Conference Proceedings
DL-CONF-96-001

Macromolecular Refinement

Proceedings of the CCP4 Study Weekend
January 1996

E Dodson M Moore A Ralph and S Bailey

August 1996

COUNCIL FOR THE CENTRAL LABORATORY OF THE RESEARCH COUNCILS

Enquiries about copyright, reproduction and requests for additional copies of this report should be addressed to:

The Central Laboratory of the Research Councils
Chadwick Library
Daresbury Laboratory
Daresbury
Warrington
Cheshire
WA4 4AD
Tel: 01925 603397 Fax: 01925 603195
E-mail library@dl.ac.uk

ISSN 1362-0223

MACROMOLECULAR REFINEMENT

Proceedings of the CCP4 Study Weekend
January 1996

Compiled by:

E Dodson University of York
M Moore University of York
A Ralph Daresbury Laboratory
S Bailey Daresbury Laboratory

CCLRC
Daresbury Laboratory
1996

CONTENTS

INTRODUCTION

The approach to the refinement of macromolecular models has changed radically in the past few years. Structures are solved more quickly, and much less time is now devoted to understanding their details. This has partly arisen because more powerful algorithms such as those encoded in 'X-PLOR' make it apparently easy to reach a satisfactory R-factor, and partly because the scientists are often now primarily biochemists with little experience in the finer points of the mathematical basis of refinement.

This meeting covered the underlying theory of different types of refinement which are commonly used and also explored new techniques. The use of maximum likelihood refinement was covered in some detail. It has recently been implemented in two new programs and appears to be a very promising new option. Insight into how to handle motion in crystals was provided by several speakers. There was discussion on indicators which can detect both gross errors and sub-optimal refinement. Various case studies were used to give examples of different types of refinement problems.

The meeting was held at Chester College again this year. The facilities available at this venue being suited to the large number of participants. There were 385 participants in total, including 99 participants from Europe and 14 from the Americas. 55 of the Europeans participants were young scientists who were able to come to the meeting due to support from the EC Human Capital and Mobility Scheme. The speakers comprised 13 from the UK, 5 from elsewhere in Europe and 6 from the USA and Canada.

The meeting was organised and supported by the BBSRC Collaborative Computational Project in Protein Crystallography (CCP4) and the EC Human Capital and Mobility Scheme. We thank the invited speakers for sharing their expertise with us and for the contributions to this booklet. We are very grateful to Daresbury Laboratory for providing organisational support.

Eleanor Dodson
Madeleine Moore
Adam Ralph
Sue Bailey

July 1996

ACKNOWLEDGMENTS

CCP4 would like to thank the EC Human Capital and Mobility Scheme for the provision of funding which allowed 55 young European scientists to attend the 1996 CCP4 meeting.

CCP4 would also like to thank the following companies for their financial contributions to the CCP4 project in the year 1995. This support was an essential contribution to the costs of the meeting.

Abbott Laboratories
Amgen Incorporated
Ariad
Banyu Pharmaceutical Company Ltd
Bristol- Myers Squibb
Chugai Pharmaceuticals
Ciba-Geigy Limited
Dupont
Genentech, Incorporated
Genetics Institute
Glaxo UK Ltd
Green Cross Corporation
Hoffman La Roche and Co
Japan Tobacco Incorporated
Kyowa Hakko
K. Thomae
Mitsubishi
Pfizer Limited
Pharmacia AB
Pharmacia SpA
Rhone Poulenc
Sandoz Pharma AG
Schering
SmithKline
Syntex (USA) Incorporated
The Wellcome Foundation Limited
Vertex
Wyeth-Ayrst Laboratories
Yamanouchi
Zeneca Pharmaceuticals
Zeneca USA

The Limits of Interpretation

Dale E. Tronrud
Howard Hughes Medical Institute
Institute of Molecular Biology
University of Oregon

Abstract

The standard method of refining a macromolecular model uses both automated and manual methods. This combination allows the best abilities of both the computer and the human to be applied to the problem. At a basic level, however, both methods are examining the same indicators of error. This paper discusses some of the properties of these indicators which limit the investigator's ability to identify errors in their models.

Introduction

Our automated refinement packages are limited in that they cannot alter the basic form of the models they are optimizing. Initially the model must be constructed . Interspersed with the automated refinement are sessions of manual intervention. During these sessions at the computer graphics workstation the crystallographic information is presented in the form of density and difference density maps. To properly interpret these maps you must have an understanding of the way errors are represented in these maps and kinds of information not shown by them.

Usually one examines a Fo-Fc map to identify errors in a model and a 2Fo-Fc map to guide the construction of the new model. Since the Fo-Fc map is used to detect the error most of this paper will be devoted to a description of the appearance of these maps.

The first order description of the signal in a Fo-Fc map is know to all crystallographers. Locations in space where there should be electrons show positive features in the map while locations where the model inappropriately contains electrons show negative features. For example, if the model is missing a bound water molecule the Fo-Fc map will show a positive peak at the location where the water molecule should be placed.

A more complicated signal is expected when an atom is modeled but is slightly misplaced. In this case you will expect to see a positive peak next to a negative peak with the atom's current location between. This feature indicates that the atom should be moved toward the positive peak.

These are the signals that all crystallographers are taught to identify. The real situation is more complicated. There are peaks in Fo-Fc maps which do not indicate that atoms should be added to the model and

sometimes atoms have errors in their positions which are not marked by pairs of peaks. The proper interpretation of a Fo-Fc map requires that you be familiar with these limitations.

Fo-Fc Map Theory and Limitations

It was shown some time ago (Crukshank, 1951) that least squares minimization and flattening a Fo-Fc map are closely related tasks. This relationship is what allows us to use Fo-Fc map refitting side-by-side with least squares refinement. It was later shown that a relatively simple transformation can convert a Fo-Fc map to the gradient of the least squares' residual (Agarwal, 1978). In fact, this is the way many refinement packages calculate the gradient today. Agarwal's result allows us to treat the Fo-Fc map and the gradient of the least squares' residual function as equivalent.

Therefore, moving atoms to cause the Fo-Fc map to become flat is the same as moving the parameters of the model down the gradient vector. This describes the steepest descent method of function minimization (Leuenberger, 1971). While the steepest descent method is quite robust it is also quite limited.

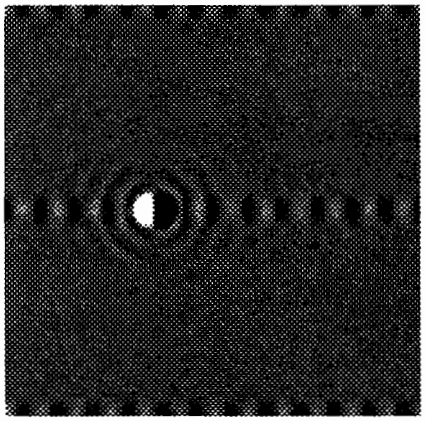
The principle omission from the steepest descent method is the lack of consideration of any second derivative information. The second derivative of the least squares expression contains several types of information about the model, including

- The precision or "significance" of each parameter, and
- The correlation and anticorrelation of pairs of parameters.

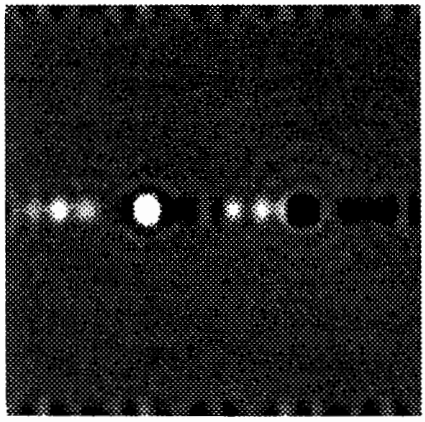
While the Fo-Fc map does not present any second derivative information all refinement packages incorporate some or all of it either directly or indirectly. XPLOR (Brünger, 1987) only includes the second derivative information indirectly via the conjugate gradient procedure (Fletcher and Reeves, 1964, Konnert, 1975). PROLSQ (Hendrickson and Konnert, 1980) uses the precision part (diagonal) of the second derivatives as well as some of the correlation part (off-diagonal) but uses this data ineffectively by using the conjugate gradient method of minimization in roughly the same fashion as XPLOR. TNT (Tronrud, et al, 1987) uses the precision part of the second derivatives with the preconditioned conjugate gradient method (Axelsson, 1985, Tronrud, 1990). While SHELXL (Sheldrick, and Schneider) can use all of the second derivative information the size of the computation required to determine the shift limits its use to small proteins.

The Effect of Parameter Correlation on the Fo-Fc Map

To demonstrate the effect of correlated errors in the parameters of a model I have constructed the following test case.



This is one section of a Fo-Fc map. Positive density is white and negative density is black. Regions with no difference density appear neutral gray. The length of each edge is 40Å. The full unit cell contains 10 atoms, all of which are in their correct position except for the atom in this section which is placed in error by 1.5Å. While the expected pair of peaks is quite evident there are a considerable number of other features in this section. Despite the complications the pair of peaks are sufficiently clear to indicate the error in the atom's position.



For comparison I have created another Fo-Fc map where I have simply added nine more atoms to the section, each of which are positioned in error by 1.5Å in the same direction. In this case there is not a pair of peaks for each atom but a single pair for the entire group of atoms. If you did not consider this group of atoms as a block you would be tempted to simply add a water molecule in the positive peak on the right and increase the B factor of the furthest atom on the left. Since some of the difference density (the three positive peaks on the far right) is fairly strong you might add water molecules there as well. These incorrect modifications of the model would lock the positions of these atoms in the wrong position. This map is very easy to misinterpret.

Since the refinement packages usually do not include second derivative information either they will not usually correct the error in this model either. When there is a concerted shift of a number of atoms you must specifically instruct the refinement package to look for such a shift. However, you will not be able to recognize the existence of this problem from looking at the map and if you perform automated refinement without precautions the computer will make inappropriate shifts and trap your model in error forever.

