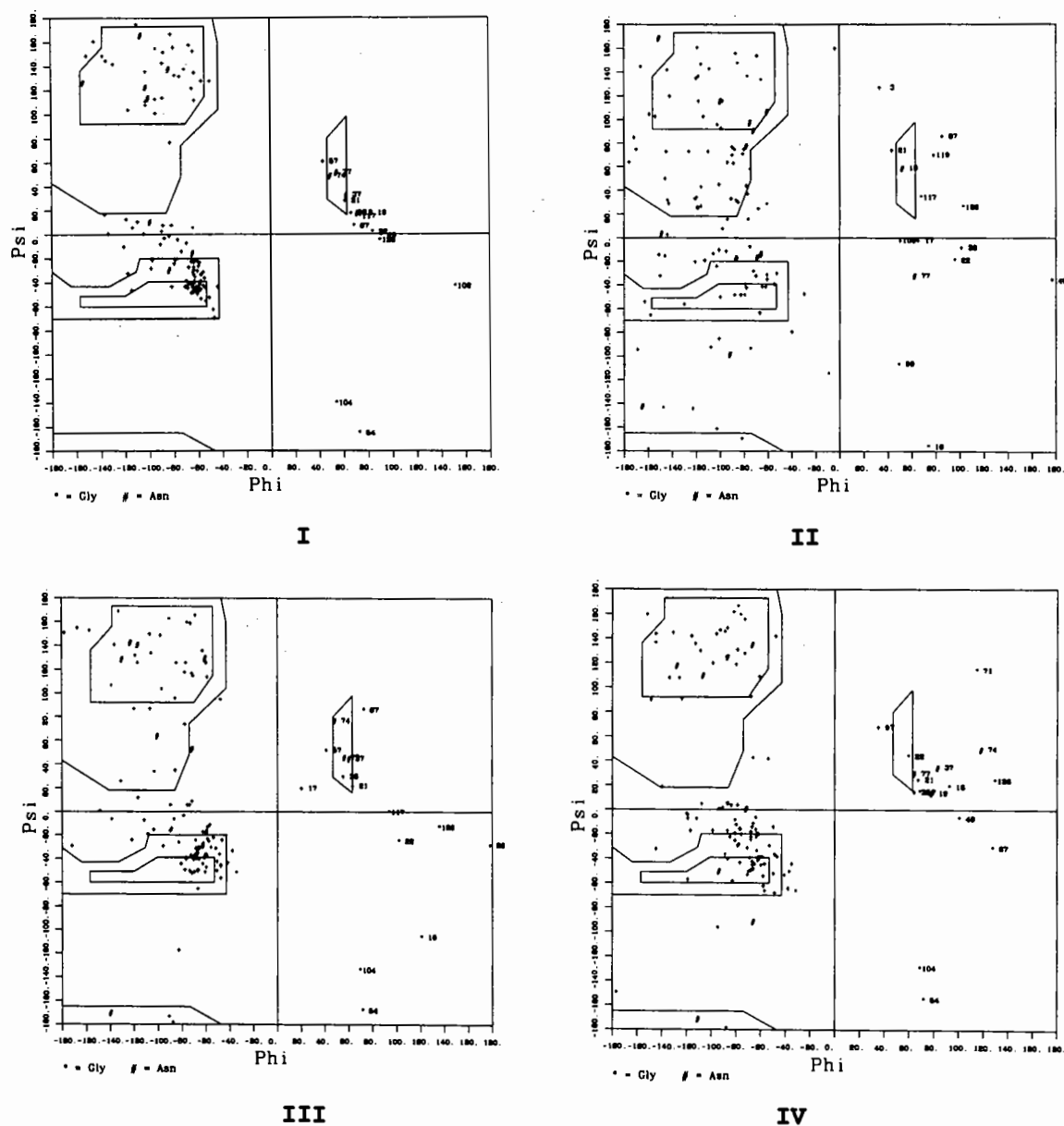


Figure 4: Ramachandran plots for the various models.

I initial model built from hen lysozyme.

II wrong model after XPLOR refinement.

III correct model after XPLOR refinement.

IV final 2.2Å H-K refined model.

structure, and the final 2.2Å refined TEL structure is slightly worse, reflecting the lower resolution. In the latter there are one or two residues in the disallowed regions, and the clustering of points in the regions of regular secondary structure is looser, especially in the α -helical region. The XPLOR refined correct model has a good plot, with tighter clustering than the final model, presumably due to hydrogen bonding geometry restraints. The incorrect refined model is poorer, with more scattered points, though not in itself so bad as to definitely indicate an incorrect solution. Many structures at intermediate resolution and refinement may give such a plot, but what is probably diagnostic of a serious problem is the break up of the clustered points in the α -helix and β -sheet regions.

Such diagnostic tests are required since at medium resolution, 3Å for instance, electron density maps, but especially R factors, are insensitive indicators of correctness. It is very common for crystallographers to pursue the lowest possible R factor as if that were the main aim of the structure determination. It is not, of course, and the aim should be to make the best possible atomic model and use it to try to learn something about the protein's function. This is often forgotten, and tricks such as leaving out all the weak data are used, solely to get a lower R factor. It can, however, be shown that omitting data less than 2 or 3 σ produces less accurate atomic models, even though the R factors might be lower (Arnberg et al, 1979). In fact, some published small molecule structures were shown to refine successfully using only the weak data the authors chose to omit in order to reduce their R factors. In resolution ranges where the majority of the reflections are significant, the weak reflections provide good observations for the refinement.

A further feature of medium resolution protein refinements is that they are underdetermined. This is well known, but the problem is often not treated with the respect it deserves. Figure 5 shows the number of experimental observations available per parameter for TEL at various resolutions. For a unique solution the observation/parameter ratio must exceed 1:1, and this

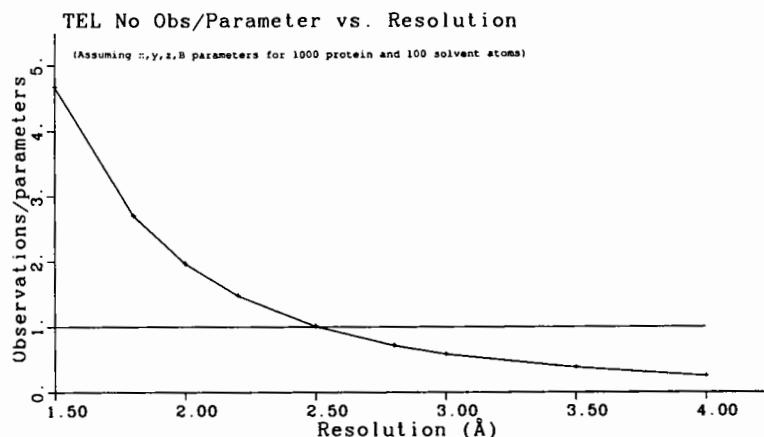


Figure 5: Plot of number of observations per parameter with resolution for turkey lysozyme, assuming x,y,z,B parameters for 1000 protein and 100 solvent atoms. No account is taken of the effect of stereochemical restraints which increase the effective number of observations but do not directly depend on the experimental data. The plot would be similar for other protein crystals of average solvent content. Note that the ratio only reaches 1:1 at 2.5Å resolution.

condition is not fulfilled until the resolution is better than 2.5Å. This is, of course, circumvented by the use of stereochemical restraints, which effectively add observations to the refinement, but these observations do not depend directly on the crystal structure and serve only to keep the model self-consistent. In fact, one can refine the model successfully at any position in space using only the restraints, corresponding to simple molecular mechanics or dynamics.

What Went Wrong?

We felt it was important to understand why the original structure solution by molecular replacement had not yielded the correct model. Bott and Sarma carried out two series of rotation searches, using the whole HEL molecule as a probe and using a model with all side chains removed beyond C β . The probes were placed in a triclinic cell with $a=b=c=120\text{\AA}$ and angles of 90° . The searches were carried out in increments of 10° initially, with fine searches of 5° around possible solutions, using only the strong intensities in the 12-6Å resolution range. We repeated the calculation using the POLARRFN program, but necessarily had to guess the intensity cutoffs used. Our searches produced at least six independent peaks with heights greater than 90 (scaled to 100 for the highest peak). The correct solution was present amongst them (not as the highest peak) but the originally reported solution was not. When we used the resolution range 8-3Å, the correct solution was clear in both rotation and translation searches.

Since the success of the translation function depends crucially on the accuracy of the molecular orientation, it is not surprising the original authors found no good solution to it, and had to rely on packing considerations. This is clearly dangerous as there can be many solutions to the packing problem, some of which are likely to be systematically related to the correct one and could give plausible results. It is not clear why the original rotation function solution could not be reproduced, but it could be due to selection of a limited number of strong terms, or possibly to errors in programs. We have found errors in widely used molecular replacement programs when working in high symmetry trigonal and hexagonal space groups.

Conclusions

There is no overall guiding principle for avoidance of errors in protein crystallography, except to take great care at every stage of the process and be sure to understand the underlying principles and the limitations of the data. In particular, the R factor should not be taken as the best measure of success. A low R is a necessary, but not sufficient condition, for a correct structure. It is more important to examine electron density maps critically, and check that the geometry of the model is acceptable. Bond lengths and angles are usually restrained in refinements, but both main and side chain torsion angles should lie almost exclusively in ideal conformations, except in particular cases where a good reason for a deviation can be found (eg. a strong hydrogen bond stabilizing an otherwise less favourable conformation).

Modern data collection and refinement technology has made life much easier, but one should always use "Black Box" programs and systems with caution. These are excellent servants but poor masters. The incorrect TEL structure could be refined quite well with XPLOR, but there were clues that something was not quite right which could have been used to detect the problem. The shortage of

experimental data for structures determined at less than 2.5Å compounds the problem, and extra care is needed in these cases. Our experience suggests that it becomes much easier to detect serious errors as the resolution approaches 2Å. When tackling structures in unusual, high symmetry space groups, where standard programs may not have been fully tested, it is always wise to run trial calculations with known cases to check there are no bugs, and to get the feel of the expected results.

Finally, one should never absolutely assume that someone else's program, or coordinates from the Data Bank are necessarily correct if one has any cause for suspicion.

Acknowledgements: We thank the SERC for studentship awards to MRP and WSS, and for use of the Daresbury SRS for data collection.

References

Amit, A.G., Mariuzza, R.A., Phillips, S.E.V. and Poljak, R.J. (1986), *Science*, **233**, 747.

Arnberg, L., Hovmöller, S. and Westman, S. (1979), *Acta Cryst.* **A35**, 497.

Bott, R. and Sarma, R. (1976), *J. Mol. Biol.* **106**, 1037.

Brünger, A.T., Kuriyan, J.K. and Karplus, M. (1987) *Science* **235**, 458.

Haneef, I. (1990), *J. Mol. Graphics*, **8**, 45.

Jones, T.A. (1985), *Methods in Enzymology* **115**, 157.

Parsons, M.R. (1988), PhD Thesis, University of Leeds.

Ponder, J.W. and Richards, F.M. (1987), *J. Mol. Biol.* **193**, 775

Ramachandran, G.N. and Srinivasan, R. (1970) "Fourier Methods in Crystallography", Interscience, Wiley, New York.

Sarma, R. and Bott, R. (1977), *J. Mol. Biol.* **113**, 555.

Sheldrick, G.M. (1976) "SHELX: A Program for Crystal Structure Determination"

Sheldrick, G.M. (1986) "SHELXS-86" University of Göttingen.

The RuBisCO Saga

Herman A. Schreuder, Paul M.G. Curmi, Duilio Cascio & David Eisenberg
Molecular Biology Institute, UCLA, Los Angeles, California 90024-1570

Introduction

Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) catalyzes the reaction between atmospheric CO₂ and ribulose-1,5-bisphosphate. The six-carbon intermediate initially formed is cleaved to yield 2 molecules of 3-D-phosphoglycerate. Since this reaction is the first step in photosynthetic carbon dioxide fixation, this reaction is of considerable importance, responsible for the annual fixation of between 10¹⁰ and 10¹¹ tons of CO₂ (Woodwell, 1978).

Two different quaternary forms of RuBisCO are known. The enzyme from the photosynthetic bacterium *Rhodospirillum rubrum* forms a dimer, consisting of two large subunits, while RuBisCO from higher plants and most other photosynthetic microorganisms is a complex of eight large subunits (mw ~ 55 kd) and eight small subunits (mw ~ 15 kd). The whole complex has a molecular mass of about 550 kd and this paper will describe some of the problems encountered during the determination of such a large structure.

Brief historic overview

With the encouragement of professor Sam Wildman, the discoverer of RuBisCO (Wildman & Bonner, 1947), David Eisenberg and coworkers at the UCLA started structural studies of tobacco RuBisCO in 1970. In 1975 they described the L8S8 stoichiometry and 422 symmetry of the RuBisCO molecule. These findings were based on coordinated electron microscopy and X-ray crystallography of form I crystals (Baker et al., 1975). Earlier workers had discussed L24 and L8S6 models. Because of the poor diffraction pattern of the form I crystals ($d_{\min} \sim 50 \text{ \AA}$), a search for better crystals continued. Form II crystals were somewhat better ($d_{\min} \sim 14 \text{ \AA}$) and in 1977, coordinated electron microscopy and X-ray diffraction (Baker et al., 1977a) produced low resolution information about the shape of the RuBisCO molecule. Also in 1977, form III crystals were discovered (Baker et al., 1977b). These were the first RuBisCO crystals that gave any hope of a structure at the atomic level. In an effort to get even better X-ray quality crystals, RuBisCO's from many different sources were tried, and in 1980 a paper was published describing the crystallization and characterization of RuBisCO's from eight different species (Johal et al., 1980). While Eisenberg and his group decided to continue with form III crystals from tobacco RuBisCO, Andersson and Brändén (1984) managed to obtain high quality crystals of CABP-inhibited spinach RuBisCO.

The first RuBisCO to be solved at atomic level turned out to be the dimeric *R. rubrum* structure (Schneider et al., 1986). Now many papers about the RuBisCO structure followed in quick succession. A few will be mentioned here: The Berlin group reported evidence in 1987 for a spectacular shift of 36 Å between two halves of the L8S8 *Alcaligenes eutrophus* H16 RuBisCO

upon activation of the enzyme (Holzenburg et al., 1987). It was immediately noted by the group of Eisenberg that their current model of tobacco RuBisCO excluded such a shift (Chapman et al., 1987a). The evidence for the 36 Å shift was later reinterpreted (Choe, H.-W. et al., 1989).

However, when the 2.6 Å structure of tobacco RuBisCO was published (Chapman et al., 1988), the authors were in turn criticized about the interpretation of the density for the small subunit (Knight et al., 1989). The group of Brändén in Sweden had a 2.4 Å model of spinach RuBisCO built into a 2.4 Å electron density map based on three heavy atom derivatives, improved by four-fold averaging. The large subunits of the tobacco and spinach models agreed reasonably well, but the models for small subunits were different. To quote the abstract of their paper: “one of these models is clearly wrong,” leaving it to the reader to decide whether the spinach or the tobacco model was wrong. However, when we read further in the paper we find: “The arrangement of the L subunits in the spinach enzyme onto four L₂ dimers around the fourfold is also similar to the tobacco enzyme... The model of the S subunit described by Chapman et al. also contains a core of four antiparallel β strands. the positions of these β strands in relation to the L subunits agree in the two structures... Similarly, the positions of two of the three α helices in the tobacco model agree with our spinach model... The connections agree at one end of the β strands in the topology diagram, but they are quite different at the other end... Most of these regions, residues 1 to 6, 38 to 42, and 114 to 123 in the tobacco model have poor density in the tobacco electron density map according to Chapman et al....”

As an aside, we note that the acronym “RuBisCO” for ribulose-1,5-bisphosphate carboxylase/oxygenase was first proposed as a jest by Eisenberg at a symposium in honor of professor S.G. Wildman at his retirement. Wildman had the dream of producing large amounts of this enzyme as a protein supplement in prepared foods: Essentially to grow tobacco as a food crop. As a joke, Eisenberg suggested that the protein supplement and the company producing it be called “RuBisCO.” Wildman responded by saying that the terms for the enzyme in current use (“fraction I protein,” and “carboxylase”) were awkward, and that he intended to use “RuBisCO” forthwith. He did, and the name caught on.

Structure determination of tobacco RuBisCO.

Most of the procedures described below have been taken from the thesis of Michael Chapman (Chapman, 1987b) and are also described to some extent by Chapman et al., (1986).

Data were collected on Xuong-Hamlin multiwire area detectors (Hamlin, 1985). In order to obtain sufficient signal to noise ratio's, a relative long exposure time per frame was used. The resulting datasets were of low redundancy and the scaling was poor. This problem was overcome by collecting a special low resolution, high redundancy data set to boot-strap the scaling (Chapman, 1987b).

A great number of heavy atom reagents and conditions were tested, but eventually only four useful derivatives were obtained:

Table 1: Phasing statistics (from Chapman et al., 1986)

resolution (Å)	10.92	6.85	5.34	4.53	3.99	3.61	3.33	3.10	all
figure of merit	0.58	0.58	0.53	0.50	0.39	0.26	0.19	0.09	0.37

1. thiomersal
2. dimethyl mercury
3. $K_2Pt(CN)_4$
4. a double derivative of (2) and (3)

The internal R_{sym} 's were all between 4.5% and 8.1%. Except for the double derivative, each dataset was merged from datasets from two different crystals. The method of merging was unconventional in that common reflections were not averaged, but the reflection with the best signal to noise ratio was picked (Chapman, 1987b).