The lack of consideration of the second derivatives of most refinement packages results in the requirement that you perform rigid body re-

finement whenever it is possible that your model contains such errors. Usually a model constructed by reference to an m.i.r., s.i.r., or m.a.d. map will not contain errors of this type. However models generated by molecular replacement or molecular substitution (isomorphous mutant or inhibitor structures) often do. In these cases you must perform rigid body refinement with first each entire molecule in a group, then each domain in a rigid group, and perhaps finishing with significant portions of domains defined as rigid groups. Only then can you proceed to individual atom refinement.

You will not see clear indications in your Fo-Fc map that such errors are present even if they are present. To be safe you must perform the rigid body refinement in all cases.

Correlation of Parameters for a Single Atom

While the difference map signals mentioned above, a pair of peaks of opposite sign indicating a positional error and a peak centered on the atom indicating a B factor error, are the form generally taught they are rarely observed in refined difference maps. This is because there is a correlation between the position and B factor of each atom.

If a model is refined and, for some reason, an atom cannot move to accommodate the diffraction data the difference map will develop a pair of peaks. However the atom does not lie halfway between the two peaks – it will be a little closer to the negative peak. Since we have assumed that the atom cannot move to correct the error the only option available to the program is to raise the B factor to attempt to remove the negative peak. By the time the map is examined all that is left to see is a positive peak near an atom. The B factor may be unusually large but that may not be recognizable given the expected fluctuation of this type of parameter.

The most common difference density feature in a refined difference map is positive density near a atom. If there is any density at the position of the atom it is due to restraints preventing the B factor from changing. The response to this density is to search for the restraint which is preventing the atom from moving. If you simply move the atom manually whatever restraint caused the problem will pull the atom back to its original location.

The density of a difference map calculated with an unrefined model will exhibit the classical features.

Series Termination in Fo-Fc Maps

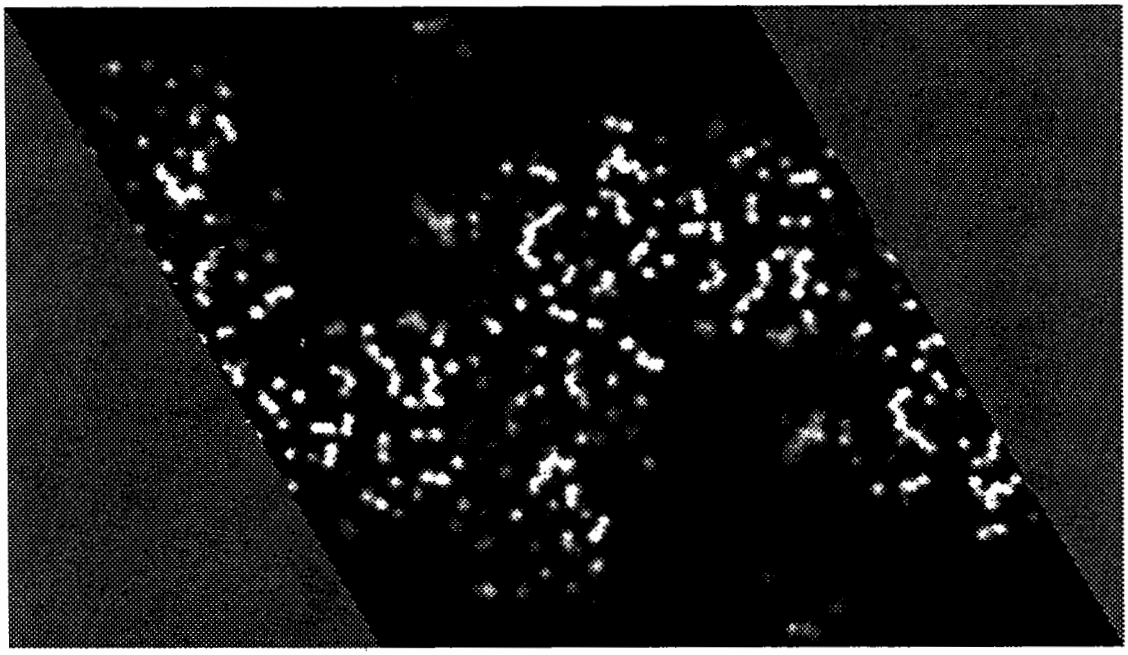
The maps above each contain two principle peaks which indicate the error in position of the group of atoms. Each map also contains a number of other peaks. These peaks are caused by series termination –

The lack of certain Fourier terms in the calculation of the maps. All density maps will contain a certain amount of series termination.

The principle cause of series termination is the incompleteness of the observed data set. While the incompleteness of a data set could have many forms usually it is described by an inner (or low) resolution and outer (or high) resolution limit. While the outer resolution limit usually exists because the crystal does not diffract with sufficient intensity to accurately measure (or the structure factors cannot be phased well enough) the inner limit is either chosen arbitrarily or imposed by the technical limitations of the data collection procedure (e.g. the beam stop). The significance of a resolution limit is determined by the amount of intensity lost from the calculation. If the outer resolution limit is caused by the weak diffraction of the crystal at that resolution this limit will not cause significant artifacts in the maps.

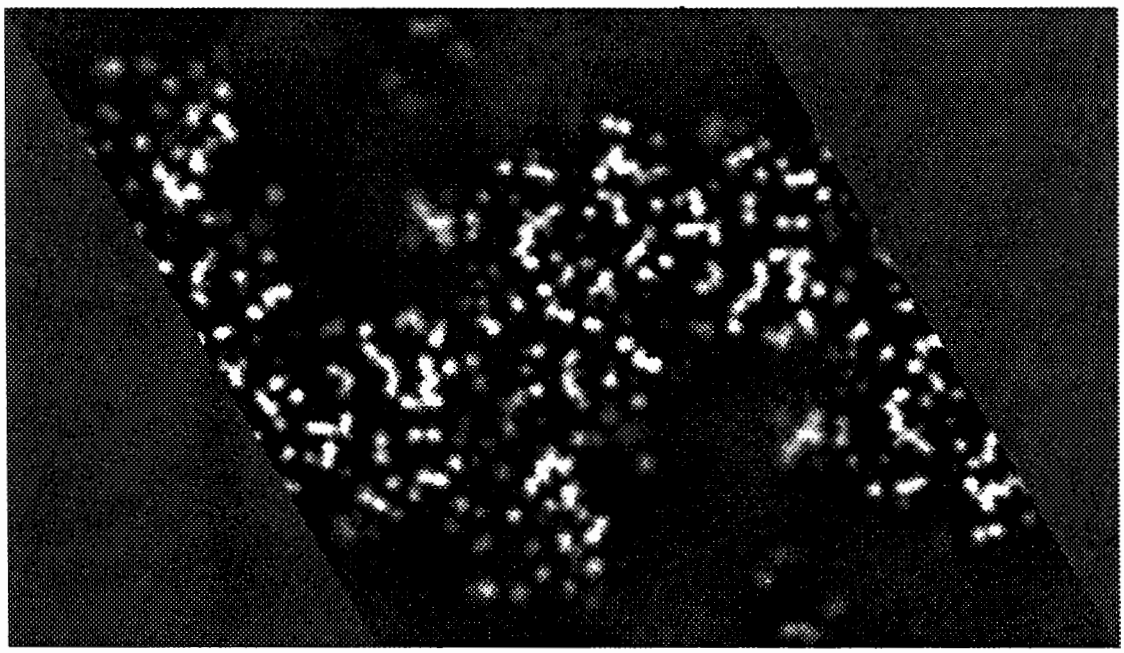
If the outer resolution limit is imposed because of phasing errors, as in a m.i.r. map with a breakdown of isomorphism at high resolution, there can be significant series termination errors. In addition the low resolution limit always excludes significant reflections and causes more errors. Since these limits are simple shapes in reciprocal space their effects are simple in real space as well. They cause every feature to be surrounded by ripples. The wavelength of the ripple will be somewhat beyond the resolution limit of the data. For example, a 3Å outer resolution limit will cause all features in a map to be surrounded by ripples with a wavelength somewhat shorter than 3Å. A 6Å inner resolution limit will cause ripples somewhat longer than 6Å.

To demonstrate the affect of series termination on the appearance of a 2Fo-Fc map I will show the results of some model calculations. The $z = 0$ section of a calculated electron density map for the protein Thermolysin (Holland, et al, 1992) is



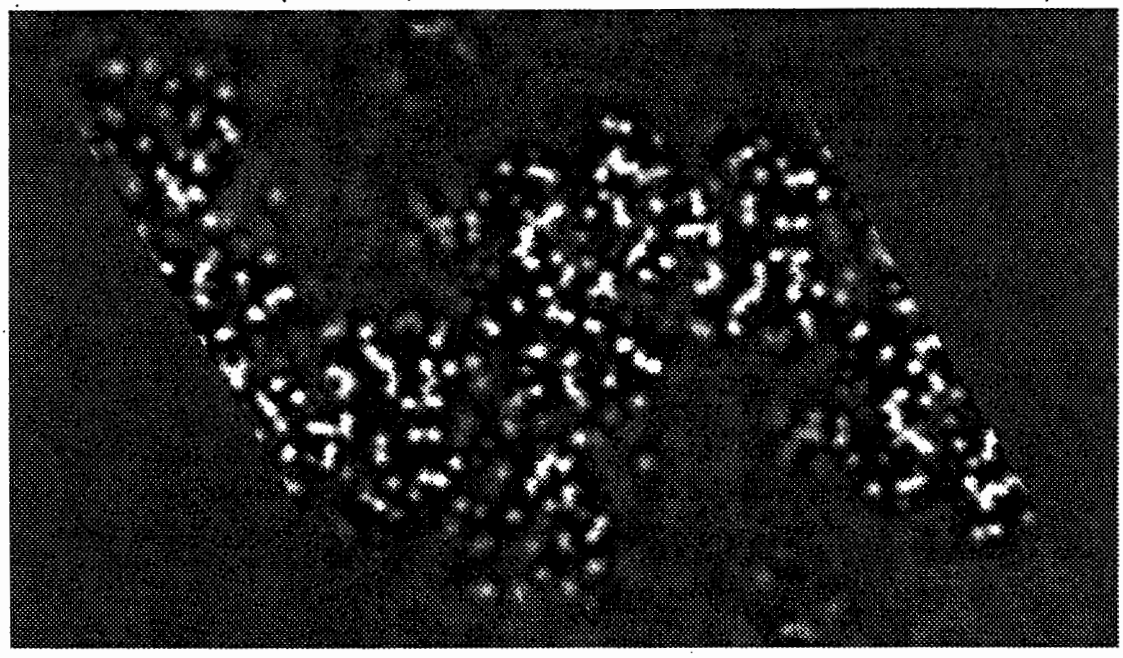
The crystal is hexagonal which explains the gray triangles on the maps sides. Since this map is simply calculated from the atomic positions it does not exhibit any defects due to resolution limits. The bulk solvent regions are devoid of density and the atoms are as resolved as well as can be expected for atoms with B factors of ~15.

A 2Fo-Fc map will never look this good. It will always be missing some of the low resolution data and most likely some of the high resolution data as well. If we recalculate the map shown above with the resolution limits 20 to 1.8Å the result is



You will note that while the solvent region now appears to contain density the principle features of the protein are still quite recognizable. This map could be used to build a model of the protein without much difficulty.

If the map is calculated again, this time with the resolution limits 6 to 1.8Å, the result is



In this map considerable density appears in the bulk solvent regions. While the core of the protein still exhibits sufficient detail to allow the positions of the atoms to be recognized the superposition of the false solvent density on the surface regions of the protein could cause regions with high B factors to be difficult to interpret. In addition there is a great temptation to interpret the "features" in the bulk solvent region as structured solvation.

One must be very cautious when interpreting weak density. There are many explanations for weak features in a map that do not involve the presence of ordered atoms.

Series Termination in Fo-Fc maps

The example shown above mimics a 2Fo-Fc map but series termination also affects Fo-Fc maps. Any error in the protein model will result in features in the Fo-Fc map. These features will be of the classical form – a pair of peaks of opposite sign for positional errors, a peak centered on the atom for a B factor error, and positive density for unmodeled protein – but will be modified by the series termination ripples.

When interpreting a Fo-Fc map you should only attempt to model the strongest features. The weaker features will be distorted by the ripples from the stronger and cannot be reliably interpreted. Once you have

corrected the major problems with your model you can calculate a new Fo-Fc map which will show a clearer image of the remaining problems.

Reducing the Parameter Uncertainty

The parameters of your final model will contain uncertainties. These uncertainties arise from the uncertainties in the measurement of the data and are modulated by the mathematical transformation required to calculate the model from those data. Since we do not know how to calculate the model from the data (we can only calculate what the data should be given a model.) the calculation of the uncertainties of our final parameters is quite difficult.

We do know the character of these uncertainties. While we usually talk about the uncertainty of a parameter by estimating a standard deviation, this list of "sigmas" does not tell the whole story. The more troublesome aspect of the uncertainty is the covariance.

The covariance of two parameters quantifies the extent that one parameter can change to compensate for a change in another. Whenever a pair of parameters have a large covariance their values have a much larger uncertainty than their individual standard deviations would indicate.

While it is quite difficult to calculate the covariance of every pair of parameters in a model there are steps which can be taken to reduce the uncertainty. The most powerful is to change the parameters of the model to another set which exhibit less correlation. Usually proteins are modeled by supplying a position and B factor for each atom. When the diffraction data only cover low resolution the parameters for neighboring atoms become highly correlated and their positions quite difficult to refine and their final values quite uncertain. If we knew the basic fold of the protein from some other source (say molecular replacement) we can redefine the parameters of the model. An example of this would be to define the parameters to be the position, orientation, and B factor of each domain in the protein and refine these parameters. Since the electron density of each domain does not overlap the covariance of these parameters will be much smaller.

This example is simply rigid body refinement and is a commonly used means of aiding refinement convergence. While these types of parameter changes are quite powerful current refinement packages are quite limited in their ability to allow parameterizations other than individual atoms and rigid groups.

Usually a new parameterization is devised to make use of some additional source of information. An analogy between the current structure and one solved in another space group provides the information used in the rigid body parameterization. The analogy from one crystal form to an-

other is usually only considered valid at low resolution and the rigid body model is abandoned when refining against high resolution diffraction data.

It would seem reasonable that an analogy between two very similar, isomorphous, structures would be valid to high resolution. If true one could redefine the parameters of the models to be more sensitive to the differences between the two structures. Terwilliger & Berendzen, (1995) have proposed a means of redefining the refinement process to emphasize the differences between the "derivative" and "native" structures (be they mutant verses wild type or inhibited verses uninhibited). While their approach appears promising it does not change the parameterization of either model. The next step would be to define a set of parameters which express the structural details of the two structures in a minimalist form.

Summary

The best source of information about the quality of your model is your maps. If a detail of the structure is not visible in the 2Fo-Fc map and a trial change in this feature of your model does not affect the Fo-Fc map then that detail is probably artifactual. You must be very careful, however, because these maps will contain features which do not arise from the true structure of the protein but are artifacts due to series termination, phase errors, incomplete data, and other sources. To achieve the best maps you must include all available data in their calculation (no omission of the low resolution data) and model all aspects of the structure, including the bulk solvent.

If you are interested in the fine details of your structure you will have to carefully choose the parameters of your model. You should not allow the model to violate facts about the structure such as the conformation of related structures. The parameters of the model should be contrived to allow variability in only those aspects which are believed to differ from known quantities. The fewer parameters the better.

Acknowledgments

This work was supported in part by NIH grant GM20066 to B. W. Matthews.

Bibliography

Agarwal, R.C., Acta Cryst **34A**(1978) 791-809

Axelsson, O., BIT, **25**(1985) 166-187

Brünger, A.T., Kuriyan, K., and Karplus, M., Science, **235**(1987) 458-460

Crukshank, D., Acta Cryst, **5** (1952) 511-518

Fletcher, R., and Reeves, C., Computer Journal, **7** (1964) 81-84

Hendrickson, W.A., and Konnert, J.H., in *Computing in Crystallography*, edited by Diamond, R., Ramasechan, S., and Venkatesan, K., 13.01-13.25 (1980), Bangalore: Indian Academy of Sciences

Holland, D.R., Tronrud, D.E., Pleyk, H.W., Flaherty, M., Stark, W., Jansonius, J.N., McKay, D.B., and Matthews, B.W., Biochemistry, **31**(1992) 11310-11316

Konnert, J.H., Acta Cryst, **32A**(1975) 614-617

Luenberger, D.G., *Introduction to Linear and Nonlinear Programming* (1973). Reading, MA. Addison-Wesley

Sheldrick, G.M., and Schneider, T.R. in *Methods of Enzymology*, edited by Sweet, B. and Carter, C., in preparation.

Terwilliger, T.C, and Berendzen, J., **51D** (1995) 609-618

Tronrud, D.E., Ten Eyck, L.F., and Matthews, B.W., Acta Cryst, **43A** (1987) 489-501

Tronrud, D.E., Acta Cryst, **48A**(1992) 912-916

PROTEIN PRECISION RE-EXAMINED: LUZZATI PLOTS DO NOT ESTIMATE FINAL ERRORS

D.W.J.CRUICKSHANK
Chemistry Department, UMIST, Manchester M60 1QD

INTRODUCTION

Almost 50 years ago E.G. Cox and G.A. Jeffrey started my interest in the accuracy of the structures of small molecules as determined by X-ray crystallography. Recently, for two reasons, I became interested in protein accuracy: first, the paper by Daopin *et al.* (1994) on the accuracy of two structures of TGF-β2 made generous remarks about error formulae of mine dating back to 1949; second, numerous protein papers are reporting the estimation of final errors by Luzzati (1952) plots. Unfortunately Luzzati developed his theory for a quite different purpose, and this use of Luzzati plots is ***wrong***. A critical discussion of Luzzati's theory will be offered later in this paper. However plots of R versus $2\sin\theta/\lambda$ can still be used to provide a statistical estimate of global uncertainty.

Even in 1967 when the first few protein structures had been solved, it would have been hard to imagine that a time would come when the best protein structures would be determined with a precision approaching that of small molecules. That time was reached some while ago. Consequently the methods for the assessment of the precision of small molecules can be extended to good quality protein structures.

The key proposition is simply stated. At the conclusion of a refinement, **the estimated variances and covariances of the parameters may be obtained through the inversion of the least-squares full matrix.** The inversion of the full matrix for a large protein is a gigantic computational task, but has been achieved for a few cases. Alternatively, approximations may be sought. Often these can be no more than rough order-of-magnitude estimates. Some of these approximations are considered below.

Caveat. Quite apart from their large number of atoms, protein structures show different features from well-ordered small molecule structures. Protein crystals contain large amounts of solvent, possibly not well ordered. Parts of the protein chain may be floppy or disordered. All protein crystals are non-centrosymmetric, hence the simplifications of error assessment for centrosymmetric structures are inapplicable. The effects of incomplete modelling of disorder on phase angles, and thus on parameter errors, are not addressed in the following analysis. Nor does this analysis address the quite different problem of possible gross errors or misplacements in a structure, other than by their indication through high B values or high coordinate e.s.d.'s.