The heavy atom positions for the three single derivatives could be deduced from Patterson maps (Suh, 1980; Chapman, 1987b). The heavy atom parameters were refined with the program HEAVY by Terwilliger and Eisenberg (1983). The initial parameters were determined from Wilson plots. Some phasing statistics are given in Table 1. A plot of the phasing power as a function of resolution indicated that not much phasing power was left beyond 4.0 Å.

As could be expected from the phasing statistics, the MIR map was not readily interpretable and it was necessary to improve the phases by solvent flattening (B.C. Wang, 1985). Because in the early stages the automatic determination of the solvent mask was not very good, cycles with automatic mask determination were alternated by cycles in which the automated mask was manually corrected, e.g. holes in the protein region and peaks in the solvent region were removed. The overall figure of merit rose to 0.6 during this procedure.

At this stage, 12 α -helices were visible and fragments of the backbone could be built. However, in many places the direction of the chain and the connectivity could not be established unambiguously. In order to fit the sequence to the density, a guess was made about the nature of the side chains on the basis of the electron density (Richardson & Richardson, 1985). A Dayhoff-like table (Dayhoff et al., 1979) was constructed with estimates of the probabilities of assigning a different residue to the density of a particular amino acid. The profile method (Gribskov et al., 1987) was used to align the fragments built in density with the chemical sequence.

Around this time, the structure of *R. rubrum* RuBisCO was published (Schneider et al., 1986). Upon recognizing that the large subunit was dominated by an α/β barrel, much of the ambiguities of inter-fragment connections were solved by following the topology of pyruvate kinase (Muirhead et al., 1986) and triosephosphate isomerase (Banner et al., 1975). The strands in the N-terminal domain were connected with a topology similar to that of the α -carbon diagram of *R. rubrum* RuBisCO.

When about 65% of the atoms had been placed in the electron density map, further progress was made by boot-strapping the structure as follows: First, the partial model was extensively refined with restrained least squares (Hendrickson, 1985). Then, in an attempt to get rid of heavy model bias, rebuilding was done in OMIT maps. To calculate an OMIT map, the asymmetric unit was divided into 64 boxes and all model atoms in a certain box were deleted and F_{omit}, ϕ_{omit} were calculated from the truncated model. The ϕ_{omit} was combined with MIR phases and solvent flattened phases and a ϕ_{comb}, F_{obs} map was calculated for that particular box. This procedure was repeated for all 64 boxes and the all boxes were combined to a complete map (Chapman, 1987b). The OMIT maps seemed to show less model bias than $2F_{obs} - F_{calc}$ type maps. Cycles of model building, refinement and OMIT map calculation were repeated until all atoms had been built into density and a 2.6 Å model with an R-factor of 28% was presented (Chapman et al., 1988).

Despite considerable effort put into refinement, no further improvement in R-factor was attained.

Comparing the correct and the incorrect model.

Since the publication of the structure of tobacco RuBisCO in 1988 (Chapman et al., 1988), a much better model has been obtained for two reasons: A 2.0 Å dataset collected at the Brookhaven synchrotron allowed the calculation of much better electron density maps and the people of the group of Brändén in Sweden agreed to collaborate and were kind enough to make the coordinates of activated spinach RuBisCO available to us.

In the following discussion, the current model of tobacco RuBisCO, with an R-factor of 18.0% based on 84% of the theoretical data to 2.0 Å, will be referred to as the "new" (presumably correct) model and the 1988 model (probably incorrect) will be referred to as the "old" model.

The C α backbones of the old and the new model are shown in Figure 1. The barrel domain and the upper part of the N-terminal domain (in Figure 1) are more or less correct, while the lower part of the N-terminal domain and the small subunit show large discrepancies.

The lower part of the N-terminal domain has very high temperature factors in the new model, which most likely led to errors in the interpretation of the poor MIR map. The main chain of the small subunit has been built in the wrong direction in the MIR density, as will be shown in the next paragraph. Nevertheless, the position of the major elements of secondary structure (α -helices and β -sheets) was essentially correct, but many of the connecting loops were totally wrong. Specially two loops protruding from the small subunit into the barrel domain were missed (see Figure 1a).

A more quantitative view of the errors in the chain tracing is given in Figure 2, where the nearest C α in the new model is plotted against the residue number in the old model. Where the two models agree, the diagram gives a horizontal line of slope 1; where they disagree, the plot deviates from this. It is clear from Figure 2 that most C α atoms of the large subunit have been placed correctly, but that the chain tracing of the small subunit was reversed and that a number of wrong connections have been made. This is the same conclusion as the one made by Knight

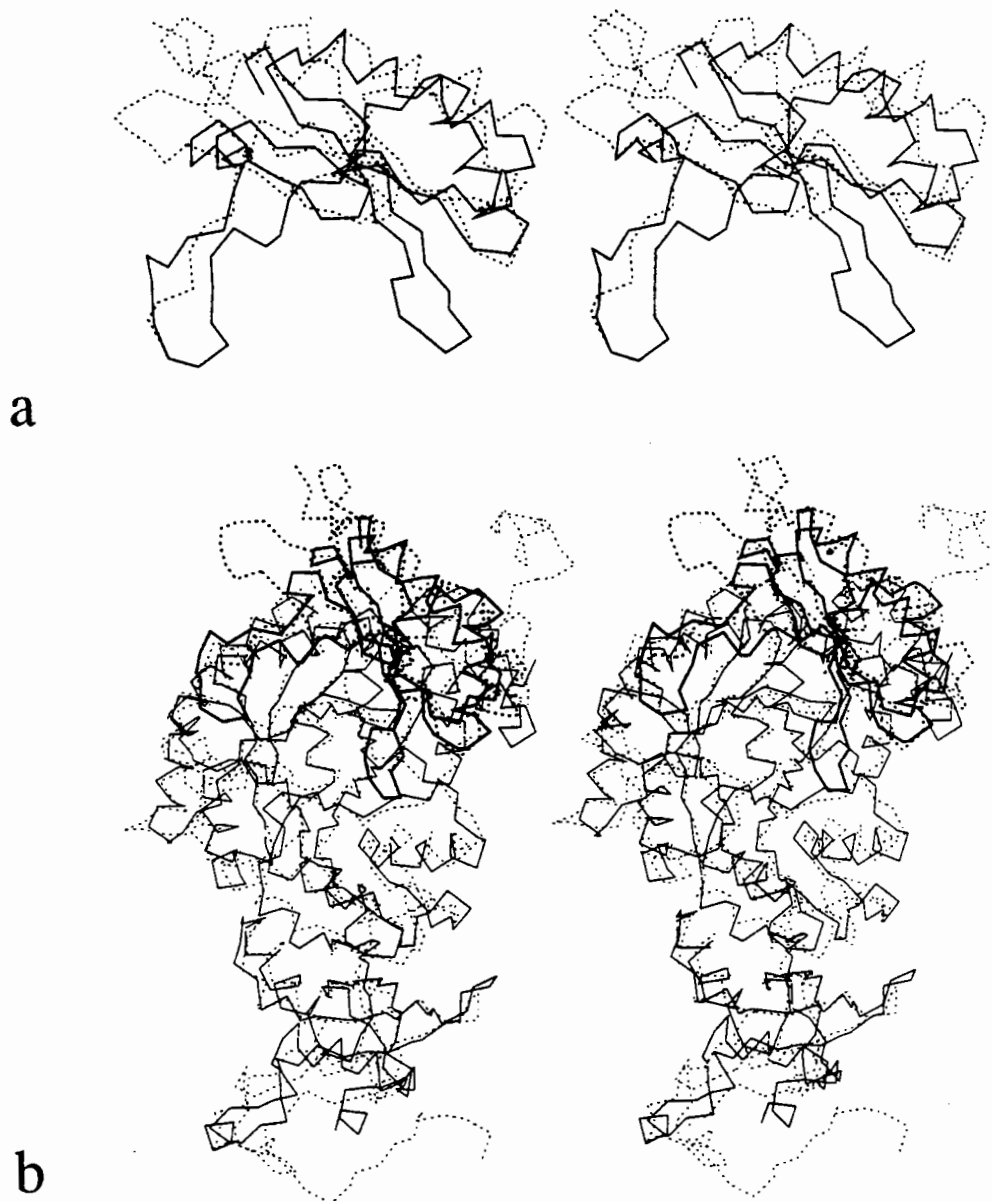


Figure 1: Comparison of the $C\alpha$ backbones of the old and the new model of tobacco RuBisCO. The backbone of the old model is indicated in broken lines, the backbone of the new model is indicated in continuous lines. a: comparison of the small subunits. b: comparison of both the large and the small subunits. The large subunits are indicated in thin lines, the small subunits are indicated in thick lines.

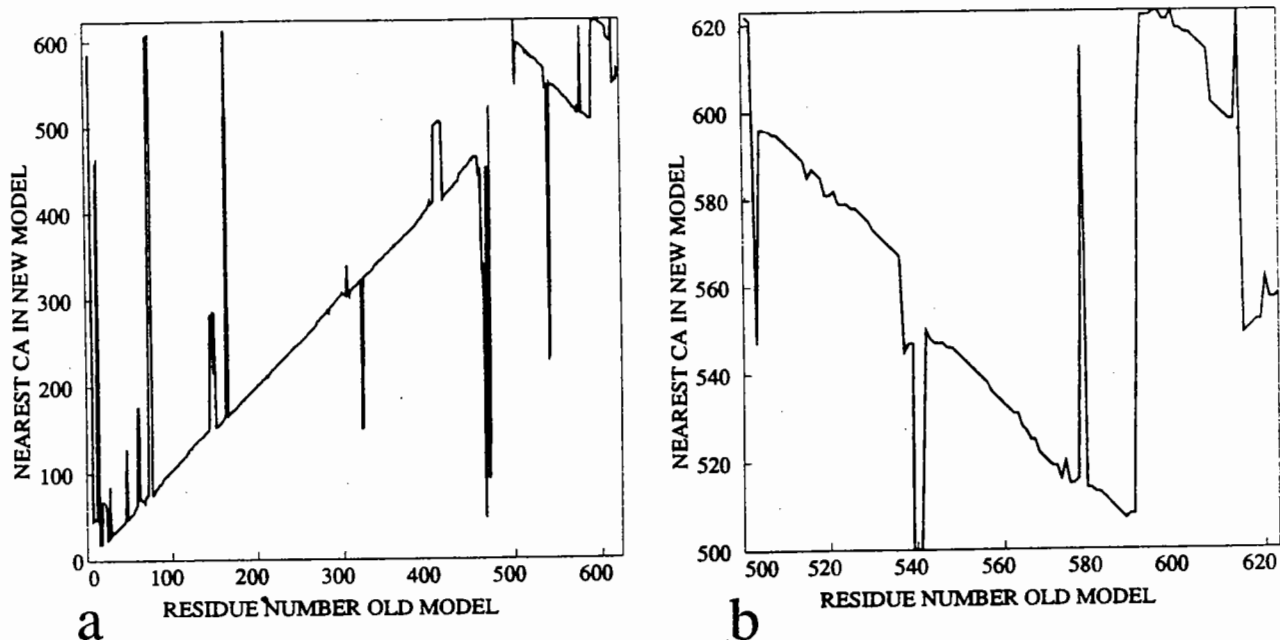


Figure 2: Nearest C α atom in the new model for each C α of the old model. The residue numbers of the small subunit have been increased by 500. a: combined plot of large and small subunit. b: plot of small subunit only

et al. (1989), as illustrated by Figure 3, taken from their paper.

To get some idea of the errors made in the assignment of the side chains, we divided the side chains somewhat arbitrarily into 5 categories and counted how often residues of different categories were mixed up. The results are given in Table 2.

Although many residues have been replaced by residues from a similar or neighbouring class, for example medium versus large, a significant number of total mismatches does exist. An example is that seven small residues have been mixed up with large aromatic residues. Table 3 gives an idea of the fit of the models to the two types of maps, used to build the model. It is

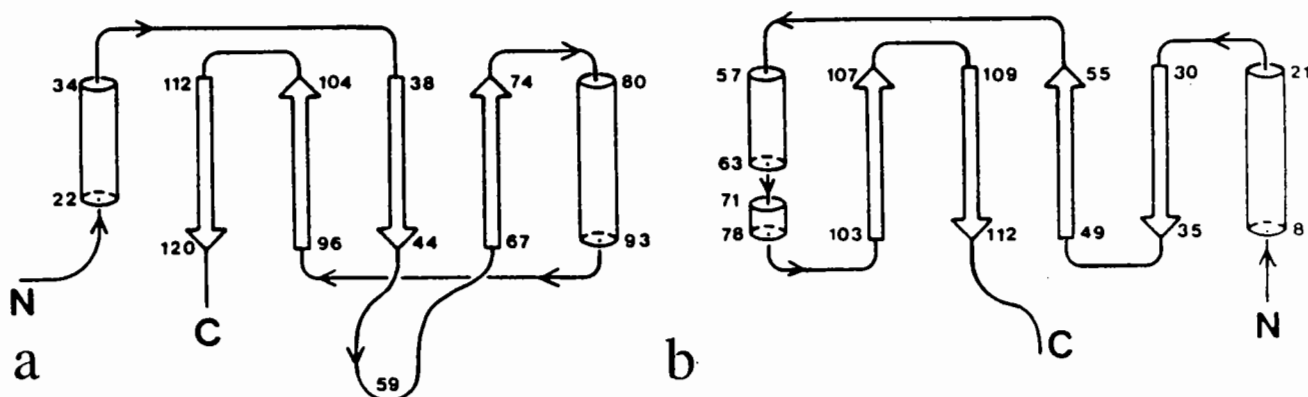


Figure 3: Comparison of the topologies of the small subunits. a: the spinach model. b: the tobacco model. Numbers refer to the amino acid sequence. Taken (with permission) from Knight et al., 1989.

Table 2: Replacement table between the old and the new model

category	1	2	3	4	5	amino acids in each category
1	1					1 very small: Gly, Ala
2	3	3				2 small: Val, Cys, Ser, Pro
3	5	17	7			3 medium: leu, Ile, Thr, Asp, Asn
4	7	9	18	15		4 large: Met, Glu, Gln, Lys, Arg, His
5	4	7	11	14	2	5 aromatic: Phe, Tyr, Trp

Table 3: Average density per residue (on an arbitrary scale)

model	MIR map		OMIT map	
	large	small	large	small
old	32.6	20.6	30.0	23.0
oriented spinach	40.2	37.1	28.2	24.3
new	44.9	40.8	32.0	26.4

clear that the oriented spinach model and the new model fit better to the MIR map than does the old model. However, the old model seems to fit the OMIT map reasonably well compared with the new model, indicating the presence of model bias in the OMIT map.

Criteria to assess the reliability of protein structures are ϕ, ψ plots (Ramachandran & Sasisekharan, 1968), shown in Figure 4. The almost random distribution of ϕ, ψ angle combinations for the old model is certainly indicative that some problems exist. However, the distribution for the large subunit is not significantly better than the distribution for the small subunit. Table 4 gives some numbers.

Table 4: Number of deviating ϕ, ψ angle combinations.

model	subunit	out of range	total	%
old	large	159	471	33.8
	small	50	123	41.0
new	large	15	447	3.4
	small	5	123	4.1

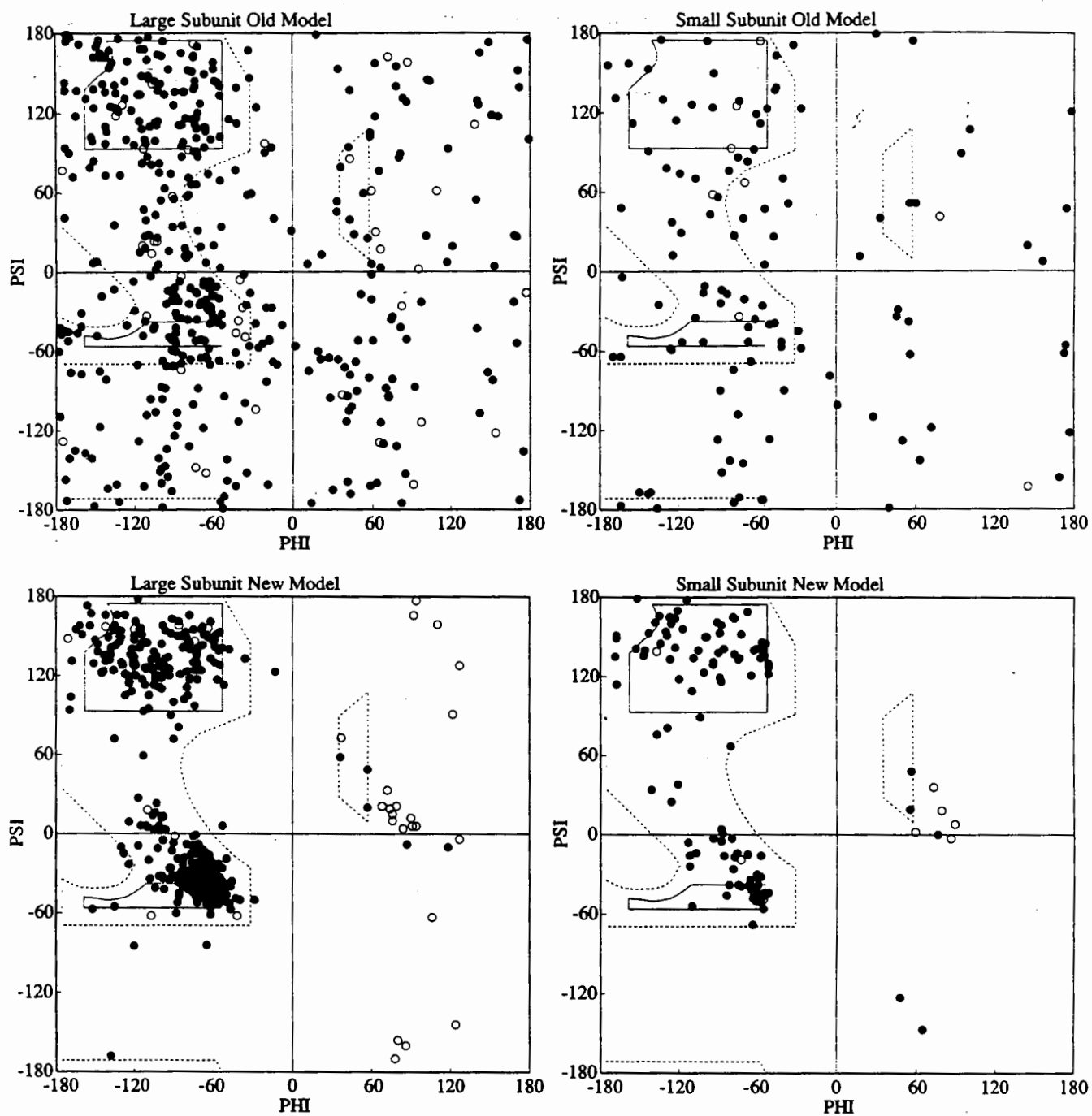


Figure 4: ϕ, ψ plots of the old and new tobacco models. Open circles denote glycine residues.