EFFECT OF TEMPERATURE FACTORS

It is useful to begin with a reminder that the Debye $B = 8\pi^2\langle u^2\rangle$ where u is the atomic displacement amplitude. If B = 80 $Å^2$, the r.m.s. amplitude is 1.01 Å. The centroid of such an atom is unlikely to be precisely determined. For B = 40 $Å^2$, the 0.71 Å r.m.s. amplitude of an atom reaches the mid-point of a C-N bond. For B = 20 or 5 $Å^2$, $\langle u^2\rangle^{1/2}$ = 0.50 or 0.25 Å. Scattering power depends on $\exp[-2B(\sin\theta/\lambda)^2] = \exp[-B/(2d^2)]$. For B = 20 $Å^2$ and d = 4, 2 or 1 Å, this factor is 0.54, 0.08 or 0.0001. For d = 2 Å and B = 80, 20 or 5 $Å^2$, the factor is 0.0001, 0.08 or 0.54. The scattering power of an atom thus depends strongly on B and on the resolution $d = 1/s = \lambda/2\sin\theta$. Scattering at high resolution (low d) is dominated by atoms with low B.

Daopin, Davies, Schlunegger and Grutter (1994) compared their two independent determinations of transforming growth factor-β2 (112 amino acids). Each refinement of TGF-β2 used the TNT package. The 1TG1 structure had d_{min} = 1.8 Å and a final residual R = 0.173. The 1TFG structure had 1.95 Å and 0.188. The authors plotted the r.m.s. position differences $\langle\Delta r\rangle$ of the C_α atoms versus residue number, and showed that these structural differences were highly correlated with the Debye B factors. This provided another direct demonstration that atomic precision in proteins depends strongly on B. They then showed (Fig. 1) that the agreement between the observed r.m.s. structure differences $\langle\Delta r\rangle$ and the errors estimated by a formula of Cruickshank (1949a, 1959) was "quite good throughout the entire range of B values".

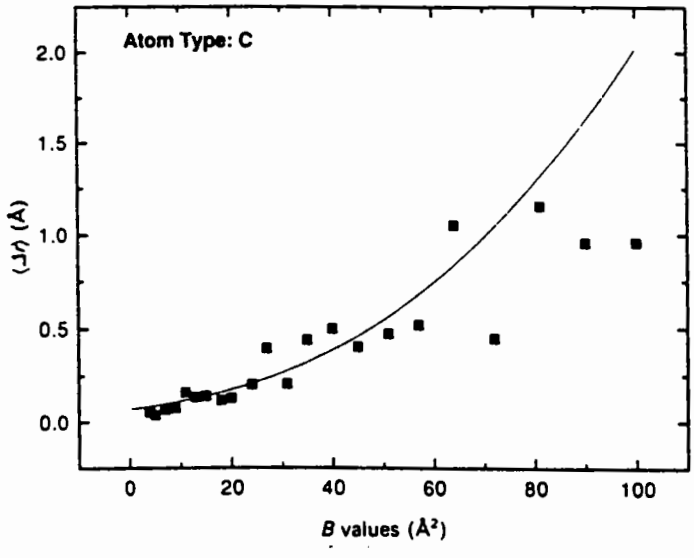


Fig.1. Daopin et al.'s comparison for all C atoms of the observed r.m.s. deviations between the two structures of TGF-β2 and the theoretical distribution curve calculated using (1).

This formula, based on a Fourier map approach, can be described approximately as

$$\sigma(x) = \sigma(\text{slope})/(\text{atomic peak "curvature"}). \qquad (1)$$

The σ(slope) term is the same for all atoms and is proportional to $[\Sigma_{obs} h^2 |\Delta F|^2]^{1/2}$. The "curvature" term, which depends on B and the atom type, is proportional to $\Sigma_{obs} h^2 f[\exp(-B\sin^2\theta/\lambda^2)][m/2]$, where m = 1 or 2 for acentric or centric reflections. Thus $\sigma(x)$ increases steadily with B as observed.

RESTRAINED REFINEMENT

Proteins are usually refined by a restrained refinement program such as PROLSQ. Here a function of type

$$R' = \Sigma w_h(\Delta F)^2 + \Sigma w_g(\Delta Q)^2 \qquad (2)$$

is minimised, where Q denotes a geometrical restraint, e.g. a bond length. Formally, all one is doing is extending the list of observations. One is adding to the protein diffraction data geometrical data from a stereochemical dictionary such as that of Engh and Huber (1991). A chain C-N bond length may be known from the dictionary with much greater precision $1/w_g^{1/2}$, say 0.02 Å, than from an unrestrained diffraction-data-only protein refinement.

In a high-resolution unrestrained refinement of a small molecule, the estimated standard deviation (e.s.d.) of a bond length A-B is often well approximated by $\sigma(l) = (\sigma_A^2 + \sigma_B^2)^{1/2}$. However in a protein determination $\sigma(l)$ is often much smaller than either σ_A or σ_B because of the excellent information available from the stereochemical dictionary which correlates the positions of A and B.

Laying aside computational size and complexity, the protein error problem is straightforward in principle. When a restrained refinement has converged to an acceptable structure and the shifts in successive rounds have become negligible, invert the full-matrix. The inverse matrix immediately yields estimates of the variances and covariances of all parameters.

A very simple protein model

Some aspects of restrained refinement are easily understood by **considering a one-dimensional protein consisting of two like atoms** in the asymmetric unit, with coordinates x_1 and x_2 and bond length $l = x_2 - x_1$. In the refinement the normal equations are of the type $\mathbf{N}\Delta\mathbf{x} = \mathbf{e}$. For two non-overlapping like atoms the ***diffraction data*** will yield a normal matrix

$$\begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix} \qquad \text{with inverse} \qquad \begin{bmatrix} 1/a & 0 \\ 0 & 1/a \end{bmatrix}$$

where $a = \Sigma w_h(\partial F_h/\partial x_i)^2$. A ***geometric restraint*** on the length will yield a normal matrix

$$\begin{bmatrix} b & -b \\ -b & b \end{bmatrix}$$ with no inverse since determinant = 0

where $b = w_g(\partial l/\partial x_i)^2$. Note $\partial l/\partial x_2 = -\partial l/\partial x_1 = 1$, and $b = w_g = 1/\sigma_g^2$, where σ_g^2 is the variance assigned to the length in the stereochemical dictionary.

Combining the diffraction data and the restraint the normal matrix becomes

$$\begin{bmatrix} a+b & -b \\ -b & a+b \end{bmatrix}$$ with inverse $[1/a(a+2b)]\begin{bmatrix} a+b & b \\ b & a+b \end{bmatrix}$

For the diffraction data alone the variance of x_i is

$$\sigma_d^2(x_i) = 1/a.$$

For the diffraction data plus restraint the variance of x_i is

$$\sigma_r^2(x_i) = (a+b)/[a(a+2b)] \quad < \sigma_d^2(x_i). \qquad (3)$$

Note that though the restraint says nothing about the position of x_i, the variance of x_i has been reduced because of the coupling to the position of the other atom.

Analysis shows that **the bond length l determined in the restrained refinement is the weighted mean of the diffraction-data only length and the geometric-dictionary length.**

For the diffraction data alone it can be shown

$$\sigma_d^2(l) = 2/a = 2\sigma_d^2(x_i), \text{ as expected.}$$

For the diffraction data plus restraint

$$\sigma_r^2(l) = 1/(a/2 + b) \quad < \quad \sigma_d^2(l). \qquad (4)$$

The centroid has coordinate $c = (x_1 + x_2)/2$. It is easily found that $\sigma_r^2(c) = \sigma_d^2(c) = 1/2a$. Thus, as expected, the restraint says nothing about the centroid of the molecule.

STATISTICAL EXPECTATION OF ERROR DEPENDENCE

From general statistical theory, one would expect the e.s.d. of an atomic coordinate determined from the diffraction data alone to show dependence on four factors:

$$\sigma(x) \propto (\text{"R"})\ [(n_{atoms})/(n_{obs} - n_{params})]^{1/2}\ (1/s_{rms}). \qquad (5)$$

"R" is some measure of the precision of the data, n_{obs} is the number of independent data, but to achieve the correct number

of degrees of freedom this must be reduced by the number of parameters determined, n_{atoms} is the recognition the information content of the data has to be shared out, and $1/s_{rms}$ is a more specialised factor arising from the sensitivity $\partial|F|/\partial x$ of the data to the parameter x. Any error estimate must show show some correspondence to these four factors.

Cruickshank (1960) published, based on a least-squares approach, a simple order-of-magnitude formula for $\sigma(x)$ in small molecules. It was intended for use in experimental design: how many data of what precision are needed to achieve a given precision in the results. The formula, derived from a very rough estimate of a least-squares diagonal element in non-centrosymmetric space groups, was

$$\sigma(x) = (1/2)\ [N/p]^{½}\ [R/s_{rms}]. \qquad (6)$$

Here $p = n_{obs} - n_{params}$, R is the usual residual $\Sigma|\Delta F|/\Sigma|F|$, and N is the number of atoms, similar to atom i whose $\sigma(x)$ is in question, needed to give scattering power at s_{rms} equal to that of the asymmetric unit of the structure, i.e. $\Sigma_j f_j^2 \equiv N f_i^2$.

For small molecules, this definition of N allowed the treatment of different types of atom with not-too-different B's. However it is not suitable for individual atoms in proteins where there is a very large range of B values and some atoms have B's so large as to possess negligible scattering power at s_{rms}.

Often, as in isotropic refinement, $n_{params} = 4n_{atoms}$, where n_{atoms} is the total number of atoms in the assymetric unit.