Discussion

As shown by the good fit of the new model to the MIR map, the MIR map was essentially correct and the error occurred in the interpretation of the map. This interpretation, however, was far from easy given the poor quality of the map. When we scanned old notebooks, we found many remarks indicating that both directions for the chain looked equally possible, or that the opposite direction from what had been built looked more plausible etc. Both chain directions were left open for quite some time during the model building. The turning point came when a match from the alignment program coincided with an independent manual fit to one of the helices. The other fragments were connected from this starting point. Unfortunately, as we know now, residues 65-70 had been fitted in reverse to the N-terminal helix (residues 22-34) and the OMIT maps did not remove enough model bias to reveal this error. OMIT maps not recombined with MIR phases, have even more model bias (data not shown). This model bias is almost certainly a result of the refinement procedure where adjustments of atoms outside the omitted box compensate for errors in the position of atoms inside the box.

In retrospect, there were some hints which could have signaled that the map interpretation was partly wrong:

- The R-factor did not drop below $\sim 27\%$, even after extensive refinement.
- The ϕ, ψ plot showed a near random distribution.
- Residues 51 to 62, which are deleted in the cyanobacterium *Anabaena*, are part of the core of the structure of the small subunit in the old model. This deletion would disrupt the core structure if the old model were true (Knight et al., 1989).

Although the MIR map was essentially correct, its effective resolution of 4 Å made the interpretation extremely difficult and hazardous. The best way to prevent this type of errors is therefore to collect better data. It is probably no accident that the data used for solving the spinach structure were collected at the Daresbury synchrotron. Other lessons to be learned are that one should not put too much confidence into OMIT maps, and that one should be extremely careful in connecting the fragments together. It may also be wise to start refinement with a poly-alanine model and not impose certain sidechains at certain positions.

Acknowledgments

We would like to thank the workers in earlier stages of structure determination of the tobacco structure: Tim Baker, Se Won Suh, Ward Smith and Michael Chapman. Their notebooks and comments provided invaluable information about the whole process. We also acknowledge the help of Carl-Ivar Brändén, Stefan Knight and Inger Andersson from Sweden for discussions about the spinach structure and for making the coordinates available to us. This research has been supported by grants from the NIH.

References

- Andersson, I. and Brändén, C.-I. *J. Mol. Biol.* **172** (1984) 363.
- Baker, T.S., Eisenberg, D., Eiserling, F.A. and Weissman, L. *J. Mol. Biol.* **91** (1975) 391.
- Baker, T.S., Eisenberg, D. and Eiserling, F. *Science* **196** (1977a) 293.
- Baker, T.S., Suh, S.W. and Eisenberg, D. *Proc. Natl. Acad. Sci.* **74** (1977b) 1037.
- Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C., Pogson, C.I., Wilson, I.A., Corran, P.H., Furth, A.J., Milman, J.D., Offord, R.E., Priddle, J.D. and Waley, S.G. *Nature* **255** (1975) 609.
- Chapman, M.S., Smith, W.W., Suh, S.W., Cascio, D., Howard, A., Hamlin, R., Xuong, N.-H. and Eisenberg, D. *Phil. Trans. R. Soc. Lond. B* **313** (1986) 367.
- Chapman, M.S., Suh, S.W., Cascio, D., Smith, W.W. and Eisenberg, D. *Nature* **329** (1987a) 354.
- Chapman, M.S. Ph. D. thesis UCLA, Los Angeles, California 90024-2569 (1987b).
- Chapman, M.S., Suh, S.W., Curmi, P.M.G., Cascio, D., Smith, W.W. and Eisenberg, D. *Science* **241** (1988) 71.
- Choe, H.-W., Georgalis, Y. and Saenger, W. *J. Mol. Biol.* **207** (1989) 621.
- Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. in "Atlas of Protein Sequence and Structure 1979," Volume 5 suppl. 3, Dayhoff, M.O. and Schwartz, R.M. (editors), National Biomedical research Foundation, Washington, D.C. (1979) 345.
- Gribskov, M., McLachlan, A.D. and Eisenberg, D. *Proc. Natl. Acad. Sci. USA* **84** (1987) 4355.
- Hamlin, R. *Methods Enzymol.* **114** (1985) 416.
- Hendrickson, W.A. *Methods Enzymol.* **115** (1985) 252.
- Holzenburg, A., Mayer, F., Harauz, G., van Heel, M., Tokuoka, R., Ishida, T., Harata, K., Pal, G.P. and Saenger, W. *Nature* **325** (1987) 730.
- Johal, S., Bourque, D.P., Smith, W.W., Suh, S.W. and Eisenberg, D. *J. Biol. Chem.* **255** (1980) 8873.
- Knight, S., Andersson, I. and Brändén, C.-I. *Science* **244** (1989) 702.
- Muirhead, H., Clayden, D.A., Barford, D., Lorimer, C.G., Fothergill-Gilmore, L.A., Schiltz, E. and Schmitt, W. *EMBO J.* **5** (1986) 475.
- Ramachandran, G.N. and Sasisekharan, V. *Advan. Protein Chem.* **23** (1968) 325.
- Richardson, J.S. and Richardson, D.C. *Methods Enzymol.* **115** (1985) 189.
- Schneider, G., Lindqvist, I., Brändén, C.-I. and Lorimer, G. *EMBO J.* **5** (1986) 3409.
- Suh, S.W. Ph. D. thesis, UCLA, Los Angeles, California 90024-2569 (1980).
- Terwilliger, T.C. and Eisenberg, D. *Acta Crystallogr. A* **39** (1983) 813.
- Wang, B.-C. *Methods Enzymol.* **115** (1985) 90.
- Wildman, S.G. and Bonner, J. *Arch. Biochem. Biophys.* **14** (1947) 381.
- Woodwell, G.M. *Sci. Am.* **238** (1978) 34.

Validation of protein structures - a case study: The small subunit of Rubisco

Stefan Knight, Inger Andersson and Carl-Ivar Brändén
Swedish University of Agricultural Sciences
Uppsala Biomedical Centre
Department of Molecular Biology
P.O. Box 590
S-751 24 Uppsala, Sweden

Introduction

Ribulose-1,5-bisphosphate carboxylase/oxygenase, Rubisco, catalyses the initial steps of two opposing metabolic pathways, carboxylation and oxygenation of ribulose-1,5-bisphosphate. The carboxylation reaction is the first step in the photosynthetic fixation of CO₂. The reaction yields two molecules of phosphoglycerate, which are partly recycled in the Calvin cycle to regenerate ribulose-1,5-bisphosphate and partly converted to starch, the main storage form of photosynthetic chemical energy. The oxygenation reaction yields one molecule each of phosphoglycerate and phosphoglycolate. Phosphoglycolate is metabolized in the photorespiratory pathway where reduced carbon is oxidized to CO₂. The energy released in these reactions is dissipated as heat. Since the photorespiratory process causes considerable loss of photosynthetic energy in plants, Rubisco is an important target for attempts to reduce photorespiration by increasing the carboxylation/oxygenation ratio of the enzyme through protein engineering.

Rubisco from all higher plants as well as blue-green algae is built up from two types of subunits, eight large (L, 55 kD) and eight small (S, 15 kD), forming an L₈S₈ molecule of molecular weight around 550 kD. In contrast the enzyme from the photosynthetic bacterium *Rhodospirillum rubrum* is a homodimer of L subunits. The L subunit is responsible for the catalytic activity, whereas the S subunit exerts a modulating effect. Removal of these subunits from the L₈ core decreases carboxylation activity by two orders of magnitude (Andrews, 1988). It is also possible that the S-subunits modulate the carboxylation/oxygenation ratio since all known L₈S₈ Rubisco molecules have a higher ratio than the L₂ enzyme from *Rh. rubrum* (Andrews and Lorimer, 1987).

Background

The possibility that a Rubisco molecule with a higher carboxylation/oxygenation ratio than found in nature might be constructed, given detailed knowledge of the reaction mechanisms of carboxylation and oxygenation in a structural framework, has prompted a number of crystallographic studies of Rubisco from various sources (Andersson and Brändén, 1984; Andersson *et al.*, 1989; Baker *et al.*, 1975, 1977a, 1977b; Barcena *et al.*, 1983; Chapman *et al.*, 1987, 1988; Choe *et al.*, 1985; Holzenburg *et al.*, 1987; Janson *et al.*, 1984; Knight *et al.*, 1989; Lundqvist and Schneider, 1988, 1989; Nakagawa *et al.*, 1986; Pal *et al.*, 1985; Schneider *et al.*, 1986a, 1986b, 1990a, 1990b). The first Rubisco structure to be reported was that of the non-activated enzyme from *Rh. rubrum* (Schneider *et al.*, 1986b).

This study revealed the two-domain structure of the L subunit with one smaller N-terminal domain and a larger C-terminal domain folded as an eight-stranded α/β barrel and located the active site to the subunit interface at the C-terminal end of the β -strands in the α/β barrel. The model of the non-activated enzyme from *Rh. rubrum* has now been refined to 1.7 Å resolution (Schneider *et al.*, 1990a) and is currently being used in the design and interpretation of site-directed mutagenesis experiments. Preliminary descriptions of the structures of the non-activated tobacco enzyme (Chapman *et al.*, 1987, 1988) and the activated spinach enzyme with a bound reaction-intermediate analogue (Andersson *et al.*, 1989, Knight *et al.*, 1989) have been published. These two models of L_8S_8 Rubisco agree on the fold of the L subunits, which in both cases is similar to that of the *Rh. rubrum* L subunit. Both studies also show the L_8S_8 molecule to be formed by a core of four L_2 dimers arranged around a four-fold axis, with clusters of four S subunits at each end of the molecule. The fold of the S chain is, however, completely different in the two models (Knight *et al.*, 1989). Although the same secondary structural elements are present in both models, the connections between these elements are different. Furthermore, the location of the N- and C-termini is completely different so that actually none of the residues in the subunit occupy equivalent positions in the two models (Figure 1). Since 77% of the residues in the tobacco and spinach S subunit are identical, there is no doubt that both subunits have the same fold. Although the published tobacco model had been partially refined, giving a crystallographic R-value of 28 % at 2.6 Å resolution (Chapman *et al.*, 1988), whereas the spinach model at that time had not been refined, a number of factors clearly showed the spinach model to be the correct one. Today there is no dispute over this matter. The UCLA group have rebuilt their model of the S subunit based on that from spinach, and refinement has proceeded to an R-value of 18.0% at 2.0 Å resolution. The model of spinach Rubisco has been refined to an R-value of 19.4% at 2.4 Å resolution.

In this paper we describe those features of the structure determination that were most important in obtaining a correct result and summarize the evidence validating the model of the spinach S subunit. None of this evidence is dependent on a refined model and instead makes use of known facts about proteins in general and the S subunit of Rubisco in particular. A detailed account of the structure determination and description of the structure of spinach Rubisco will be given elsewhere (Knight *et al.*, 1990).

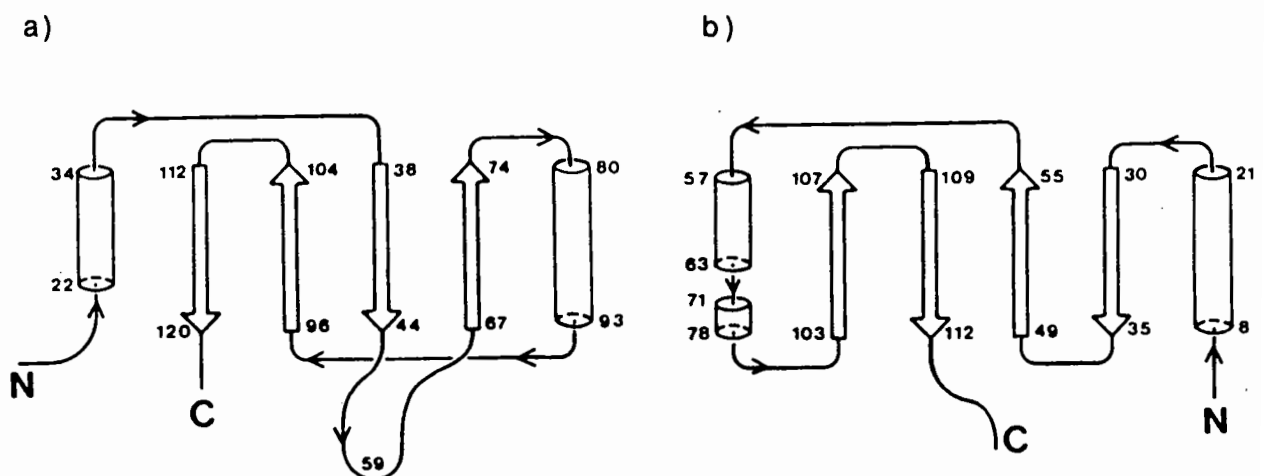


Figure 1. Topology diagrams of (a) the spinach model and (b) the tobacco model (Chapman *et al.*, 1988) of the S subunit of Rubisco. Arrows denote β -strands and cylinders α -helices. Numbers refer to residue numbers in the amino acid sequence.

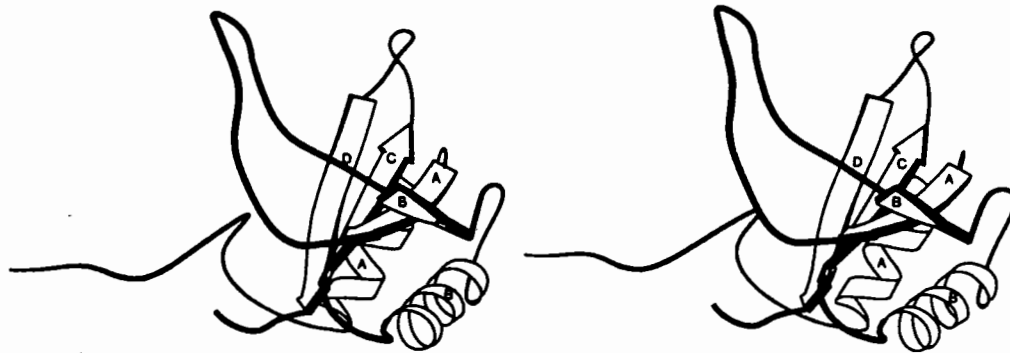


Figure 2. Computer-generated ribbon diagram (Priestle, 1988) of the S subunit of spinach Rubisco.

The fold of the S subunit

The S subunit of Rubisco is folded as a four-stranded anti-parallel beta-sheet of topology (+1, -2x, -1), covered on one side by two helices (Figure 2). The first 20 residues at the N-terminus form an irregular arm, which extends from the main body of the structure to a neighbouring small subunit in the L_3S_8 molecule. The edge strands in the beta-sheet, βB and βD , are somewhat irregular, particularly strand βB where residues 71 - 73 form a bulge and do not form hydrogen bonds to strand βA . The two strands βA and βB are joined by a long loop, comprising residues 46 to 67, which protrudes into the central solvent channel in the L_3S_8 molecule.