A very rough extension of (6) for application in proteins to an atom with $B = B_i$ is

$$\sigma(x_i) = k\ (N_i/p)^{½}\ [g(B_i)/g(B_W)]\ C^{-1/3}\ d_{min}\ R, \qquad (7)$$

where k is about 1.0, $N_i = \Sigma Z_j^2/Z_i^2$, B_W is the Wilson B for the structure, and C is the fractional completeness of the data to d_{min}. In deriving (7) from (6) $1/s_{rms}$ has been replaced by $1.3d_{min}$, and the factor (1/2) x 1.3 = 0.65 has been increased to 1.0 as a measure of caution in the replacement of a full-matrix by a diagonal approximation. $g(B) = 1 + a_1B + a_2B^2$ is an empirical function to allow for the dependence of $\sigma(x)$ on B. Based only on the data of Daopin *et al.* (1994), a provisional estimate for g(B) is

$$g(B) = 1 + 0.04B + 0.003B^2. \qquad (8)$$

A useful comparison of the relative precision of different structures may be obtained by comparing atoms with $B = B_W$ in the different structures. (7) then reduces to

$$\sigma_W(x) = 1.0\ (N_i/p)^{½}\ C^{-1/3}\ d_{min}\ R. \qquad (9)$$

The smaller d_{min} and the smaller R the better the precision of the structure. (9) is not to be regarded as having absolute validity. It is a quick and rough guide for the diffraction-data-only error contribution for an atom with temperature factor equal to the Wilson B for the structure. We shall call it the **Diffraction-data Precision Indicator**, or DPI. It contains none of the restraint data.

TABLE 1. DIFFRACTION-DATA PRECISION INDICATOR $\sigma_W(x)$

$$\sigma_W(x) = 1.0\ (N_i/p)^{½}\ C^{-1/3}\ d_{min}\ R$$

Protein	n_{atoms}	n_{obs}	$[N/p]^{½}$	d_{min}	R	σ_W	σ_{Luzz}
Crambin	504	23759	0.160	0.83	0.090	0.013	0.03
Rubredoxin	479	18532	0.170	1.0	0.160	0.029	0.08
TGF-β2							
1TGI	948	14000	0.305	1.8	0.173	0.094	0.13
1TFG	974	11000	0.370	1.95	0.188	0.14	0.14
Plastocyanin							
295 K	849	14303	0.279	1.33	0.149	0.061	0.09
173 K Sydney	928	7393	0.502	1.6	0.132	0.12	0.08
173 K Hamburg	910	7393	0.492	1.6	0.153	0.14	-
S protease D	4295	23249	0.841	2.0	0.188	0.32	0.14
Thaumatin C2	1552	(4622 +5274)	(0.649)	2.6	0.184	(0.34)	0.16

EXAMPLES OF DIFFRACTION-DATA PRECISION INDICATOR

Table 1 shows some details of the application of the Diffraction-data Precision Indicator (9) to proteins of differing precision, starting with the best. In all the examples, N_i has been set equal to $N = n_{atoms}$, the total number of atoms. The first entry is for crambin at 0.83 Å resolution and 130 K (Stec, Zhou and Teeter, 1995). Their results were obtained from an **unrestrained full-matrix anisotropic refinement.** Their inversion of the full-matrix gave e.s.d's 0.0096 A for backbone atoms, 0.0168 Å for side chain atoms and 0.0409 Å for solvent atoms, with an average for all atoms of 0.022 Å. The DPI gives $\sigma_W(x) = 0.013$ Å, which is satisfactorily between the full-matrix values for the backbone and side chain atoms.

The next entry in Table 1 is for rubredoxin at 1.0 Å (Dauter, Sieker and Wilson, 1992). They carried out both unrestrained and restrained isotropic refinements. Details

are given for the unrestrained refinement. They did not make formal calculations of e.s.d's, but from the deviations of the bond lengths from the dictionary values, they suggested the r.m.s. errors in the coordinates of the well-ordered atoms were about 0.04 Å. The DPI gives $\sigma_W(x)$ = 0.029 Å for B_W = 5.9 Å^2; the mean main-chain B was 8.0 Å^2.

The next entries concern the two lower resolution studies of TGF-β2 (Daopin *et al.*, 1994). The DPI gives $\sigma_W(x)$ = 0.094 Å for 1TGI and 0.14 Å for 1TFG. This indicates an r.m.s. position difference between the structures for atoms with B = B_W of $(0.094^2 + 0.14^2)^{½}/3$ = 0.29 Å. Daopin *et al.* reported the differences between the two determinations, omitting poor parts, as $\langle\Delta r\rangle_{rms}$ = 0.15 (main chain) and 0.29 Å (all protein).

The next entries concern poplar plastocyanin at 295 K (Guss *et al.*, 1992) and 173 K (Fields *et al.*, 1994). For 173 K a single set of Hamburg synchrotron data was refined quite independently in Sydney and Hamburg. The r.m.s. difference in position between the two models was 0.12 Å (excluding six outliers). The DPI expectation is $(0.12^2 + 0.14^2)^{½}/3$ = 0.32 Å, which is much larger. This is not surprising, since these were two refinements of the same data. The higher resolution room temperature study was more precise.

Serine protease factor D (Carson *et al.*, 1994) is an example of a large protein at lower resolution with a high value of $[N/p]^{½}$, leading to σ_W = 0.32 Å.

Three crystal forms of thaumatin were studied by Ko *et al.* (1994). The orthorhombic and tetragonal forms diffracted to 1.75 Å, but the monoclinic C2 form diffracted only to 2.6 Å with 4622 reflections. The structures with 1552 protein atoms were successfully refined with XPLOR and TNT. For the monoclinic form the number of parameters exceeds the number of diffraction observations, so [N/p] is negative and no estimate of the diffraction-only error is possible. However the refinement introduced 5274 geometrical restraints, and if, improperly, these are regarded as additional diffraction observations, we may derive $\sigma_W(x)$ = 0.34 Å. Perhaps such a calculation should be called simply a Precision Indicator. The usual DPI formula (9) gives 0.17 and 0.16 Å for the orthorhomic and tetragonal forms.

EFFECT OF RESTRAINTS

At the end of a restrained refinement the proper estimates of precision should be based on the normal matrix including the restraint contributions.

Some aspects of restrained refinement were illustrated in the simplest possible two-atom model discussed above. In a protein the restraints are applied between the atoms in each peptide and in each side-chain group. Thus there are numerous

significant off-diagonal terms in the full matrix. A simple algebraic approximation to the inversion does not seem possible. Approximation (7), based only on the diffraction contribution to the matrix diagonal, overestimates true **coordinate** errors. Final errors in **bond lengths** will be often more nearly represented by the original geometric weights.

Geometric dictionaries typically use bond length weights based on $\sigma_g(1)$ of around 0.02 or 0.03 Å. Table 1 shows that even 1.5 Å resolution studies have diffraction-only errors $\sigma_W(x)$ of 0.08 Å and upwards. Only for resolutions of 1.0 Å or so are the diffraction-only errors comparable with the dictionary weights. Of course, the dictionary offers no values for many geometric/configurational parameters of the protein structure, including the centroid and orientation.

If the protein main chain were represented by a chain of **rigid** peptide groups, 12 new coordinate parameters per successive group would be reduced to 2 torsion angles per group. Off-diagonal terms between these variables would be relatively less significant. Even if successsive peptide groups were treated as not coupled at the C_α atoms, each group could be specified by 3 coordinates and 3 orientation angles. Thus one may suspect in 1.5 Å and lower resolution restrained refinements that the true atomic $\sigma(x)$ in such groups may be between $(2/12)^{\frac{1}{2}}$ and $(6/12)^{\frac{1}{2}}$ times the values given by (7). Similar arguments apply to side-chain groups.

Monoclinic C2 thaumatin with 2.6 Å data illustrated a low-resolution problem. There are fewer observations than parameters, so no diffraction-only error estimate is possible. But thanks to the extensive geometric dictionary, coupled with the diffraction observations, the structure has been validly determined. The diffraction data will have added practically nothing to the dictionary knowledge of dictionary parameters, but the true atomic $\sigma_W(x)$ may be better than the 0.34 Å calculated by the, improper, Precision Indicator.

CRITIQUE OF LUZZATI PLOTS

Luzzati (1952) did not provide a theory for estimating errors at the end of a refinement. He provided a means for estimating how far the refinement still had to go to reach R = 0.

(1) His theory assumed that the F_{obs} had no errors, and that the F_{calc} model (scattering factors, thermal parameters, etc.) was perfect apart from coordinate errors.

(2) The Gaussian probability distribution for these coordinate errors was assumed to be the ***same for all atoms*** , independent of B or Z.

(3) The atoms were not required to be identical, and the position errors were not required to be small.

Luzzati gave families of curves for R vs. $\sin\theta/\lambda$ for varying $\langle\Delta r\rangle$ for both centrosymmetric and non-centrosymmetric

structures. The curves do not depend on the number N of atoms in the cell. They all rise from R = 0 at $\sin\theta/\lambda = 0$ to the Wilson values 0.828 and 0.586 for random structures at high $\sin\theta/\lambda$. In a footnote (p.807) Luzzati suggested that at the end of a refinement (with R non-zero due to experimental and model errors, etc.) the curves would indicate an upper limit for $\langle\Delta r\rangle$. As examples, the Luzzati plots for TGF-β2 are shown in Fig. 2. They suggest average Δr around 0.21 and 0.23 Å.