The model of spinach Rubisco was built in high-quality electron density

The model of spinach Rubisco was built in an electron density map obtained by cyclic non-crystallographic averaging (Bricogne, 1976) of an m.i.r. map, based on three heavy-atom

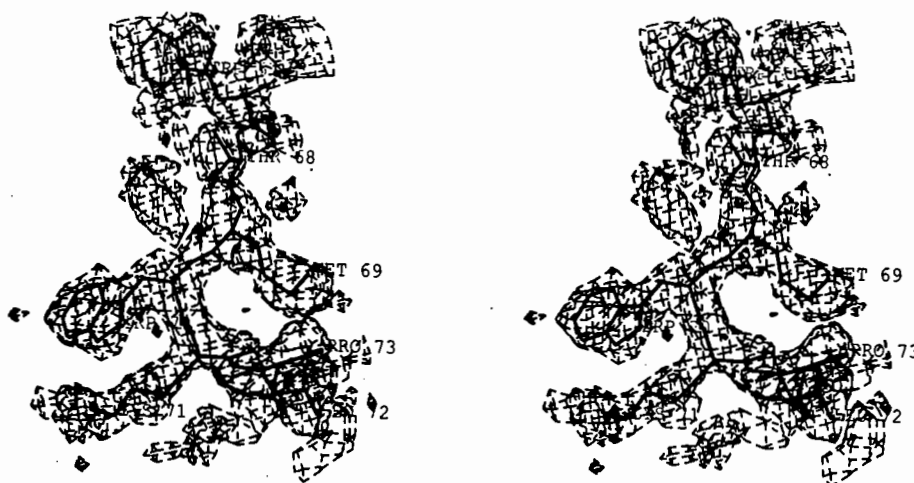


Figure 3. (a) Example of electron density in the averaged map with residues 67 - 73 in the S subunit superimposed. The contour level is at one standard deviation of the map.

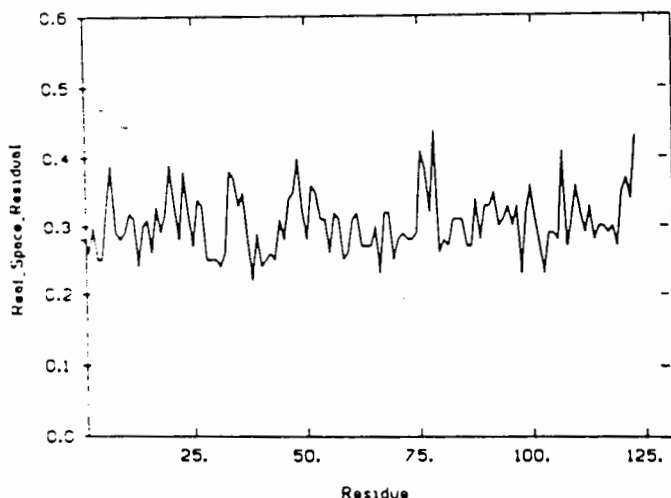


Figure 3. (b) Main-chain real-space R-factor (Brändén and Jones, 1990) for the S subunit in the averaged electron density map.

quality (Figure 3a) and we had no problems in unambiguously tracing the fold of the polypeptide chains, although a few short breaks in the continuous electron density were present in loop regions. Chain-tracing for the small subunit was further facilitated by the presence of an unusually large number of aromatic side-chains, as well as three methionine residues, that could be used as "handles". Figure 3b shows the fit of the model of the S subunit to the averaged electron density map as evidenced by the real-space R-factor per residue (Brändén and Jones, 1990). Only one L subunit and one S subunit were built and then used to generate the whole molecule using the four-fold local symmetry. The R-factor for this initial model was 43.0% for all observed reflections between 8.0 and 2.4 Å resolution.

Modelbuilding was performed using fragments from well refined structures

Model building was performed on an Evans & Sutherland PS330 vector graphics display using FRODO (Jones, 1978; Jones and Thirup, 1986). The L subunit was built starting from the *Rh. rubrum* model. The S subunit as well as those parts of the L subunit that were ill-defined in the *Rh. rubrum* structure, or where the two structures were obviously very different, were built using the BONES option of FRODO (Jones and Thirup, 1986). With this option, a skeletonized representation of the electron density can be edited to give a chain-tracing of pseudo-atomic positions. The skeleton representation of the backbone was used to build a poly-alanine model using short amino-acid fragments from a data-base of 32 refined structures as described by Jones and Thirup (1986). This approach to modelbuilding ensures that there are no gross errors in the stereochemistry of the resulting model. Figure 4 shows the Ramachandran plot for the initial, unrefined, model of the S subunit from spinach Rubisco. As can be seen, a majority of the residues have ϕ - ψ angles in allowed regions, although a fair number also fall in disallowed regions of the plot. These latter residues are those that connect the fragments from the database or are found in regions of the structure for which no fragment could be found in the database. The side-chains were fitted manually by dihedral rotations and fragment moves.

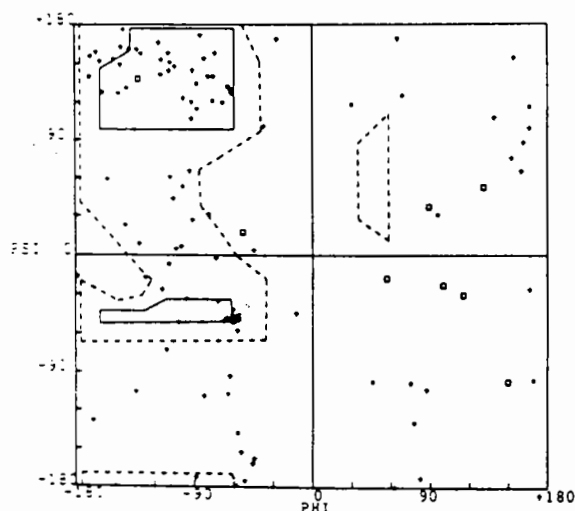


Figure 4. Ramachandran plot for the initial, unrefined, model of the spinach Rubisco S subunit.

Heavy-atom binding sites have chemically sound ligands

Heavy-atoms used to obtain heavy-atom derivatives usually bind to specific ligands. For example, it is well established that mercurials in most cases bind to cysteins. In the structure determination of spinach Rubisco, two mercury derivatives, $K_2Hg(CN)_4$ and ethyl mercuri thiosalicylate, were used. These derivatives contained a fairly large number of sites with eight distinct mercury binding sites per Rubisco L-S protomer, four on each L subunit and four on each S subunit. All of the sites on the L subunit were located close to cystein residues. Although the published sequence of the spinach S subunit (Martin, 1979) contains only one cystein residue, the presence of three cystein residues at positions 44, 77 and 112 has been established by isolating and sequencing all cystein-containing peptides of the spinach S subunit (G. Lorimer and B. Ranty, personal communication). The electron density in our map also strongly suggested that there should be cysteins in these positions. All of the mercury-binding sites on the S subunit were found close to these three residues, thus giving independent confirmation of the fold of the subunit.

The S subunit has a well defined hydrophobic core

The interior of proteins usually contain a well defined core of hydrophobic residues. This core is generally not conserved but can instead tolerate a large number of substitutions, as long as the residues remain hydrophobic. The hydrophobic core in the small subunit is formed by packing the two alpha-helices against the anti-parallel beta-sheet. Residues from helix αA interact mostly with beta-strand βD whereas helix αB is packed against beta-strands βA and βB . Residues involved in this core are listed in Table I. All these residues except A97 and V99 are invariably hydrophobic in all known sequences of the small subunit.

Table I. Amino-acid residues in the core of the small subunit.

V30	L42	V90	I101
L33	V83	A97	F115
P40	V87	V99	

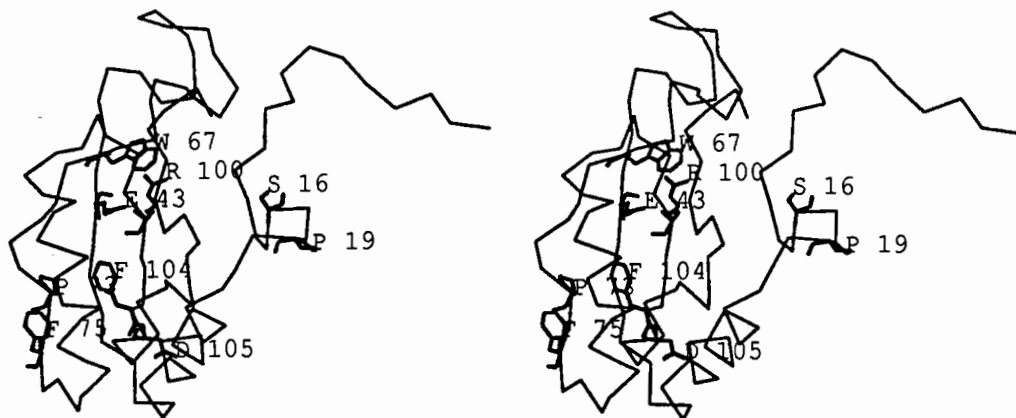


Figure 5. Strictly conserved residues in the S subunit cluster in three distinct areas on the surface of the subunit.

Conserved residues are found at the S-L subunit interfaces

Residues that are conserved between species may be conserved for structural reasons or for functional reasons. Usually, as the number of available sequences goes up, the number of strictly conserved residues goes down. If the thirty or so sequences of the S subunit that have been published to date are examined, only nine invariant residues are found. These nine strictly conserved residues, from different regions along the polypeptide chain, are clustered in three distinct regions on the surface of the small subunit (Figure 5). All these residues except one are involved in interactions with conserved regions of the L subunits in the L_8S_8 molecule. If regions of the S-chain where there is a high degree of homology are examined, these are also found to interact with L subunits, whereas highly variable regions are located on the outside of the molecule, exposed to solvent.

A deletion in cyanobacterial S subunits occurs within a loop

It is well known that deletions and insertions in proteins occur in loop regions between secondary structural elements. S subunits from different species vary considerably in length. For example, in small subunits from cyanobacteria, residues 52 to 63 are deleted, whereas in some algae there are instead insertions in this region. In the model of the spinach S subunit, this region of the sequence is found within the long loop between strands βA and βB . As can be seen in Figure 6, residues 51 and 64 are juxtaposed in the spinach S subunit. The distance between the $C\alpha$ atoms of these two residues is only 4.4 Å and model building has shown that the additional residues in plant S subunits can be deleted, and residues 51 and 64 connected, without disrupting the rest of the subunit structure.

Although the presence of deletions/insertions in loop regions does not constitute proof of a correct fold; if deletions/insertions occur within secondary structural elements this strongly indicates that the model is not correct.

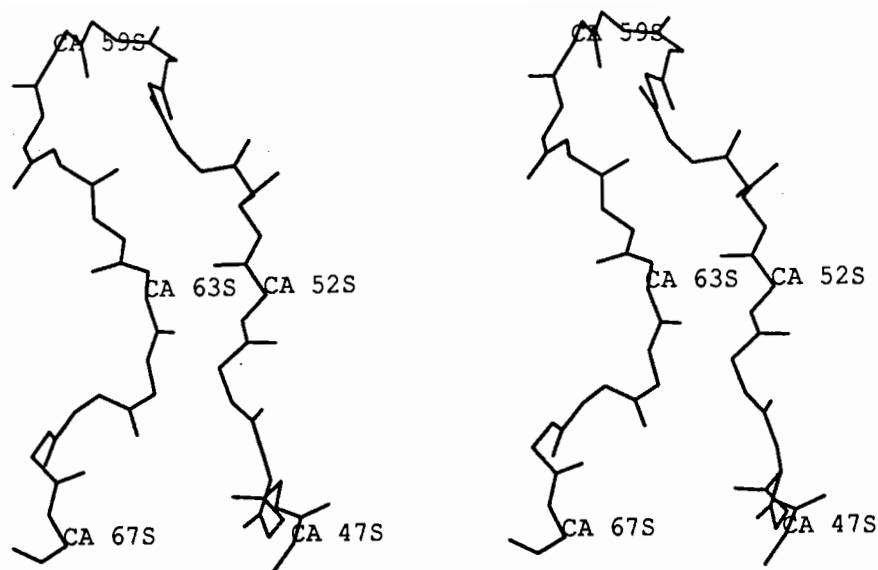


Figure 6. Stereodiagram showing the backbone of the long loop between strands βA and βB in the S subunit. In S subunits from cyanobacteria, residues 52 to 63 are deleted.

Concluding remarks

The model of the S subunit of spinach Rubisco was built in an electron density map of high quality, thus making chain-tracing quite easy. Model-building using fragments from a database of well-refined structures minimized the stereochemical errors in the initial model. Heavy-atom binding sites were found close to the expected ligands. The subunit shows a normal distribution of residues, with a well defined hydrophobic core and conserved residues at the surface where they interact with the L subunit in the L_8S_8 molecule. A deletion present in cyanobacterial S subunits removes a loop that is present in plant S subunits.

The initial, unrefined model of the S subunit of spinach Rubisco could thus be validated using physical, chemical and biological information. Information of this kind is generally present and should be used to check the initial model.

References

- Andersson, I. & Brändén, C.-I. (1984). *J. Mol. Biol.* **172**, 363-366.
- Andersson, I. Knight, S., Schneider, G., Lindqvist, Y., Lundqvist, T., Brändén, C.-I. & Lorimer, G.H. (1989) *Nature* **337**, 229-234.
- Andrews, T.J. (1988). *J. Biol. Chem.* **263**, 12213-12219.
- Andrews, T.J. & Lorimer, G.H. (1987). In "The Biochemistry of Plants" (Hatch, M.D., ed.), vol. 10, pp. 131-218, Academic Press, Orlando.

- Baker, T.S., Eisenberg, D., Eiserling, F.A. & Weisman, L. (1975). *J. Mol. Biol.* **91**, 391-399.
- Baker, T.S., Eisenberg, D. & Eiserling F. (1977a). *Science* **196**, 293-295.
- Baker, T.S., Suh, S.W. & Eisenberg, D. (1977b). *Proc. Natl. Acad. Sci. U.S.A.* **74**, 1037-1041.
- Barcena, J.A., Pickersgill, R.W., Adams, M.J., Phillips, D.C. & Whatley, F.R. (1983). *EMBO J.* **2**, 2363-2367.
- Brändén C.-I. & Jones, T.A. (1990). *Nature* **343**, 687-689.
- Bricogne, G. (1976). *Acta Crystallogr. sect. A* **32**, 832-847.
- Chapman, M.S., Se Won Suh, Cascio, D., Smith, W.W. & Eisenberg, D. (1987). *Nature* **329**, 354-356.
- Chapman, M.S., Se Won Suh, Curmi, P.M.G., Cascio, D., Smith, W.W. & Eisenberg, D.S. (1988). *Science* **241**, 71-74.
- Choe, H.-W., Jakob, R., Hahn, U. & Pal, G.P. (1985). *J. Mol. Biol.* **185**, 781-783.
- Holzenburg, A., Mayer, F., Harauz, G., van Heel, M., Tokuoka, R., Ishida, T., Harata, K., Pal, G.P. & Saenger, W. (1987). *Nature* **325**, 730-732.
- Janson, C.A., Smith, W.W., Eisenberg, D. & Hartman, F.C. (1984). *J. Biol. Chem.* **259**, 11594-11596.
- Jones, T.A. (1978). *J. Appl. Cryst.* **11**, 268-272.
- Jones, T.A. & Thirup, S. (1986). *EMBO J.* **5**, 819-822.
- Knight, S., Andersson, I. & Brändén, C.I. (1989). *Science* **244**, 702 - 705.
- Knight, S., Andersson, I. & Brändén, C.I. (1990). *J. Mol. Biol.* in press.
- Lundqvist, T. & Schneider, G. (1988). *J. Biol. Chem.* **263**, 3643-3646.
- Lundqvist, T. & Schneider, G. (1989). *J. Biol. Chem.* **264**, 7078-7083.
- Martin, P.G. (1979). *Aust. J. Plant Physiol.* **6**, 401-408.
- Nakagawa, H., Sugimoto, M., Kai, Y., Harada, S., Miki, K., Kasai, N., Saeki, K., Kakuno, T. & Horio, T. (1986). *J. Mol. Biol.* **191**, 577-578.
- Pal, G.P., Jakob, R., Hahn, U., Bowien, B. & Saenger, W. (1985). *J. Biol. Chem.* **260**, 10768-10770.
- Priestle, J.P. (1988). *J. Appl. Cryst.* **21**, 572-576.
- Schneider, G., Brändén, C.-I. & Lorimer, G. (1986a). *J. Mol. Biol.* **187**, 141-143.
- Schneider, G., Knight, S., Andersson, I. Brändén, C.-I., Lindqvist, Y. & Lundqvist, T. (1990b). *EMBO J.* in press.
- Schneider, G., Lindqvist, Y., Brändén, C.-I. & Lorimer, G. (1986b). *EMBO J.* **5**, 3409-3415.
- Schneider, G., Lindqvist, Y. & Lundqvist, T. (1990a). *J. Mol. Biol.* **211**, 989-1008.

Hallmarks of a Wrong Structure

C. D. Stout
Department of Molecular Biology
Research Institute of Scripps Clinic
La Jolla, CA 92037

A number of incorrect protein structures have been reported recently as attested to by this volume and summarized in *Nature* (C. Brändén and A. Jones, 1990). The structure of a 7Fe ferredoxin from *A. vinelandii*, which I originally reported, is one of these. In this paper I list characteristics of the structure as first reported which were indicative of an incorrect solution. An independent structure determination of this protein, which proved my original model to be incorrect, has been published (G.H. Stout, et al., 1988). Subsequently, I have independently redetermined the structure with new data sets (C.D. Stout, 1988) and refined the structure at 1.9Å resolution (C.D. Stout, 1989). The reader is referred to these papers for crystallographic details of the structure determinations.

The first error in the original analysis was due to inversion of positive and negative peaks in a Bijvoet difference Fourier map (C.D. Stout, 1979). The space group assignment as $P4_12_12$ or $P4_32_12$ and coordinates of a single-site Pt derivative were made correctly. However, in the calculation of the Bijvoet difference Fourier map (G. Strahs and J. Kraut, 1968) the signs of the coefficients were inverted. This must have occurred in the calculation because data collected in point group 422 (or 222) in a right-handed system necessarily retain the correct sense of ($|F^+|$) and ($|F^-|$) (e.g. A.H. Robbins and C.D. Stout, 1985). The effect of this error on the map is detailed in Table 1. Consequently the wrong hand of the structure was chosen (C.D. Stout, 1979). Therefore, an error of the entire analysis was to place undue weight on this one result. In the structure redetermination, the hand of the structure was confirmed with a second isomorphous derivative, as well as with a Bijvoet difference Fourier map (C.D. Stout, 1988).

TABLE 1

Space Group Determination in P4₁2₁2/P4₃2₁2
Using a Bijvoet Difference Fourier Map

<u>Space Group</u>	<u>SIR Phase Calculation Pt Coordinates</u>	<u>SIR Phase Set</u>
P4 ₁ 2 ₁ 2	+X +Y +Z	++
P4 ₁ 2 ₁ 2	-X -Y -Z	(a)
P4 ₃ 2 ₁ 2	+X +Y +Z	(a)
P4 ₃ 2 ₁ 2	-X -Y -Z	--

Bijvoet Difference Fourier Map

<u>SIR Phase Set</u>	<u>Coeficients</u>	<u>Sign</u>	<u>Peaks Z Coordinates^(b)</u>	<u>Space Group Indication</u>
++	(F ⁺ - F ⁻)	+	0.12, 0.22	P4 ₁ 2 ₁ 2 ^(c)
--	(F ⁺ - F ⁻)	-	0.38, 0.28	P4 ₁ 2 ₁ 2 ^(c)
++	(F ⁻ - F ⁺)	-	0.12, 0.22	P4 ₃ 2 ₁ 2
--	(F ⁻ - F ⁺)	+	0.38, 0.28	P4 ₃ 2 ₁ 2

(a) Incorrect assignments of the Pt position with respect to the space group (C.D. Stout, 1979).

(b) Two Fe-S clusters at 0.12, 0.22 in the correct solution. Inverted solutions in this space group pair are related by xyz, xy 1/2-z.

(c) Correct space group of the re-determined structure.

The original structure determination proceeded from the incorrect SIR phase set, i.e. (--) in P4₃2₁2 (Table 1). The SIR map indicated the presence of a previously unrecognised type of Fe-S cluster with three Fe atoms (C.D. Stout, et al., 1980; M.H. Emptage, et al., 1980). The co-discovery of this cluster stimulated a free interpretation of the protein structure in spite of rather poor 3.0Å resolution electron density (D. Ghosh, et al., 1981); consequently, the model deduced lacked overall homology with the known structure of a 8Fe ferredoxin (E.T. Adman, et al., 1976). Therefore, two major problems at this stage of the analysis were a lack of appreciation of the importance of sequence homology in protein structure, and a lack of experience in interpreting protein electron density.

Subsequently, the incorrect structure was also refined (D. Ghosh, et al. 1982). Indications that this model was an incorrect solution have been pointed out (G. H. Stout, et al., 1988). These factors are listed in here. (1) The R-factor at 2.0Å was 0.36 after refinement when waters in the model and individual B-factors were omitted. In the correct structure this R-factor is 0.26 at 1.9Å; addition of waters to the model reduces R only 2.5% while refinement of isotropic B's reduces R only another 2.5% (C. D. Stout, 1989). In addition, the incorrect R-factor was calculated for data with a σ -cutoff; the correct R-factor is for all the data. (2) Far too many waters (>200) were added to the incorrect model; the correctly refined structure has 21 waters. (3) The incorrect structure had an abnormal distribution of B-factors, e.g. $B > 30.0 \text{Å}^2$ on some Fe atoms. (4) The distribution of ϕ, ψ angles was abnormal, whereas the correct structure has a normal distribution in spite of the presence of the novel Fe-S cluster. (5) The Fe-Fe distances in the refined incorrect structure, $\sim 4.0 \text{Å}$, were not in agreement with EXAFS results, $\sim 2.7 \text{Å}$, for a similar 3Fe cluster in another ferredoxin (M.R. Antonio, et al., 1982). (6) While omit maps returned electron density for the incorrect model with respect to the amino-acid sequence (J.B. Howard, et al., 1983) these maps were influenced by bias in the phases should not have been trusted. The correct structure leads to chemically reasonable results for site-directed mutant variants of the sequence and structure (A.E. Martín, et al., 1990).

The crystallographic lessons of this exercise are obvious. Correct structures obey good stereochemistry and sequence-structure homology in

spite of novel features, and correct structures have honestly low R values without extra waters, unrestrained B's, unusual weighting schemes, truncated data sets, etc. The philosophical lesson is to resist interpreting data when the results are coupled to rewards such as tenure, government grants, Ph.D. theses, etc., unless the data are substantive science as judged by those with experience in protein crystallography.

Acknowledgements. I am grateful to B.W. Matthews for advice and encouragement prior to the redetermination of the structure. Financial support has been provided by the National Institutes of Health.

References

1. Adman, E.T., Sieker, L.C. and Jensen, L.H. *J. Biol. Chem.* **251** (1976) 3801.
2. Antonio, M.R., Averill, B.A., Moura, I., Moura, J.J.G., Orme-Johnson, W.H., Teo, B.K., and Xavier, A.V. *J. Biol. Chem.* **257** (1982) 6646.
3. Brändén, C. and Jones, T.A. *Nature* **343** (1990) 687.
4. Emptage, M.H., Kent, T.A., Huynh, B.H., Rawlings, J., Orme-Johnson, W.H. and Münck, E. *J. Biol. Chem.* **255** (1980) 1793.
5. Ghosh, D., Furey, W., O'Donnell, S. and Stout, C.D. *J. Biol. Chem.* **256** (1981) 4185.
6. Ghosh, D., O'Donnell, S., Furey, W., Robbins, A.H. and Stout, C.D. *J. Mol. Biol.* **158** (1982) 73.
7. Howard, J.B., Lorschach, T.W., Ghosh, D., Melis, K. and Stout, C.D. *J. Biol. Chem.* **258** (1983) 508.
8. Martín, A.E., Burgess, B.K., Stout, C.D., Cash, V.L., Dean, D.R., Jensen, G.M. and Stephens, P.J. *Proc. Natl. Acad. Sci.* **87** (1990) 598.
9. Robbins, A.H. and Stout, C.D. *J. Biol. Chem.* **260** (1985) 2328.

10. Strahs, G. and Kraut, J. *J. Mol. Biol.* **35** (1968) 503.
11. Stout, C.D. *Nature* **279** (1979) 83.
12. Stout, C.D., Ghosh, D., Patabhi, V. and Robbins, A.H. *J. Biol. Chem.* **255** (1980) 1797.
13. Stout, G.H., Turley, S., Sieker, L.C. and Jensen, L.H. *Proc. Natl. Acad. Sci.* **85** (1988) 1020.
14. Stout, C.D. *J. Biol. Chem.* **263** (1988) 9256.
15. Stout, C.D. *J. Mol. Biol.* **205** (1989) 545.

A Tale of Four Iron-Sulfur Proteins: Sequence Errors and Other Matters

Elinor T. Adman
Department of Biological Structure SM-20
University of Washington, Seattle, WA, USA 98195

The subtitle for this presentation should be "There Are Still Errors Even With Good R-factors". We have now refined four (albeit small) iron sulfur proteins in this laboratory, and in each have found more than we bargained for. (A fifth, an even smaller rubredoxin not actually discussed at this meeting, has been refined by Ron Stenkamp using diffractometer data to an R of 0.093 for data from 5 to 1.5Å¹. A disordered residue, identified chemically to be a cysteine, resembled a valine.) I am going to present three cases where we have found sequence differences to varying degrees. The fourth, azotobacter ferredoxin I, was surprising in the initial development of the correct structure, detailed elsewhere², but has not produced any further surprises.

The four cases are as follows:

1. *P. aerogenes* ferredoxin 8Fe,8S 54 amino acids (55, corrected)

P2₁2₁2₁ a=30.52 b=37.75 c=39.37 Å
refined with 2 Å data (diffractometer, two crystals)³
R=0.188 over 535 atoms (180 solvent)
when corrected: R=0.137 over 484 atoms (89 solvent)

2. *Desulfovibrio vulgaris* rubredoxin 1Fe 52 amino acids⁴

P2₁ a=19.993 b=41.505 c=24.404 Å β=107.6
refined with 1.5 Å data (diffractometer, one crystal)
R=0.098 575 atoms (176 solvent)
sequence error detected at R= 0.109

3. *Desulfovibrio gigas* ferredoxin II 3Fe,4S 57 amino acids (58 corrected)⁵

C2 a=40.87 b=45.28 c=26.47 Å β=104.7
refined with 1.7 Å data (diffractometer, 3 crystals)
R=0.157 484 atoms (55 solvent)
sequence error detected during refinement

4. *Azotobacter vinelandii* ferredoxin I 7Fe,8S 106 amino acids²

P4₁2₁2 a=b=55.6 c=95.51 Å
refined with 2.3 Å data (diffractometer)
R=0.170 954 atoms (98 solvent)

Each of these presents a slightly different story. I have discovered no particular clue as to what to look for in general with respect to potential errors, other than careful attention to detail, ample caution, and a firm belief that there is an enormous amount of information in good data sets.

Case 1. *P. aerogenes* ferredoxin. This was the second protein to be refined in our laboratory, the first for me. We refined the model, developed from an MIR map, using differential difference Fourier techniques, alternating cycles of X-ray refinement with idealization. The structure, at an R of 0.188 (unidealized, .206, idealized) was reported in 1976³. In that publication we noted difficulties in two places. Residue 23, according to our electron density maps was an internal Ile, but was identified chemically as a Gln. A sharp turn at residues 26-27, had always been difficult to fit, and an extra residue inserted earlier in the chain had been tried, but discarded.

Recently we decided to re-refine the structure using PROLSQ/PROTIN⁶ with which we have had much experience. In addition we now had FRODO⁷ on an E&S PS340 whereas the earlier refinement checking had been done with electron density plotted on plastic sheets. We lost track of the solvent positions (they had not been deposited along with the protein in the Brookhaven Data Bank (!)), so that we started with phases computed from the protein model alone, which turned out to be to our advantage. A few cycles of PROLSQ without solvent reduced R to about 0.23 from 0.28. The largest peaks in a difference map at this point showed two new features. One of the largest peaks occurred at the C γ -C δ of Ile 22, even though side chain atoms of that isoleucine had entirely reasonable B-values. The second region of interest was a large shift peak associated with the turn at residues 26-27. Looking at a compilation of ferredoxin sequences⁸ (published subsequent to the earlier refinement) showed that *P. aerogenes* ferredoxin had a deletion in position 22 relative to others, and that in some cases that residue was a cysteine. We immediately hypothesized that in fact what we had identified as Ile 22 was a cysteine, in turn moving Ile to an internal residue (23), as it appeared in the electron density. This also necessitated an insertion of a Gln into density previously modeled as solvent, after residue 24, and a subsequent repositioning of the turn at 26-27. Whereas the previous chemical sequence was -CPVN IQQG- and our previous X-ray sequence was -CPVNIIQG-, the corrected sequence became -CPVNCIQQG-. The sequence has since been confirmed chemically as well (LeTrong, Sieker and Adman, unpublished results). The bend at residues 26-27 moved into density as indicated by the difference map and produced a better fit there as well. The refinement then proceeded quite smoothly to an R of 0.137, with 89 solvent positions. We note that towards the end of the refinement it was possible to 1) use σ weights, 2) to refine using B-values, and 3) to keep the model reasonably tightly restrained (rms deviation 0.01Å although restrained to only 0.02Å). The average B-value over all atoms is 15.6Å², although that over protein and iron and sulfur atoms is more like 7-10Å². The plot of R vs $\sin\theta/\lambda$ indicates that more could probably be squeezed out of the low resolution data. There are no longer any significant

residual peaks near protein atoms.

Case 2. Rubredoxin from *Desulfovibrio vulgaris*. Although this structure was solved, partially refined, and reported at 2Å resolution in 1977⁴, the refinement using 1.5 Å data has just been completed (not because it was particularly difficult, but it has just been sort of an orphan project). Using PROLSQ, much model checking with graphics, and redetermination of solvent structure three times, the R-value now stands at 0.098 for infinity to 1.5Å data. We thought the model was complete at an R of 0.106; the solvent model seemed complete, geometry of the protein entirely reasonable. The remaining troublesome feature in the 'final' difference map occurred at residue 21. According to the sequence it was a threonine, which we believed was probably disordered since we had tried at least two alternative positions for it. A difference map with that side chain removed indicated clearly that in fact this side chain is an aspartic acid, which, had we been cognizant of the literature, had already been demonstrated from gene sequencing⁹. At the same time we also found a valine which had incorrectly labeled C γ atoms (indicated by a large shift peak at the C β position; again something which could have been prevented by closer attention to the voluminous output of PROLSQ). Correcting both of these, and judicious use of partial shifts in the refinement allowed the R-value to go to 0.098. As we argue in a separate publication soon to be submitted, it is still unlikely that we have reached the limit of the data. The Luzzati plot shows the expected lack of agreement for low order reflections due to incomplete solvent model. The take home lesson from this experience, aside from the fact that it is possible to refine a protein to a low R-value, is that even at an R of 0.106 there was still identifiable error.

Case 3. Ferredoxin II from *Desulfovibrio gigas*. This is not a case of residual error being found at the end of a refinement, but rather that during the refinement, modification of the chemical sequence was necessary, and that an unusual modification of a side chain was found. *D. gigas* ferredoxin was solved using the resolved anomalous phasing method¹⁰. Dr. Charles Kissinger worked on the development and refinement of the structure for his doctoral thesis work⁵. Refinement of the model did not commence until the model was nearly complete: that is, partial structure information was used only in developing combined phases, and judged useful only when new information came from the next map. During the refinement, again using PROLSQ (X-PLOR was tried on the structure at the end of the refinement, but produced no significant difference in the results), we found that a residue needed to be inserted after residue 55 in order that the subsequent residues fit density more closely. The new residue was judged to be a valine rather than a threonine from the side chain B-values. Similarly, density appended to Cys-11 was judged to be a S-CH₃, both from B-values and from atom distances. The final R for this structure is 0.157 for data from 10 to 1.7 Å. Unit weighting was used for this data, which possibly reduces the outer resolution shell R-values somewhat since higher resolution data will be weighted more heavily than lower in this case. The B-values for the protein center around 11-12 Å², and look reasonable in terms of the structure.

Case 4. *Azotobacter vinelandii* ferredoxin I. Refinement of the model for this ferredoxin has been carried out by Lyle Jensen, Hugh Stout, and more recently Ethan Merritt, mainly in order to confirm that their independently derived model in the correct space group was in fact correct. The details of how this came about has already been reported², and Dave Stout's contribution to this Study Weekend is much more revealing of why the earlier determination went astray. The refinement, at the lowest resolution reported in the present quartet of structures, has not revealed anything particularly unusual about it. A minimum number of solvent atoms has been included; the B-values for irons and sulfurs now seem in line with the rest of the protein. Experience with the previous three structures does suggest that there may still be more information in this data, but at 2.3Å it is a much harder job to extract it.

A question many of us would like to answer is how far is far enough to refine a protein model? It usually will be limited by the diffraction quality of the crystal being studied, but in cases where diffraction data is possible to high resolution, is one compelled to use it all? Or, turning the question around, when is an R-value low enough for the data in hand? Inasmuch as $\langle\sigma\rangle$ is a measure of expected $\langle\Delta F\rangle$, none of our refinements yet approaches the limit of the data.

I believe the answer lies not in the R-value itself, but in other "quality checks", some of which have been alluded to by others in this symposium. The behaviour of the refinement is a clue. A good model will refine to a low R-value with fairly tight distance restraints, e.g. as low as 0.015Å (at least at 2Å resolution). In our hands, using PROLSQ, progress in minimization also depends on using partial shifts, seemingly such that the actual shifts are on the order of magnitude of the expected errors for that R-value. Progress also depends on appropriate relative weighting of the structure factors to restraints. In some cases I have found that, again, with a good model, structure factors can be weighted even more than the commonly prescribed $\langle\sigma_{\text{applied}}\rangle = 1/2 \times \langle\Delta F\rangle$.

The behaviour of the R-value with resolution, e.g. the appearance of the Luzzati plot, yields some clues. The kind of weighting used will affect the appearance of this plot to some extent. For example, if the average error of higher resolution data is lower than that for lower resolution data, using unit weights tends to weight higher order data more than using σ weights would, and the higher order R-values may appear better than they should, leading to a lower estimate of coordinate error than is justifiable.

Certainly a check of reasonableness of geometry is called for, as is a check of close contacts within and between molecules. Some of these checks might be subsumed by over-tight restraints, but in turn the refinement should 'hang-up' if the tight restraints are not appropriate.

Most importantly, the final difference map should be checked carefully for significant peaks. In most of the cases described above, the sequence errors were detected upon evaluating "final" difference maps, and to the extent that these are interpretable, they are worth the most attention.

The reason for minimizing the R-value is to obtain a model with the least amount of error, so that one can then begin to pay attention to real deviations from expected structure. Even then it may not be possible. For example, I believe there is a significant distortion of τ angles at four C α carbons in the vulgaris rubredoxin structure, but the rms deviations in angles are on the order of the suspected distortion. Even with R-values as low as cited above, it is only on comparison of these well-refined structures that the distortion becomes evident as a systematic distortion.