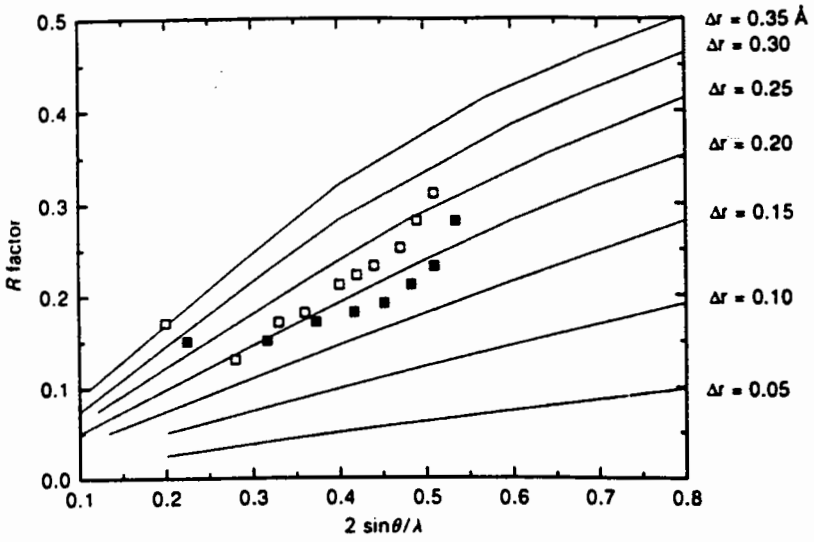


Fig. 2. Luzzati plots for TGF-β2 showing the refined R factor as a function of $2\sin\theta/\lambda$ for 1TGI (solid squares) and 1TFG (open squares), from Daopin *et al.* (1994).

For proteins, there are obvious difficulties with Luzzati's assumption (2). Errors do increase markedly with B. In the high-angle data shells, atoms with large B's contribute neither to ΔF nor to $|F|$ and so have no effect on R in these shells. In an important paper on protein accuracy, Chambers and Stroud (1979) said "the [Luzzati] estimate derived from reflections in this range applies mainly to [the] best determined atoms." Thus it has seemed that a Luzzati plot provides some sort of cautious statement about the good parts of a structure, but that it gives no indication for the poor parts except possibly in the lowest resolution shell.

Unfortunately there is a more fundamental objection to the use of Luzzati plots. This is that the Luzzati theory applies to **incomplete** refinements and estimates the r.m.s. shifts needed to reach R = 0. In the least-squares method **the equations for shifts are quite different from the equations to estimate variances** in completed refinements.

However Luzzati-style plots of R versus $\sin\theta/\lambda$ **can be re-interpreted** to give statistically based estimates of $\sigma(x)$.

During Cruickshank's 1960 derivation of the approximate estimate (6) for $\sigma^2(x)$ in diagonal least squares, he reached an intermediate equation

$$\sigma^2(x) = N / (4 \Sigma_{obs}[s^2/R^2]. \qquad (10)$$

He then assumed R to be independent of s, and took R outside the summation to reach (6) above.

Luzzati (1952) calculated the acentric residual R as a function of s and $\langle\Delta r\rangle$, the average radial error of the atomic positions. His Appendix 3 shows that $\langle\Delta r\rangle = (8/\pi)^{1/2}\sigma(x) = 1.60\ \sigma(x)$, where $\sigma(x)$ is the r.m.s. error in the x direction. We may call this quantity $\sigma_{Luzz}(x)$. Luzzati's main analysis shows that R is a linear function of s and $\sigma_{Luzz}(x)$ for a substantial range of $s.\sigma$, with

$$R(s,\sigma) = 4\ s\ \sigma_{Luzz}(x). \qquad (11)$$

The theoretical Luzzati plots of R are nearly linear for small to medium $s = 2\sin\theta/\lambda$ (see Fig. 2). If we substitute this R in the least-squares estimate (10), a little manipulation , including the earlier cautionary factor 1/0.65, then gives

$$\sigma_{LS,Luzz}(x) = 3.0\ [N/p]^{1/2}\ \sigma_{Luzz}(x), \qquad (12a)$$

or,
$$\sigma_{LS,Luzz}(x) = 0.75\ [N/p]^{1/2}\ [R(s_m)/s_m], \qquad (12b)$$

where $R(s_m)$ is the value of R at some chosen, presumably high, value of $s = s_m$.

Subject to the stated approximations, (12a,b) is a proper statistical estimate of the diffraction-data-only $\sigma(x)$ when the final residual behaves as a function of s in the manner considered by Luzzati. As expected statistically, the numbers of observations and parameters and the proportionate scattering power of a single atom enter into the result. These terms are absent in Luzzati's estimate of $\sigma_{Luzz}(x)$ from R(s).

Authors using Luzzati plots, as in Fig. 2, usually follow Luzzati and give values of $\langle\Delta r\rangle$, the average positional error. For the examples of Table 1 the reported values of $\langle\Delta r\rangle$ have been converted to r.m.s. coordinate errors $\sigma_{Luzz}(x) = \langle\Delta r\rangle/1.60$. From (12a) we see that $\sigma_{Luzz}(x) = \sigma_{LS,Luzz}(x)$ when $[N/p]^{1/2} = 1/3.0 = 0.33$. Here $\sigma_{LS,Luzz}(x)$ is the correct statistical estimate for the atom type implied in N when R behaves as a function of $s = 2\sin\theta/\lambda$ in the manner envisaged by Luzzati.

We can now see why reported Luzzati values of $\langle\Delta r\rangle$ were often plausible. For many proteins $[N/p]^{1/2}$ is around 0.4, and σ_{Luzz} is then comparable with σ_W. For large $[N/p]^{1/2}$ Luzzati **underestimates** the error. Conversely for atomic resolution studies, with $[N/p]^{1/2} < 0.2$, the Luzzati plot **overestimates** the

errors (as Luzzati had supposed in 1952).

Actually the equality of σ_{Luzz} and σ_W at 0.14 Å in Table 1 for 1TFG is a little lucky. It required that

$$d_{min}\ R = 0.75\ [R(s_{max})/s_{max}], \qquad (13)$$

i.e. $R = 0.75\ R(s_{max})$. While one expects $R(s_{max})$ to be about this much larger than a conventional averaged R, the actual value of $R(s_{max})$ depended on where the original authors placed their Luzzati line among the scattered experimental points.

Equation (12b) provides a means of making a statistical estimate of error for an atom with $B = B_W$ from a plot of R versus $2\sin\theta/\lambda$. If the original published Luzzati values of $\langle\Delta r\rangle = 1.60\ \sigma_{Luzz}$ are substituted in (12a), the resulting $\sigma_{LS,Luzz}(x)$ are within 20% of the $\sigma_W(x)$ in Table 1 for all the proteins.

CONCLUSION

The previous use of Luzzati plots to estimate final coordinate errors has been shown to be invalid. Paradoxically, (12a,b) allows the Luzzati parameter to be converted to a valid statistical form. Indeed the crystallographer has a choice of using R as a constant in (9) or $R(s_{max})$ as a single point from a linear plot as in (12b) ! Examination of Fig. 2 suggests the assumption R(s) proportional to s as in (12b) could sometimes be a better approximation than assuming R to be constant as in (9). If so, the protein crystallographer should continue to plot R as a function of $2\sin\theta/\lambda$, but should interpret the results by (12b) rather than by Luzzati's Table.

It must be stressed that the error estimate (7), the Diffraction-data Precision Indicator (9), and (12b) are only very rough formulae for the diffraction-data contribution to coordinate precision. None the less Table 1 shows that the DPI gives useful order-of-magnitude results for the global average precision of structures.

The question of local precision is obviously of great importance. Equation (7) offers the possibility of a simple formula for diffraction-data error estimates for atoms of given B and Z. It has a plausible dependence on B and Z but has not been tested yet.

An attractive (Fig. 1) and less approximate estimate for the diffraction-data error of an individual atom is offered by Daopin et al's (1994) use of equation (1). This formula (Cruickshank, 1949a, 1959) requires the summation of various series over all (h,k,l) observations; such calculations are not customarily provided in protein programmes. However due to the fundamental similarities between Fourier and least-squares methods demonstrated by Cochran (1948),

Cruickshank (1949b) and Cruickshank and Robertson (1953), similar estimates of the errors of individual atoms can be obtained from the reciprocals of the diagonal elements of the diffraction-data least-squares matrix. These elements will often already have been calculated within the protein refinement programs, but possibly never output. Such estimates could be routinely available. They will usually be overestimates of the true errors. Correspondingly, the reciprocals of the diagonal elements of the diffraction-cum-restraints matrix may be underestimates of the true coordinate errors.

Perhaps more efforts should be made to approach the proper method, which involves the inverse of the diffraction-cum-restraints full matrix. As far back as 1973 Watenpaugh *et al.* in a study of a rubredoxin at 1.5 Å resolution effectively inverted the diffraction full-matrix in 200 parameter blocks to obtain individual e.s.d's. A comparable scheme might be to calculate blocks for each residue, and for the block interactions between successive residues. Then invert the matrices in running groups of three successive residues, using only the inverted elements for the central residue as the estimates of its variances and covariances.

References

Carson, M., Bugg, C.E., DeLucas, L.J. and Narayana, S.V.L. Acta Cryst, D50 (1994) 889.

Chambers, J.L. and Stroud, R.M. Acta Cryst, B35 (1979) 1861.

Cochran, W. Acta Cryst, 1 (1948) 138.

Cruickshank, D.W.J. Acta Cryst, 2 (1949a) 65.

Cruickshank, D.W.J. Acta Cryst, 2 (1949b) 154.

Cruickshank, D.W.J. and Robertson, A.P. Acta Cryst. 6 (1953) 698.

Cruickshank, D.W.J. Int. Tables X-ray Cryst, 2 (1959) 318.

Cruickshank, D.W.J. Acta Cryst, 13 (1960) 774.

Daopin, S., Davies, D.R., Schlunegger, M.P. and Grutter, M.G. Acta Cryst, D50 (1994) 85.

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

Engh, R.A. and Huber, R. Acta Cryst, A47 (1991) 392.

Fields, B.A., Bartsch, H.H., Bartunik, H.D., Cordes,F., Guss, J.M. and Freeman, H.C. Acta Cryst, D50 (1994) 709.

Guss, J.M., Bartunik, H.D. and Freeman, H.C. Acta Cryst, B48 (1992) 790.

Ko,T-P., Day, J., Greenwood, A. and McPherson, A. Acta Cryst, D50 (1994) 813.

Luzzati, V. Acta Cryst, 5 (1952) 802.

Stec, B., Zhou, R. and Teeter, M.M. Acta Cryst, D51 (1995) 663.

Watenpaugh, K.D., Sieker, L.C., Heriott, J.R. and Jensen, L.H. Acta Cryst, B29 (1973) 943.

Review: Cross-Validation and the Free-R

Dr. Kevin Cowtan

Protein Structure Group, Department of Chemistry,
University of York, Heslington, York, YO1 5DD. England

Abstract

Macromolecular crystallography is an example of a scientific measurement problem which starts with a diffraction pattern and ends with an atomic model. The path in between is sufficiently complex that it is hard to relate the precision of the final model to the original measurements. As a result it is hard to distinguish between an improvement to the model which fits a genuine feature of the signal, and one which merely fits noise in the data.

A statistical technique, Cross Validation, is available to counter this problem. Its crystallographic application has been pioneered by Axel Brünger in the form of the Free-R factor (Brünger, 1992).

1 Introduction

In fitting an atomic model to a set of experimental structure factors we refine the model parameters in order to best reproduce the measured structure factors (in the least-squares approximation). However, in the case of macromolecules, the refinement is frequently poorly determined, since it is common to work at resolutions where the number of atomic parameters exceeds the number of measured intensities. Thus additional parameters (for example - individual atomic B's, multiple conformations, solvent atoms) must be introduced with extreme care, since it is quite possible that new parameters introduced to improve the fit of the model to the data, will simply fit the noise in the data rather than revealing any genuine features of the structure.

If the purpose of the model were simple to represent the electron density in a compact form, then there would be no problem. However, structures are usually solved with a view to learning about biological function. If erroneous features are introduced into a model in order to fit noisy data, then it is entirely possible that those erroneous features will lead to incorrect conclusions about function.

The traditional indicator of structure quality is the R-factor:

$$R = \frac{\sum_{allh} ||F_{obs}| - |F_{calc}||}{\sum_{allh} |F_{obs}|} \quad (1)$$

This quantity measures the agreement between the observed structure factor magnitudes and the values predicted from the model. It is therefore a measure of how well the model fits the data. Unfortunately, an arbitrarily good fit, and therefore an arbitrarily low R-factor, may be obtained by introducing enough parameters into the model - although in practice we generally restrict ourselves to introducing parameters with some plausible chemical interpretation.

Thus there is a problem in deciding which model parameters may be adequately determined from a noisy data set. In the case were the parameter-to-observable ratio is low, the introduction of excess parameters may lead to a functionally misleading model.

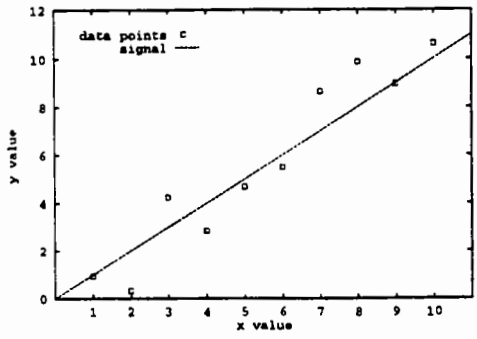


Figure 1: True signal and Obs data

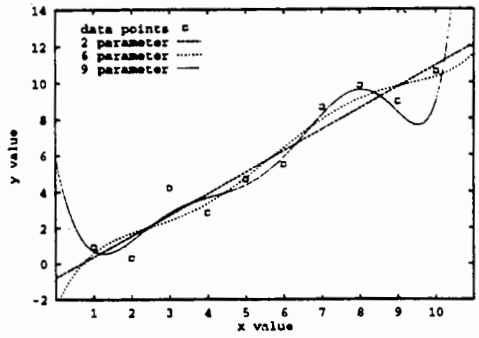


Figure 2: Polynomial fits to data

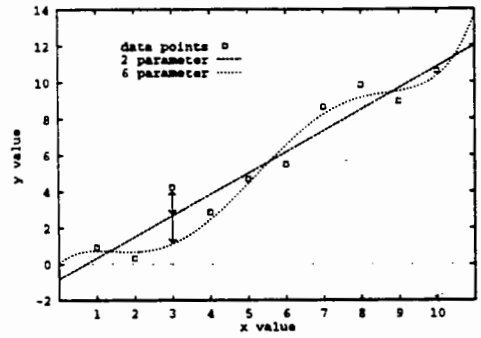


Figure 3: Polynomial with missing pt.

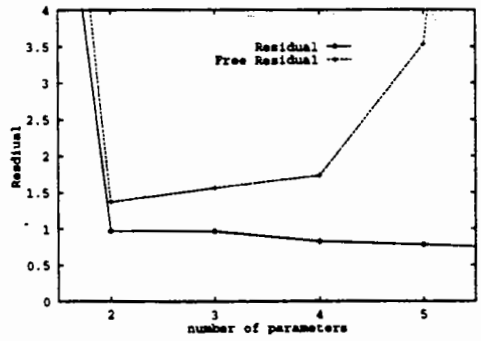


Figure 4: Residual vs no. of params.

2 Cross Validation

Fortunately, a statistical technique is available to provide precisely the discrimination between modelling signal and modelling noise which we require. This technique is Cross Validation, and will be illustrated by a simple example (inspired by Brünger, 1995).

An experiment is used to determine the dependence of a single variable on another variable. The variables are linearly dependent, however the measurement process is noisy and so the observations have some scatter about the straight line (Figure 1).

The observer does not know the form of the functional dependence, and so tries to fit a variety of polynomials of different orders. A constant is a poor fit to the data, however (as is expected), a straight line fits quite well (Figure 2). Higher order curves pass closer to the data points, as the curve begins to model the noise in the data. The R-factor does not help in identifying the best model, since each higher order polynomial gives a better fit to the data and so a lower R-factor.

The cross validation approach is to omit one of the data points from the fitting of the polynomial coefficients. In Figure 3, the third datapoint from the left has been omitted from the fitting. It can be seen from the graph that while the higher order polynomials make a better fit to the included points, the fit to the omitted point is *worse*.

What is happening here? The key point is that the signal is common across all the data, whereas the noise is independent for each data point. If increasing the number of parameters gives a better fit to the signal part of the data, then we would expect the fit to both included and excluded points to improve. If however new parameters provide are simply fitting the noise, then only the fit to

the included data will improve.

To incorporate this concept into a statistic, it is necessary to combine information from more than one data point. The normal approach in this sort of case would be to omit each point in turn, refine a function of the desired form against the remaining points, and calculate a cross-validated, or 'Free' residual over the points omitted from each calculation.

The conventional and free residuals for the test case are shown in Figure 4. It can be seen that the residual always decreases as the number of model parameters is increased, although the drop in the residual is much smaller when more than two parameters are used. The free residual however has a minima when the model has two parameters, correctly indicating the best model.

3 Crystallographic Application

In a conventional crystallographic refinement, the model is refined against all the data, and then a test (the R-factor) is performed against the same data. Any increase in the number of model parameters must therefore lead to a reduction in R-factor (Figure 5).

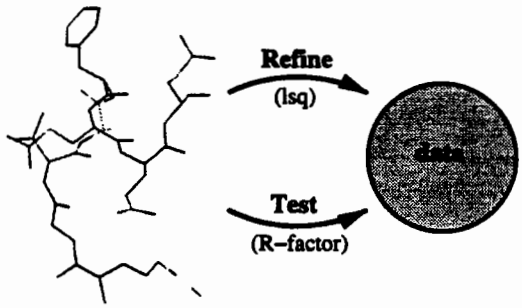


Figure 5: Conventional refinement procedure

In the crystallographic case, the number of observations is very large. It is therefore usually possible to simply omit a small fraction of the data which includes enough reflections to provide statistically meaningful information. Thus only one refinement calculation is required, during which one subset of the data is omitted (Figure 6).

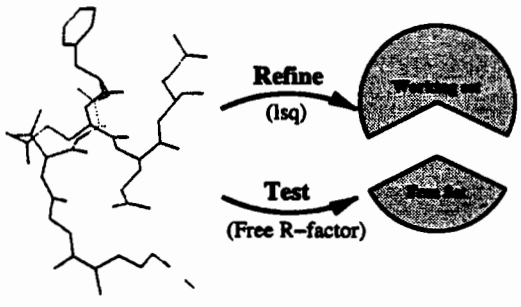


Figure 6: Cross-validated refinement procedure

This omitted, or Free set need only contain about 1000 reflections to provide a reasonable statistical indicator (Dodson, Klegwegt & Wilson, 1996) - typically in the protein case this will be 5% of the data or less. The Free-R is defined in the same way as the R-factor, but is calculated

over the reflections in the Free set alone:

$$R = \frac{\sum_{h \in \text{free-set}} ||F_{obs}| - |F_{calc}||}{\sum_{h \in \text{free-set}} |F_{obs}|} \quad (2)$$

Typically the Free-R factor for a refined structure will be higher than the R-factor, since it is impossible to completely separate parameters which fit signal and noise. Brünger suggests upper bounds beyond which a structure is almost certainly wrong of $R = 25\%$ and $R_{\text{free}} = 40\%$. These figures would be expected from a structure with mean coordinate error > 1Å.

How many parameters may be realistically refined? Dodson, Klegwegt & Wilson (1996) give some guidelines, beyond which any additional parameters should be justified by cross-validation:

1. At worse than 2.8Å resolution, there are less observed intensities than coordinates (assuming 50% solvent), and so coordinate refinement is unrealistic. This limit may be relaxed by introducing NCS constraints, torsion angle refinement, or refinement against phases; otherwise the model should be refined as a series of rigid domains only.
2. Between 2.8Å and 2.5Å it is possible to perform coordinate refinement, with *up to* 2 B-factors per residue. It is probably wise to start with a single overall B-factor.
3. At greater 2.5Å it is possible to refine individual atomic B's.