At the risk of stating the obvious, because of the nature of the diffraction experiment, one has to obtain a good model even for the "non-interesting" parts of the structure, i.e. away from active sites, etc., in order to obtain a good model for the interesting parts. If the entire model is good, the protein structure is rich enough in information that perhaps *someone* will find something interesting even in the parts presently deemed "uninteresting"!

References

1. Stenkamp, R.E. Sieker, L.C. and Jensen, L.H. submitted
2. Stout, G.H., Turley, S., Sieker, L.C., and Jensen, L.H. Proc. Natl. Acad. Sci. 85 (1988) 1020-1022.
3. Adman, E.T., Sieker, L.C., and Jensen, L.H. J. Biol. Chem. 251 (1976) 3801-3806.
4. Adman, E.T., Sieker, L.C., Jensen, L.H., Bruschi, M., and LeGall, J. J. Mol. Biol. 112 (1977) 113-120.
5. Kissinger, C.R., Adman, E.T., Sieker, L.C., Jensen, L.H., and LeGall, J., FEBS Lett. 244 (1988) 447-450.
6. Hendrickson, W.A., and Konnert, J.H. In "Computing in Crystallography", Diamond, R., Ramaseshan, S., Venkatesan, K., Eds., Bangalore: Indian Institute of Science (1980) 13.01-13.33
7. Jones, T.A., J. Appl. Crystall. 11 (1978) 268-272.
8. Bruschi, M., and Guerlesquin, F., FEMS Microbiology Reviews, 54 (1988) 155-176.
9. Voordouw, G. Gene, 67 (1988) 75-83.
10. Hendrickson, W.A. and Teeter, M.M. Nature 290 (1981) 107-113.

UNCERTAINTY AND BIAS IN DIFFERENCE FOURIER MAPS

by

**Urszula Derewenda, Eleanor Dodson, Guy Dodson,
Dorothy Hodgkin and Helen Swift**

**Department of Chemistry, University of York, Heslington,
York, YO1 5DD, UK**

Traditionally difference Fourier maps calculated with coefficients $F_o - F_c$ or $2F_o - F_c$ have been a very powerful tool in the X-ray analysis of small molecules and macromolecules. In macromolecular studies difference Fourier maps are calculated after refinement, often essentially automatic, of starting coordinates in order to locate errors and to identify new structures. Experience has shown that if the atoms used in phasing are generally correct and if the resolution of the data extends to 2.5 Å or better the difference Fourier maps are very helpful. There are however problems. First, there is bias in the density in which ghosts of wrongly placed atoms used earlier in refinement can reappear. As the structure converges this effect will reduce. There are moreover methods of reducing this problems, using variously constructed Fourier coefficients ($3F_o - 2F_c$ for example) systematic difference Fourier or 'omit' maps and so on. When the atomic positions are substantially incorrect or the protein molecule is wrongly positioned, then the ghosting of excluded structures can become seriously misleading since it appears to lend confirmation for the positioning of these atoms.

There is a slightly different though connected problem when extending refinement to poorly defined atoms. Here the distinction between genuine peaks and spurious peaks can be very difficult. This uncertainty arises first because the electron density of poorly ordered side chains and solvent structures has a wide variation in peak shape and appearance. Thus any peak can be interpreted as a physically sensible structure and therefore the normal criteria of bond length and peak height cannot be easily used to decide on the validity of these low level peaks. Secondly the appearance of the low level density can change markedly during the refinement calculations requiring careful reinterpretation of atomic positions.

Bias in Difference Fourier Maps

α -Amylase - an example of bias associated with partial errors

Partial Errors

Molecular replacement presents a special problem since a significant part of the model structure can be in error. If this model is in addition subjected to refinement before inspection of the difference Fourier maps the incorrect segments can reappear with a relatively convincing appearance.

This kind of bias was observed in the TAKA -amylase refinement (Swift, DPhil Thesis, 1990). Here there proved to be a considerable error of several Ångstroms in the positioning of one of the three domains in the molecule. The rotation and translation calculations

appeared quite convincing and consequently refinement calculations were undertaken on the model with 2Å spacing data. The R factor (defined $R_{\text{cryst}} = \frac{\sum ||F_o| - |F_c||}{\sum |F_o|}$) fell to .28 after considerable refinement. Inspection of the 2Fo-Fc map revealed occasional breaks in main chain density and poor density for some of the side chains. It was moreover not often obvious how to proceed with the rebuilding. Figure 1 illustrates a portion of the electron density in the 2Fo-Fc map on the domain that was subsequently found to be wrongly positioned. The density for the incorrect structure can be seen to fit the atomic coordinates partially. There are few indications for the correct structure.

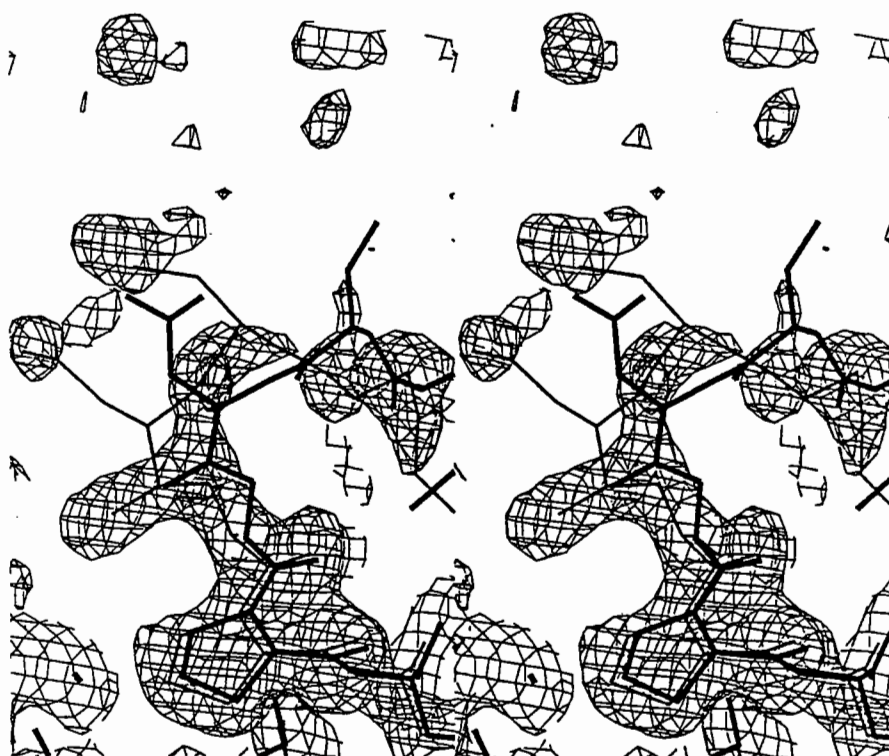


Figure 1. The 2Fo-Fc electron density shown in stereo for the TAKA α -amylase sequence pro-gly-asp-ser-gly, at $R = .28$. The incorrect coordinates used in the refinement calculation are shown as thin lines. The coordinates derived from the XPLOR calculations are connected by thick lines.

The high R factor (.28) after refinement and the appearance of the electron density together suggested the existence of errors in the model structure. A series of simulated annealing calculations were therefore carried out using the XPLOR package (Brunger, 1989). There were substantial movements in the one domain which were associated with a decrease in the R factor, to .24, and a much clearer 2Fo-Fc map. Figure 2 illustrates this map and shows how the density now corresponds well to the atomic positions first determined by the XPLOR calculations and then further refined. There is no sign of the original wrongly placed atoms - the dynamic simulations and the subsequent refinement calculations have succeeded in removing the bias.



Figure 2. The 2Fo-Fc electron density for TAKA α -amylase in stereo, showing the same features as Figure 1. It was calculated with coordinates obtained after the XPLOR simulation and followed by further least squares minimisation refinement; the R = .24. The coordinates derived from the original molecular replacement calculations are connected by thin lines; those from the XPLOR calculations and subsequent refinement by thick lines.

Dimeric insulin - an example of bias associated with a mispositioned molecule

The molecular replacement carried out on a dimeric insulin crystal illustrated the existence of bias in Fourier maps calculated from incorrect positioning of the whole molecule in the cell. In this orthorhombic crystal the asymmetric unit contained a dimer. The dimer structure that forms part of the 2Zn insulin hexamer was considered a suitable model for molecular replacement calculations. Data were collected to a resolution of 2.5 Å spacing on a CAD4 diffractometer (3 crystals) with a merging R of .047. The rotation function indicated a reasonably clear solution but the translation function calculated with projection data suggested more than one solution, two of which were most promising. Review of the crystal contacts made with these solutions showed both to be acceptable. Refinement of both solutions was then undertaken in the hope that a correct structure could be identified.

The initial crystallographic agreements for the two models were .595 and .615. After considerable refinement these reduced to .27 and .25 respectively. The inability to reduce the R value more indicated there were substantial errors in the model and a series of difference Fouriers were calculated to assess this. In one experiment the cysteine residues were removed from the phasing and the remaining atoms subjected to 4 cycles of least square refinement. The difference Fourier maps, Fo-Fc, showed the electron density for the cysteines reasonably well. It was not obvious from these maps that the models were incorrect. Peaks which had the appearance and contacts expected for water molecules were also present. Two factors made it likely however that there was an error in the molecular replacement solution. First was the inability to improve the agreement between the observed and calculated structure amplitudes to better than .25. Secondly the electron density for the main chain was occasionally broken and sometimes did not correspond nicely to the side chain shape.

A new data set extending to 2.0 Å spacing resolution was collected on the area detector. This data had an internal consistency (Rmerge) of .05. The molecular replacement calculations were repeated, this time using all the 3 dimensional data for the translational search in the whole SEARCH unit (not sections). While the rotational parameters were not changed the translational parameters were, indicating in this case a unique solution (see Table 1). Refinement with the dimer positioned with these new values began with Rcryst at .58 and led rapidly to a value of .22. Further refinement reduced the Rcryst to .17 and a clearly correct structure.

Table 1

The R search parameters

	x	y	z Å	Rcryst*	
				Initial	Final
Incorrect solution	20.32	14.58	6.33	.61	.25
Correct solution	8.63	23.71	6.33	.58	.17

* The Rcryst is calculated for the refinement carried out with the 2.5 Å data for the incorrect solution and 2.0 Å data for the correct solution.

In order to assess the behaviour of difference Fourier calculations based on correct and incorrect coordinates difference Fourier maps were recalculated on the 2.0 Å data set. For both the incorrect and correct solution the cysteines were removed at the stage in the refinement where Rcryst was .29, i.e. similar to the earlier calculations on the incorrect model with 2.5 Å data. In the same way as in the earlier incorrect solutions, the tailored structures were subjected to 4 cycles of refinement, to give Rcryst of about .32. These difference Fourier maps are illustrated in Figure 3(a) (the incorrect solution) and 3(b) (the correct solution). The correct structure produces a map which is less noisy and in which the cysteines are somewhat more clearly represented in the density (Figure 3b). The distinction might have been further strengthened by omitting water molecules which were introduced in the calculations at Rcryst below .30. Nonetheless in the absence of a map phased from a correctly placed model, it would not be easy to reject the incorrectly placed models on the basis of their difference density.

Uncertainty in Difference Fourier Maps

In the 2Zn insulin refinement we noticed that the appearance of difference density was markedly affected by the selection of atoms included in the phasing. Figure 4 illustrates the electron density at the 2Zn insulin hexamer centre which contains a network of connected water molecules. In spite of these water molecules being contained in a channel of only some 8 Å across, they are mostly poorly defined. The series of difference Fourier maps calculated during the refinement illustrated in Figure 4 shows how the density for the waters altered as the phasing improved. This is to be expected, but it was our experience that spurious peaks such as that labelled SW frequently refined with physically sensible thermal parameters and occupancies. In this

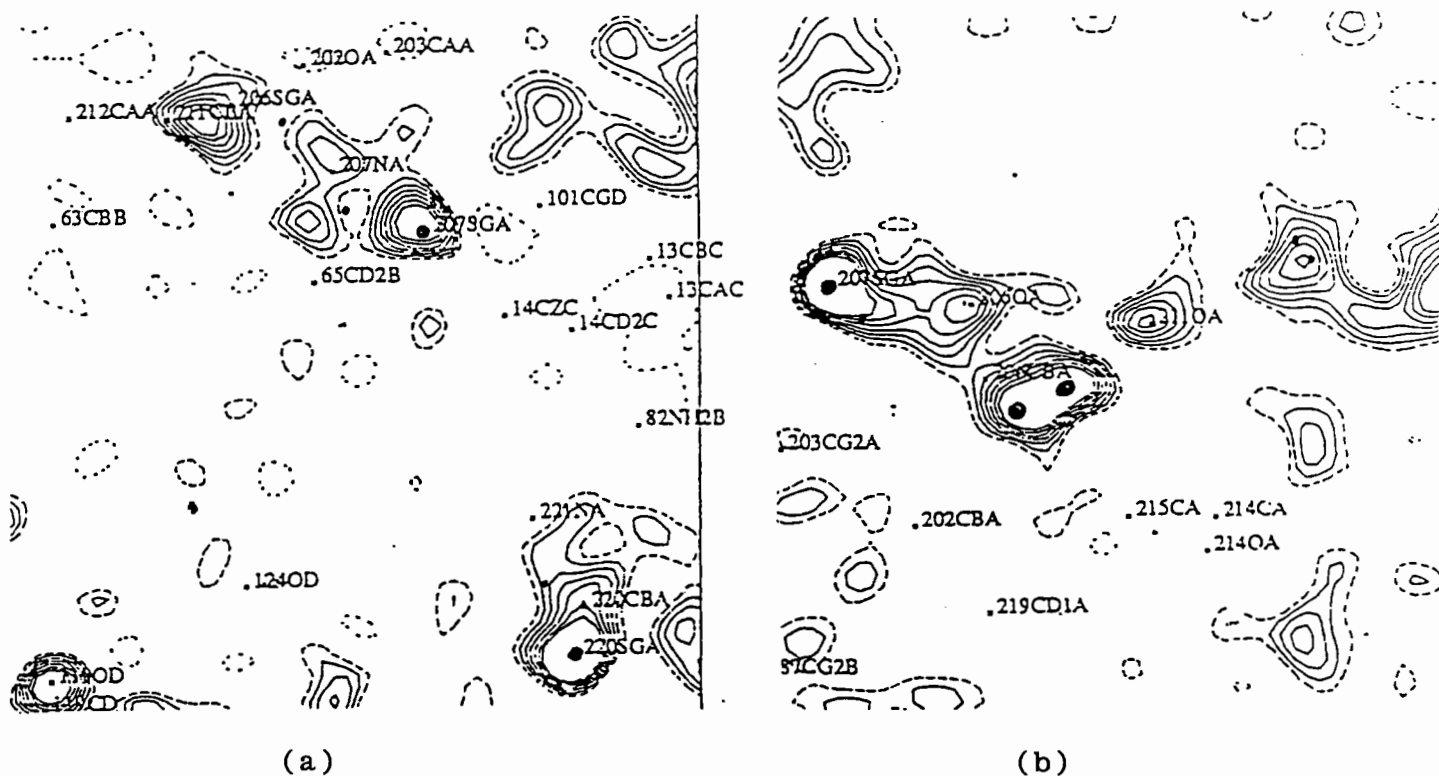


Figure 3. The difference Fourier density calculated for the incorrect and correct molecular replacement solutions of the insulin dimer in the orthorhombic crystal. In these calculations the 6 cysteines (A6 - A11, A7 - B7, A19 - B20) in both molecules were excluded, the remainder of the molecule and associated water structure subjected to 4 cycles of further refinement. The contour levels are $-.3e/\text{\AA}^2$ (- -) and $.14$ (in steps to $.14$) to $.56e/\text{\AA}^2$. The standard deviation in the maps is $.14e/\text{\AA}^3$.

Labelled atoms appear on the section. Unlabelled atoms correspond to atoms on the adjacent sections. Sulphur atoms are represented by large circles.

(a) An incorrect solution. The Rcryst is $.33$ and the map is calculated at section $z = 8/80$ with the 2.0 data set. The difference density shown is for atoms from cysteins A6 - A11 (206 - 211), A7 (207), A20 (220).

(b) The correct solution. The Rcryst is $.32$ and the map calculated on section $z = 5/80$ with the 2.0 \AA data set. The difference density is shown for the cysteines A6 - A11 (206, the sulphurs are on the adjacent section, 6, peptide atoms only), A7 (207).

particular case the water molecule positioned on the peak refined with a B of 37\AA^2 with full occupancy. The sensitivity of the low level difference density to the selection of atoms used in phasing seemed to be particularly pronounced near to atoms included in phasing. For this reason the 'omit' map was constructed in which typically $\frac{1}{8}$ of the asymmetric volume was excluded from the phasing in a systematic way. A complete difference 'omit' map could then be constructed. In our hands the map was generally noisier than a difference Fourier calculated with good phasing and thus did not always simplify the analysis of poorly defined atoms. On the other hand when combined with other difference Fouriers common peaks could be identified that were almost always correct. The density in omit maps calculated (at cycle 8) when atoms with $B \leq 12\text{\AA}^2$ were included in the excluded volume can be seen to change quite significantly however and enough to alter interpretation of water positions (Figure 5). It is significant that the spurious peak (SW) is greatly reduced by the inclusion of accurately determined atoms. It was clear that including the well defined and correctly placed atoms nearby has helped to improve the definition of the solvent positions considerably, e.g. (6W2).

Conclusion

The molecular replacement studies on the dimeric insulin illustrate the nature of the problems that can arise with refinement calculations on wrong solutions. It is clear that bias is a very real phenomenon in difference Fouriers phased on partially refined or wrongly refined macromolecular structures.

The discussion on the water density on the 2Zn insulin structure refers to a refinement in which the protein atom coordinates were generally correct and where uncertainties and variability in the appearance of the electron density occur in the low level density, i.e. $0.6e/\text{\AA}^3$ or less. The large thermal parameters and partial occupancies that characterise poorly defined side chain and solvent atoms means that it is possible to refine atoms in this density - right or wrong - quite acceptably. This uncertainty makes it essential that analysis and refinement of poorly ordered atoms is carried out with great care.

Bias in difference Fourier maps is the existence of peaks which correspond to incorrectly-placed atoms previously incorporated wrongly in the refinement calculation. The uniqueness of a crystallographic solution in the case of small molecules generally arises from the very similar curvature of the atoms (B values range typically between 2 - 5\AA^2), their individual resolution and their having proper

bond length and angles. In the absence of these constraints any electron density, no matter how phased, can be approximated to by atoms of variable atomic number (occupancy), variable thermal parameters and which have no required bond lengths and angles. Manipulation of atomic number (occupancy), B factor and position can therefore generate acceptable agreement between observed and calculated amplitudes. This structure will return in difference Fourier..

In proteins the range of curvatures and the limited resolution means that a reasonable match between the electron density and the model can almost always be achieved, even if the map contains serious errors or is, in the extreme case, quite wrong. Following refinement in which B values and molecular geometry are relaxed and water molecules are used to fill in difference density, the model can all too easily accommodate to the electron density map and appear to refine. Any part of this map or of the model will return to some extent in a difference Fourier and give the impression of correctness. It is however only a consistency that has been demonstrated. It is worth noting here that with data of resolution better than 2.5 Å the rms variation in the B values and the geometry of the atoms are valuable indicators for the physical sense and therefore correctness of a model.

The molecular replacement studies on the dimeric insulin illustrate the nature of the problems that can arise with refinement calculations on wrong solutions. It is clear that bias is a very real phenomenon in difference Fourier phased on partially refined or wrongly refined macromolecular structures.

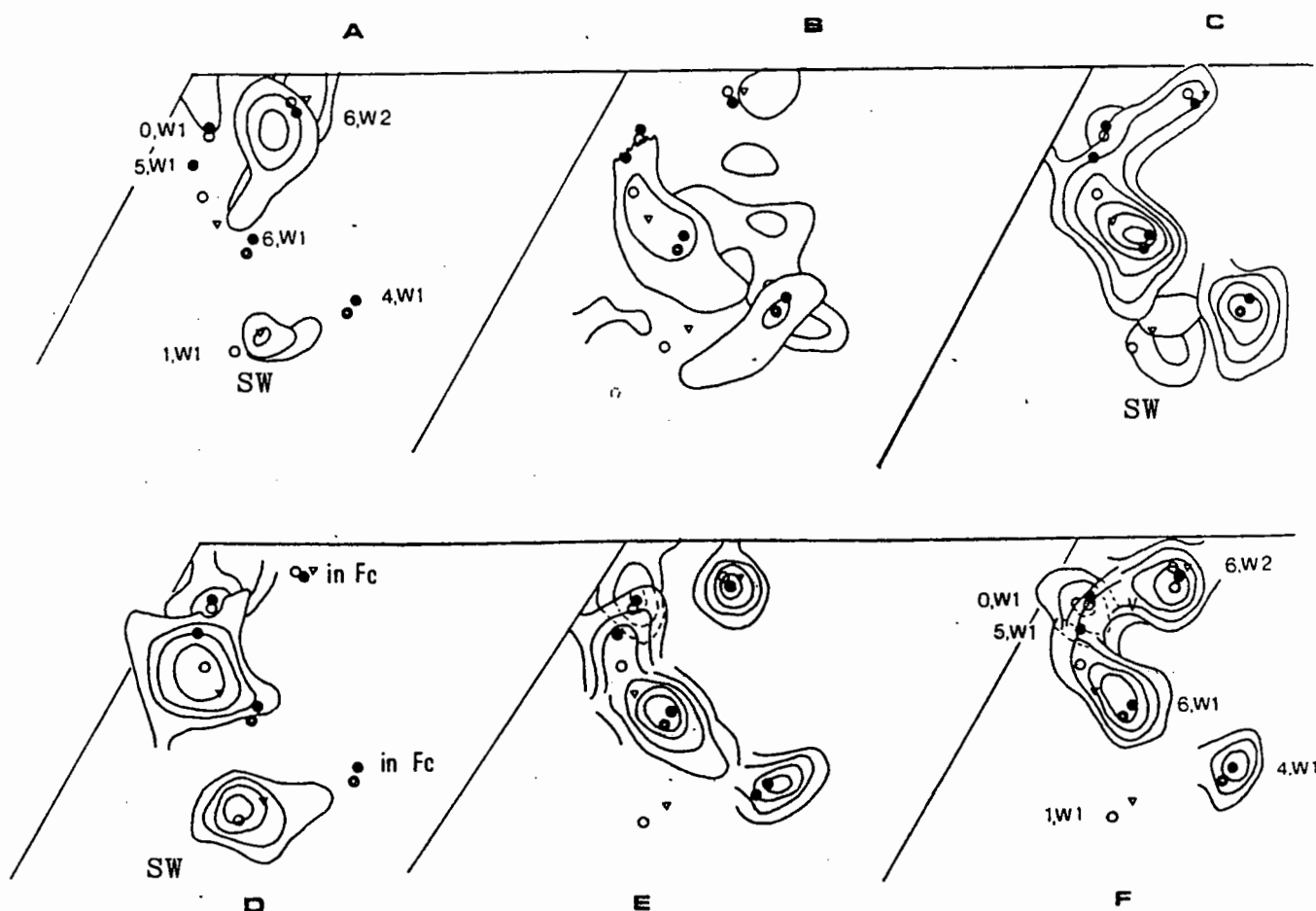


Figure 4. The electron density at the centre of the 2Zn insulin hexamer showing the appearance of the water molecule density at different stages of the analysis.

- A. The 1.9 Å spacing isomorphously phased map.
- B. Round 1; difference Fourier, no water in the phasing.
- C. Round 8; systematic difference Fourier (omit) map (atoms with $B \leq 12\text{\AA}^2$ molecules in the phasing of the excluded volume ($\frac{1}{8}$))
- D. Round 9; difference Fourier map (4W1, 6W2) included in the phasing. The large peak (SW) is spurious.
- E. Round 15; difference Fourier, all water excluded.
- F. Round 18; systematic difference Fourier map (atoms with $B \leq 12\text{\AA}^2$ included in excluded volume).

Note the persistence of the peak SW until round 15 (E). A water molecule positioned on this peak refined at RD14 with a thermal parameter B of 37\AA^2 .

The final water positions are indicated by black circles. The crosses make water positions placed in round 3. The open circles mark water positions in round 10.

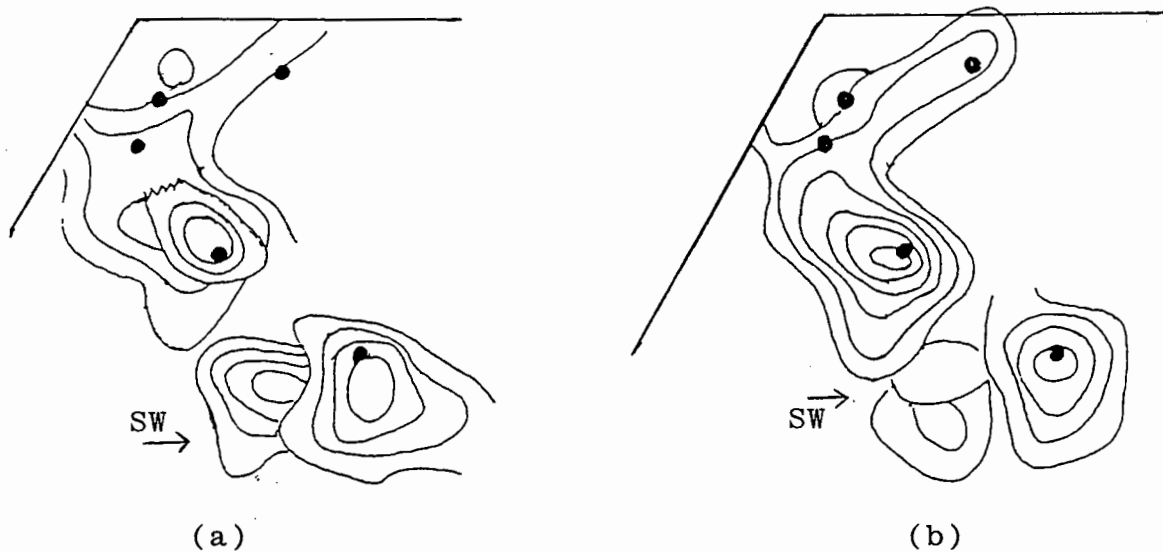


Figure 5. The effect of including well defined atoms ($B \leq 12\text{\AA}^2$) in the excluded volume of an omit map. The density associated with the spurious peak SW is larger in (a) where all atoms are excluded than in (b) where the well defined atoms are still included in the phases. The circles represent the final positions of water molecules.

Accuracy and Reliability of Macromolecular Crystal Structures

Closing Remarks by D.M. Blow

Blackett Laboratory, Imperial College, London SW7 2BZ

Over the past two days we have heard a series of fascinating papers. In order to stimulate discussion, I am trying to bring together some of the topics which might be discussed in conclusion.

When a protein structure is published, we would like to be able to assess whether it is accurate and reliable. Authors wish to provide data which will demonstrate the quality of their work. Referees have to decide whether the work merits publication, and might wish to set some minimum standards. Ideally, quantitative indicators of quality might be set, but in practice all the possible indicators interact. For example, it may be easy to idealise the stereochemistry of a peptide chain, but may be difficult to avoid unsatisfactory non-bonded distances at the same time.

There is a danger that if specific criteria are set to judge the quality of a structure, authors, or their computer programs, will "learn" to perform well at these criteria, and the problems of a structure will be revealed only by examining different criteria.

The Table summarises some of the indicators of quality which have been mentioned during the meeting. A future task for CCP4 will be to consider whether practice ought to be standardised in calculating or providing some of these quality indicators. It is important that our programme suite makes the right criteria available, and that they are calculated in the proper way.

One vexed question concerns "unobserved" reflections - reflections which are too weak to measure. We have come a long way since reflections which were measured to be weaker than the surrounding backgrounds were simply omitted, or set to zero intensity; analysis by the Bayesian statistical approach provides a weak positive intensity as the "expected" intensity, with a standard deviation of comparable size. But when it comes to presenting reliability factors (R factors) of various kinds, it seems that some authors prefer to ignore their weak or "unobserved" reflections.

A related question arises over the "resolution" of the data. Should we measure every observable reflection, even though the outermost bin contains hundreds of "unobserved" reflections for every one which has a significant intensity? It is arguable that all those unobservable reflections have a significant information content. But most workers prefer to truncate their data at a point where most of their reflections are measurable, and before large numbers of very weak intensities begin to degrade the overall R-factor.

Similar problems arise about the resolution limit for isomorphous derivatives. Isomorphous differences often appear to increase at high θ , but the r.m.s. lack of closure (" E ") increases even more rapidly. A generally accepted criterion is that isomorphous data are not worth having when the lack of closure exceeds the differences.

The most risky stages of a structure determination are at the refinement. There are plenty of examples where structural refinement has "stuck" with an R-factor in the high twenties or low thirties, until some error of interpretation has been discovered. But what should be done when the refinement sticks in the low twenties? How can that elusive error be found? Or, is it not more likely that there is no error, just problems due to disorder or bad crystals? Sometimes we will never know, but surprisingly often a new approach comes along, yielding a new crystal form or better data, and leading to a better structure. Was the earlier structure "wrong", or just a bit worse?

Table

Some Quality Indicators for Crystal Structure Data

<u>X-ray data:</u>	Quality: R_{merge} , R_{iso} Overall B- how determined? - resolution limit - where to stop taking data Completeness
<u>Isomorphous replacement</u>	Scaling - Heavy atom sites - anomalous data Criteria for map quality of mir map Figure of merit, phasing power vs resolution E (r.m.s. "lack of closure")
<u>Molecular replacement</u>	Resolution range - peak significance
<u>Interpretation and refinement</u>	Chain tracing "rules" Resolution range - early and late How to monitor progress Difference maps/residual maps When to rebuild? Problems of specific refinement programs Omit maps How to handle probable sequence errors
<u>Structural quality</u>	Ideal geometry Distribution of B's Ramachandran and χ distributions R-factor - how calculated? distribution with $\sin \theta$ and $ F $ Phase tests: isomorphous and ligand difference maps "Reasonability" Comparison with identical or homologous molecules Bad contacts- intramolecular - intermolecular
<u>Data base</u>	Avoidance of blunders Errors in matrices Levels of checking Do we want a true historic record or the best approximation to the truth?

It is inevitable that the labour intensive, detailed examination of maps and models, required at this stage of structural refinement will be replaced by automatic computational procedures. At the moment I think that XPLOR is much more expensive, and less likely to be successful, than meticulous and critical study by the human brain. It is right to be very suspicious of new automatic procedures until they have thoroughly proved themselves.

When we come to the final structure as it is to be deposited in the data bank, there is a wealth of different criteria for judgement. A striking feature which has emerged in the last few years is how, in a high quality structure determination, the α -helical conformational angles cluster tightly onto the ideal angles, and the other conformational angles remove themselves from the "forbidden" regions. Though we all know that α -helices have kinks in

them, and that other conformations are sometimes strained, it is remarkable how Pauling's and Ramachandran's classical ideas are vindicated.

When the structure becomes a mature one, one of a family of homologues whose structure has been determined, or which has been determined many times in different complexes, new truths will emerge. New criteria are then available to show which of a family of structures is more likely to be highly accurate, which are less so, and which (yes, it happens!) are rubbish. Never forget that in ten or twenty years' time, your brand new structure will stand for comparison with a dozen similar ones, and it will be obvious how good it is (or not).

The talk by Frances Bernstein on the Protein Data Bank brought home a number of points I had not thought about. One could make such a data bank in a number of ways. It could attempt to tell you the real structures, updating every structure in the light of the latest information and methods of analysis. It could be judgmental, telling you clearly which structures are reliable, and which are not; what features of older structures ought to be reassessed in the light of newer structural data using better techniques. Or it could be strictly historic, depositing the co-ordinates for a particular structure determination in an immutable form, an archive of the development of our trade. The trustees of the data bank have adopted none of these approaches. The data bank contains what the author wants you to know about his structure. It is changed when he wants, and nobody else changes it - not, at least, without his permission. It is important to be aware of this, so that the data bank is used in the proper way.

Some may argue a need for different types of data bank. It would be nice to have a bank of "true" data, but that is impossible. There may be a need for a simplified data bank which can be used for certain purposes by users who are not capable of being critical about the information which they are trying to use. But our bankers have adopted a clear and simple policy, which it is important for us to understand.

This meeting has pointed to a number of areas where the CCP can itself improve the quality and criticality of the information which we produce. I hope the working groups will find ways to achieve this.

LIST OF DELEGATES

- Abeyasinghe, I.S.B.
Department of Biochemistry, University of
Sheffield, Sheffield S10 2TN.
- Achari, A.
Genex Corporation, 16020 Industrial Drive,
Gaithersburg, MD 20877, U.S.A.
- Adams, M.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Adams, P.D.
Department of Biochemistry, University of
Edinburgh, Hugh Robson Building, George
Square, Edinburgh EH8 9XD.
- Adman, E.T.
Department of Biological Structure, School
of Medicine, University of Washington,
Seattle, WA 98195, USA.
- Artymiuk, P.
Krebs Institute, Department of
Biochemistry, University of Sheffield,
Sheffield S10 2TN.
- Bailey, S.
Department of Physics, University of
Keele, Keele, Staffs ST5 5BG.
- Baker, P.J.
Department of Molecular Biology &
Biotechnology, University of Sheffield,
Sheffield S10 2TN.
- Barford, D.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Barrett, T.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Barton, G J
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Basak, A.K.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Beavil, A.
Department of Biomolecular Sciences,
King's College London, 26 Drury Lane,
London WC2B 5RL.
- Bernstein, F.C.
Department of Chemistry, Brookhaven
National Laboratory, Upton, New York 11973,
USA.
- Bing, X.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Bloomer, A.C.
Laboratory of Molecular Biology, Medical
Research Council, Hills Road,
Cambridge CB2 2QH.
- Blow, D.
Blackett Laboratory, Imperial College,
Prince Consort Road, London SW7 2BZ.
- Blundell, T.
Department of Crystallography, Birkbeck
College, Malet Street, London WC1E 7HX.
- Bolognesi, M.
Department of Crystallography, University
of Pavia, Via Taramelli 16, I-27100 Pavia,
Italy.
- Bordas, J.
Daresbury Laboratory.
- Brick, P.
Blackett Laboratory, Imperial College,
Prince Consort Road, London SW7 2AZ.
- Britton, L.
Department of Biology and Biotechnology,
University of Sheffield, Sheffield S10 2TN.
- Brown, D.G.
Institute of Cancer Research, Block F,
Royal Marsden Hospital, 15 Cotswold Road,
Sutton, Surrey SM2 5NG.
- Brownlie, P.
Department of Crystallography, Birkbeck
College, Malet Street, London WC1E 7HX.
- Brzozowski, A.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Cameron, A.D.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Campbell, J.W.
Daresbury Laboratory.
- Caves, L.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Cervi, A.R.
Department of Chemistry, University of
Manchester, Manchester M13 9PL.
- Cleasby, A.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Clifton, I.J.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.

- Daly, P.J.
Daresbury Laboratory.
- Davies, C.
Department of Biochemistry, School of
Medical Sciences, University of Bristol,
Bristol BS8 1TD.
- Davies, G.
Department of Biochemistry, University of
Bristol, Bristol BS8 1TD.
- Derewenda, U.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Derewenda, Z.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Derrick, J.P.
Department of Biochemistry, University of
Leicester, Leicester LE1 7RH.
- Dideberg, O.
Institute of Physics, B5 Crystallography,
University of Liege, 4000 Liege, Belgium.
- Dodson, E.J.
Department of Physics, University of York,
Heslington, York YO1 5DD.
- Dodson, G.G.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Driessen, H.P.C.
Department of Crystallography, Birkbeck
College, Malet Street, London WC1E 7HX.
- Duke, E.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Eaton, J.
Department of Chemistry, University of
Glasgow, Glasgow G12 8QQ.
- Edwards, D.J.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Edwards, K.J.
Institute of Cancer Research, Block F,
Royal Marsden Hospital, 15 Cotswold Road,
Sutton, Surrey SM2 5NG.
- Eliopoulos, E.
Department of Biophysics, University of
Leeds, Leeds LS2 9JT.
- Evans, P.R.
Medical Research Council, Laboratory of
Molecular Biology, Hills Road,
Cambridge CB2 3QH.
- Evans, G.
EMBL, Notkestrasse 85, Building 25A,
D-2000, Hamburg 52, FRG.
- Flower, D.R.
Astbury Department of Biophysics,
University of Leeds, Leeds LS2 9JT.
- Ford, G.C.
Krebs Institute, University of Sheffield,
Sheffield S10 2TN.
- Freemont, P.S.
Imperial Cancer Research Fund, Protein
Structure Laboratory, 44 Lincolns' Inn
Fields, London WC2A.
- Freer, A.
Department of Chemistry, University of
Glasgow, Glasgow G12 8QQ.
- Frigerio, F.
Department of Crystallography, University
of Pavia, Via Tarmelli, 16, I-27100 Pavia,
Italy.
- Fulop, V.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Gamblin, S.
Department of Biochemistry, University of
Bristol, Bristol BS8 1TD.
- Garman, E.F.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Geddes, A.
Astbury Department of Biophysics,
University of Leeds, Leeds LS2 9JT.
- Ghosh, M.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1
3QU.
- Gibbs, M.R.
Laboratory of Molecular Biology, Medical
Research Council, Hills Road,
Cambridge CB2 2QH.
- Glover, I.
Department of Physics, University of Keele,
Keele, Staffs ST5 5BG.
- Glumoff, T.
Laboratory of Biochemistry 1, Swiss Federal
Institute of Technology, ETH-Zentrum,
Universitat strasse 16, CH-8092 Zurich,
Switzerland.
- Goldberg, J.D.
Blackett Laboratory, Imperial College,
Prince Consort Road, London SW7 2AZ.
- Gonzalez, A.
Daresbury Laboratory.

- Gordon, E.J.
Department of Biochemistry, University
of Edinburgh, George Square,
Edinburgh EH8 9XD.
- Gover, S.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Green, T.P.
Department of Molecular Biology and
Biotechnology, University of Sheffield,
Sheffield S10 2TN.
- Griest, R.E.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Groom, C.R.
Astbury Department of Biophysics,
University of Leeds, Leeds LS2 9JT.
- Gros, P.
Laboratory of Chemical Physics, University
of Groningen, Nyenborgh 16, 9747 AG
Groningen, The Netherlands.
- Guthrie, N.
Department of Chemistry, University of
Glasgow, Glasgow G12 8QQ.
- Habash, J.
Department of Chemistry, University of
Manchester, Manchester M13 9PL.
- Hajdu, J.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Harding, M.M.
Department of Chemistry, University of
Liverpool, PO Box 147, Liverpool L69 3BX.
- Harlos, K.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Harris, G.W.
AFRC Institute of Food Research, Reading
Laboratory, Shinfield, Reading,
Berks RG2 9AT.
- Harris, D.
Department of Chemistry, University of
Glasgow, Glasgow G12 8QQ.
- Harrop, S.J.
Department of Chemistry, University of
Manchester, Manchester M13 9PL.
- Hasnain, S.S.
Daresbury Laboratory.
- Hemmings, A.M.
Department of Crystallography, Birkbeck
College, Malet Street, London WC1E 7HX.
- Henrick, K.
Daresbury Laboratory.
- Hilgenfeld, R.
Central Research/Biotechnology, Hoechst AG,
PO Box 800320, D-6230 Frankfurt 80, FRG.
- Holden, P.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Hovmoller, S.
Department of Structural Chemistry,
University of Stockholm, S-10691 Stockholm,
Sweden.
- Hu, S-H.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Hunter, W.N.
Department of Chemistry, University of
Manchester, Manchester M13 9PL.
- Isaacs, N.W.
Department of Chemistry, University of
Glasgow, Glasgow G12 8QQ.
- Islam, S.
Imperial Cancer Research Fund, Lincoln's
Inn Fields, London WC2
- Ito, N.
Astbury Department of Biophysics,
University of Leeds, Leeds LS2 9JT.
- Janes, W.
Hoffman-La Roche AG, Abt ZFE/PHY BAU65/301,
CH4002 Basel, Switzerland.
- Jeffrey, P.
Squibb Institute for Medical Research, PO
Box 4000, Princeton, New Jersey 08543-4000,
USA.
- Jenkins, J.
Laboratoire de Biologie Physicochimique,
Batiment 433, Universite Paris-Sud, 91405
Orsay Cedex, France.
- Jhoti, H.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richard Building,
South Parks Road, Oxford OX1 3QU.
- Jiang, J.S.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Jones, Y.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Knight, S.
Swedish University of Agricultural
Sciences, Uppsala Biomedical Center,
Department of Molecular Biology, Box 590,
Sweden

- Kokkinidis, M.
Institute of Molecular Biology and
Biotechnology, P O Box 1527, GR-711 10
Iraklion, Crete, Greece.
- Lahm, A.
EMBL, Postfach 10.2209, 69 Heidelberg,
FRG.
- Leslie, A.G.W.
MRC, Laboratory of Molecular Biology, Hills
Road, Cambridge CB2 2QH.
- Lewit-Bentley, A.
LURE, Batiment 209D, Universite Paris Sud,
F-91405 Orsay, France.
- Li, J.
MRC Laboratory of Molecular Biology, Hills
Road, Cambridge CB2 2QH.
- Lindahl, M.
Molecular Biophysics Chemical Centre,
University of Lund, Box 124, S-22100,
Sweden.
- Lindley, P.
Department of Crystallography, Birkbeck
College, Malet Street, London WC1E 7HX.
- Littlechild, J.A.
Department of Biochemistry, University of
Bristol, University Walk, Bristol BS8 1TD.
- Littlejohn, A.
Department of Chemistry, University of
Glasgow, Glasgow G12 8QQ.
- Logan, D.T.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Lutter, R.
MRC Laboratory of Molecular Biology, Hills
Road, Cambridge CB2 2QH.
- MacArthur, M.W.
Laboratory of Molecular Biology, Department
of Crystallography, Birkbeck College, Malet
Street, London WC1E 7HX.
- Mallinson, P.R.
Department of Chemistry, University of
Glasgow, Glasgow G12 8QQ.
- Martin, J.L.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Mason, S A
Institut Laue Langevin, BP 156X, 38042
Grenoble Cedex, France.
- McAlpine, A.S.
Department of Biochemistry, University of
Edinburgh, Hugh Robson Building, George
Square, Edinburgh EH8 9XD.
- McDonald, N.
Department of Crystallography, Birkbeck
College, Malet Street, London WC1E 7HX.
- McLaughlin, P.J.
Medical Research Council, Laboratory of
Molecular Biology, Hills Road,
Cambridge CB1 3QH.
- Moody, P.C.E.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Morais Cabral, J.
Department of Biochemistry, University of
Edinburgh, Hugh Robson Building, George
Square, Edinburgh EH8 9XD.
- Moss, D.S.
Department of Crystallography, Birkbeck
College, Malet Street, London WC1E 7HX.
- Murphy, L.M.
Daresbury Laboratory.
- Murray-Rust, J.
Department of Crystallography, Birkbeck
College, Malet Street, London WC1E 7HX.
- Muskett, F.W.
Department of Biochemistry, University of
Edinburgh, Hugh Robson Building, George
Square, Edinburgh EH8 9XD.
- Myles, D.
Department of Physics, University of Keele,
Keele, Staffs ST5 5BG.
- Nave, C.
Daresbury Laboratory.
- Neil, T.K.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Newman, M.P.
Department of Crystallography, Birkbeck
College, Malet Street, London WC1E 7HX.
- Noble, M.
EMBL, Postfach 10.2209, Meyerhofstrasse 1,
D-6900 Heidelberg, West Germany.
- North, A.C.T.
Astbury Department of Biophysics,
University of Leeds, Leeds LS2 9JT.
- Nunn, R.
Department of Molecular Biology and
Biotechnology, University of Sheffield,
Sheffield S10 2TN.
- Ogg, D.J.
Symbicon AB, Glunten, Uppsala S-75183,
Sweden.
- O'Hara, B.
Department of Crystallography, Birkbeck
College, Malet Street, London WC1E 7HX.

- Onesti, S.C.E.
Blackett Laboratory, Imperial College,
Prince Consort Road, London SW7 2AZ.
- Pelosi, G.
Institute of General & Inorganic Chemistry,
University of Parma, Viale delle Scienze,
I-43100 Parma, Italy.
- Petratos, K.
EMBL, Notkestrasse 85, Building 25A,
D-2000, Hamburg 52, FRG.
- Pfeffer, S.
EMBL, Notkestrasse 85, Building 25A,
D-2000, Hamburg 52, FRG.
- Phillips, K.
Astbury Department of Biophysics,
University of Leeds, Leeds LS2 9JT.
- Phillips, S.
Astbury Department of Biophysics,
University of Leeds, Leeds LS2 9JT.
- Popov, S.
Institute of Crystallography, Moscow
Academy of Sciences, Moscow, U.S.S.R. and
Department of Chemistry, University of
Manchester, Manchester M13 9PL.
- Rafferty, J.B.
Astbury Department of Biophysics,
University of Leeds, Leeds LS2 9JT.
- Raftery, J.
Department of Structural Chemistry,
University of Manchester,
Manchester M13 9PL.
- Rao, Z H.
Sir William Dunn School of Pathology,
University of Oxford, South Parks Road,
Oxford OX1 3RE.
- Rawas, A.
Department of Biochemistry, University of
Bristol, University Walk, Bristol BS8 1TD.
- Ren, J.
Laboratory of Molecular Biophysics,
University of Oxford, South Parks Road,
Oxford OX1 3QU.
- Rice, D.W.
Department of Molecular Biology and
Biotechnology, University of Sheffield,
Sheffield S10 2TN.
- Richard, V.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Rizkallah, P.J.
Daresbury Laboratory
- Rob, A.
Department of Biochemistry, University of
Sheffield, Sheffield S10 2TN.
- Rodgers, H.F.
Department of Biology and Biotechnology,
University of Sheffield,
Sheffield S10 2TN.
- Rollett, J.S.
University Computing Lab, University of
Oxford, 8-11 Keble Road, Oxford OX1 3QD.
- Roth, M.
Institut Laue-Langevin, BP 156X, 38042
Grenoble Cedex, France.
- Rule, S.
Department of Physics, University of Keele,
Keele, Staffs ST5 5BG.
- Sanderson, M.R.
Institute of Cancer Research, Block F, 15
Cotswold Road, Sutton, Surrey SM2 5NG.
- Sawyer, L.
Department of Biochemistry, University of
Edinburgh, Hugh Robson Building, George
Square, Edinburgh EH8 9XD.
- Schreuder, H.
Molecular Biology Institute, University of
California, Los Angeles, CA 90024, USA.
- Sharff, A.J.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.
- Shaw, A.
Biochemistry Section, Department of Biology
and Biotechnology, University of Sheffield,
Sheffield S10 2TN.
- Shrive, A.
Department of Physics, University of Keele,
Keele, Staffs ST5 5BG.
- Skarzynski, T.
Blackett Laboratory, Imperial College,
Prince Consort Road, London SW7 2BZ.
- Skelly, J.V.
Institute of Cancer Research, Block F,
Royal Marsden Hospital, 15 Cotswold Road,
Sutton, Surrey SM2 5NG.
- Smerdon, S.J.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Smit, J.D.G.
Laboratorium fur Biochemie, ETH Zurich,
Universitastrasse 16, CH-3092 Zurich,
Switzerland.
- Somers, D.
Laboratory of Molecular Biophysics,
University of Oxford, Rex Richards
Building, South Parks Road, Oxford OX1 3QU.

- Somers, W.S.
Astbury Department of Biophysics,
University of Leeds, Leeds LS2 9JT
- Steigemann, W.
Max-Planck Institut für Biochemie am
Klopferspitz, 8033 Martinsried, FRG.
- Stein, P.
Department of Haematology, Medical Research
Council, Hills Road, Cambridge CB2 2QH.
- Stillman, T.J.
Department of Biology and Biotechnology,
University of Sheffield, Sheffield
S10 2TN.
- Stout, C.D.
Department of Molecular Biology, Research
Institute of Scripps Clinic, 10666 N Torrey
Pines Road, La Jolla, California 92037,
USA.
- Strathdee, S.D.
Astbury Department of Biophysics,
University of Leeds, Leeds LS2 9JT.
- Sussman, J.
Crystallography Laboratory, NCI-Frederick
Cancer Research Facility, Box B, Frederick,
Maryland 21710, USA.
- Sutton, B.J.
Department of Biomolecular Sciences, King's
College London, 26 Drury Lane,
London WC2B 5RL.
- Swift, H.J.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Tame, J.
Laboratory of Molecular Biology, Medical
Research Council, Hills Road,
Cambridge CB2 2QH.
- Thomas, D.
Department of Biochemistry, University of
Sheffield, Sheffield S10 2TN.
- Thompson, A.W.
Daresbury Laboratory.
- Thornton, J.M.
Laboratory of Molecular Biology, Department
of Crystallography, Birkbeck College, Malet
Street, London WC1E 7HX.
- Tickle, I.J.
Department of Crystallography, Birkbeck
College, Malet Street, London WC1E 7HX.
- Tierney, S.
Department of Chemistry, University of
Glasgow, Glasgow G12 8QQ.
- Todd, R.
Laboratory of Molecular Biology, Medical
Research Council, Hills Road,
Cambridge CB2 2QH.
- Tranqui, D.
CNRS, Laboratoire de Cristallographie, 166X
F38042 Grenoble, France.
- Tucker, P.A.
Biological Structures Programme, EMBL,
Postfach 10.2209, Meyerhofstrasse 1, 6900
Heidelberg, FRG.
- Turkenburg, J.P.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Turkenburg, M.G.W.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.
- Turner, M.
Department of Biochemistry, University of
Edinburgh, Hugh Robson Building, George
Square, Edinburgh EH8 9XD.
- Veerapaneni, N.
Department of Crystallography, Birkbeck
College, Malet Street, London WC1E 7HX.
- Vidgren, J.
Molecular Biophysics Chemical Centre,
University of Lund, Box 124, S-22100,
Sweden.
- Vrielink, A.
Blackett Laboratory, Imperial College,
Prince Consort Road, London SW7 2AZ.
- Wallace, B.A.
Department of Crystallography, Birkbeck
College, Malet Street, London W1E 7HX.
- Waller, D.A.
Astbury Department of Biophysics,
University of Leeds, Leeds LS2 9JT.
- Walls, P.H.
Laboratory of Biomolecular Modelling,
PO Box 123, Lincoln's Inn Fields,
London WC2A 3PX.
- Wan, T.
Department of Biomolecular Sciences, King's
College London, 26-29 Drury Lane,
London WC2B 5RL.
- Watson, H.C.
Department of Biochemistry, University of
Bristol, University Walk, Bristol BS8 1TD.
- Webster, G.D.
Institute of Cancer Research, Block F,
Royal Marsden Hospital, 15 Cotswold Road,
Sutton, Surrey SM2 5NG.
- White, J.
Biological NMR Centre, University of
Leicester, Leicester LE1 7RH.
- Whittingham, J.L.
Department of Chemistry, University of
York, Heslington, York YO1 5DD.

Wigley, D.B.
Department of Biochemistry, University of
Leicester, Leicester LE1 7RH.

Wonacott, A.
Blackett Laboratory, Imperial College,
Prince Consort Road, London SW7 2BZ.

Yewdall, S.J.
Department of Molecular Biology and
Biotechnology, University of Sheffield,
Western Bank, Sheffield S10 2TN.

Young, R.
Department of Biomolecular Sciences, King's
College London, 26 Drury Lane, London
WC2B 5RL.

Young, F.
Department of Chemistry, University of
Glasgow, Glasgow G12 8QQ.