Whenever new parameters are introduced to the model and refined, the R-factor and Free-R should be examined. If neither drops significantly, the the parameters are useless in improving the fit to the data and should be removed. If the R-factor drops but the Free-R does not, then over-fitting is occurring. The additional parameters are fitting noise rather than signal and should be removed. Only if both the R-factor and Free-R drop should the additional parameters be regarded as meaningful. In marginal cases where the Free-R falls far less than the R-factor, it may be worth seeking an alternative parameterisation, for example a bulk solvent model instead of individual solvent molecules.

In the CCP4 suite the recommended procedure for using the Free-R factor is show in Figure 7. The Free-R flags are assigned immediately after data reduction stage (program '**freerflag**'), so that the same free-set may be used throughout the structure solution process. It is recommended that the program '**unique**' is used to create empty records for any missing reflections in the file. This avoids the problem of calculating Free-R flags for new reflections if they are collected at a later stage, and in the future will allow the extrapolation of those reflections when appropriate software becomes available. The programs '**rstats**' and '**sfall**' both calculate the Free-R factor.

4 Problems

The Free-R factor provides a good indicator of over-fitting, however it should be used with care. The following points should be borne in mind:

1. The 'Memory Effect'. Once a model has been refined against the whole set of reflections, further refinement against a working set alone does not perturb the fit to the free-set. Thus in this case R_{Free} is seriously underestimated. It is therefore important to exclude the Free-set throughout the whole refinement.

   Once the model is complete, it can be refined against all the data, but no new parameters should be introduced. Brünger suggests that the bias to the free-set caused by refinement against all the data can be removed by applying simulated annealing from a moderate temperature.

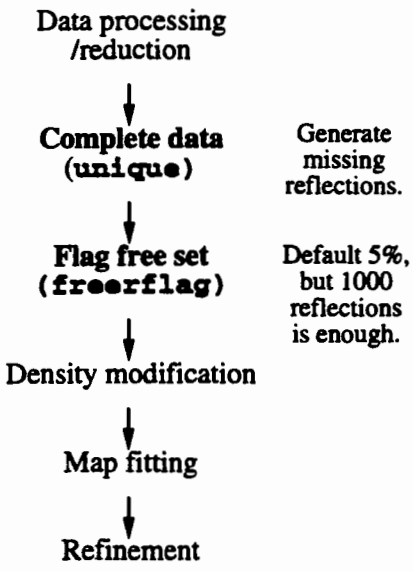


Figure 7: Free-R calculation within CCP4

2. Underestimation of R_{Free} due to NCS. Non-crystallographic symmetry in the crystal will lead to strong correlations between reflections in reciprocal space. Reflections in the free-set may be strongly correlated with other reflections in the working set, and thus no longer provide an independent indication of over-fitting. In this case R_{Free} will be systematically underestimated.

   It may be possible to obtain an unbiased estimate of R_{Free} by picking the Free-set reflections in thin resolution shells, thus ensuring that all reflections correlated to a Free-set reflection will also be in the Free-set.

3. The Free-R validates the protocol, rather than the model. It can be used to determine whether a set of observed data justifies the modelling of solvent, thermal motion, etc., but not whether an individual structural feature is correct.

4. Should the Free reflection be used in map calculations? Omitting the free-set may produce spurious map features, however including the reflections and fitting to the resultant density may bias R_{Free}. If a map fitting tool is used which performs real space refinement of fitted residues, then probably it is best to omit the free-set from the map calculation.

   Omitting the free-set from a map used in a phase improvement/density modification calculation has potentially more serious consequences (Cowtan & Main, 1996)

5 Conclusion

The statistical technique of Cross Validation, as implemented in the form of the Free-R factor, provides a more reliable validator of structure refinement than the R-factor because it is sensitive to over-fitting against noisy data. Its behaviour provides a strong indication of what parameters may be reasonably determined from a given data set. However, it does not provide a numerical indication of the magnitudes of coordinate errors in the model, and some care must be taken to avoid inadvertently biasing the free-set reflections with their experimental values. The problem of distiguishing between errors in observed data and errors in an atomic model is far from being solved, but the Free-R factor is an important step in the right direction.

References

[1] Brünger, A. T. (1992) *Nature* **355** 472-474.

[2] Brünger, A. T. (1995) *Method. Enzym.* invited paper, in press.

[3] Cowtan, K. D. and Main, P. (1996) *Acta Cryst.* **D52**, 43-48.

[4] Dodson, E.; Kleywegt, G. J. and Wilson, K. (1996) *Acta Cryst.* **D52**, 228 - 234.

SELF-VALIDATION : AN EXTENDED HAMILTON TEST

Alessia Bacchi, Victor S. Lamzin, Keith S. Wilson

European Molecular Biology Laboratory (EMBL)
c/o DESY, Notkestraße 85, 22603 Hamburg, Germany

INTRODUCTION

Observations and parameters

The refinement of a crystal structure is aimed to minimise the difference between the experimental electron density, obtained by Fourier transforming the observed structure factors obtained by measured amplitudes and calculated phases, and the electron density of the structural model represented by a set of variable parameters, which are refined. The refinement is usually performed in reciprocal space by least-squares minimisation of the differences between observed and calculated structure factors.

In X-ray crystallography a molecule is typically described as an ensemble of spherical atoms which oscillate harmonically around an equilibrium position, identified by the atomic fractional coordinates (3 parameters per atom). The atomic thermal displacement is anisotropic and is described by vibrational ellipsoids (6 parameters per atom). Static or dynamic disorder is accounted for by means of the site occupancy factor (1 parameter per atom). Ideally, ten parameters should be determined for each atom in the structure, including hydrogens. Moreover the modeling of solvent continuum, extinction or absorption effects requires the introduction of some additional parameters in the refinement. Unfortunately, very seldom experimental data contain enough information to allow a reliable determination of all the parameters which ideally fully describe the structure. It is therefore necessary to reduce the complexity of the model in order to gain reliability. The introduction of isotropic hydrogen atoms in calculated positions is a very common example of model simplification. Usually the crystallographic refinements start from a simple model, to which more and more detailed features are added progressively.

In general increasing the complexity of the model means to increase the number of independent parameters and produces a better agreement between calculated and experimental data, at the expense of a loss in the number of ways in which experimental errors can be accounted for.

As a very simple example, Figure 1 shows that although it is possible to fit perfectly three experimental (x,y) points with a parabola (obtaining an apparently very satisfactory R-factor=0.0%), it would be probably more realistic to choose a model with less parameters, *i.e.* a line, giving a worse R-factor (1.7%), but allowing for experimental errors.

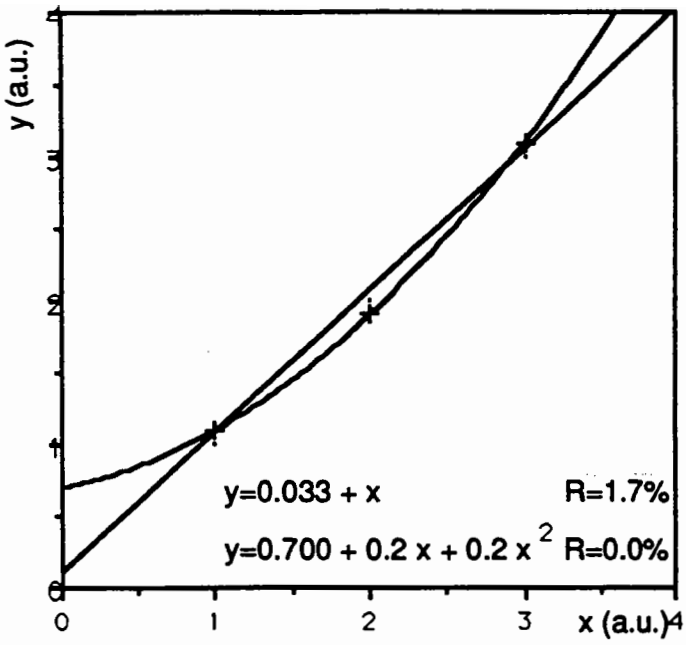


Figure 1.

It is convenient to express the number of degrees of freedom of the refinement as the number of observations minus the number of parameters (these are the degrees of freedom of the noise, and are not to be confused with the degrees of freedom of the molecule!). Since in the crystallographic experiment the number of parameters that effectively best describe the model is not known *a priori*, a validation procedure is required to avoid overfitting. This problem is particularly severe for proteins, where the number of observations available from the diffraction experiment is low if compared to the number of parameters typically used for the description of the structure. At resolution lower than 2.5 Å the number of observations no longer exceeds the number of parameters.

To overcome the problem of low data/parameters ratio, it is required to increase the number of degrees of freedom of the refinement, either by augmenting the number of observations or by decreasing the number of parameters used to describe the model. The former condition implies introduction of additional observational equations containing *a priori* information about the model in the form of restraints. Typically these are expressed as geometric or energetic conditions which the crystallographer might think the structure should satisfy (Waser, 1974; Sussman, Holbrook, Church & Kim, 1977; Konnert & Hendrickson, 1980). Restraints are included in the design matrix of the experiment as extra equations and are treated as observed data. Examples are: restraining bond distances to approach target values derived from accurately determined structures, as those contained in the Cambridge Structural Database (Allen, Kennard and Taylor, 1983); planarity restraints (Urzhumtsev, 1991); restraining bonded atoms to have comparable anisotropic thermal motion along the bond direction (rigid bond approximation, Hirshfeld, 1976), imposing non-crystallographic symmetry between stereochemically equivalent fragments (Bricogne, 1974).

The second way to increase the number of degrees of freedom of the refinement is to diminish the number of refined parameters. This is achieved by constraining them to obey exact conditions (Prince, Finger and Konnert, 1995). Examples are: imposing the space group crystallographic symmetry; introducing hydrogen atoms at their

calculated positions, riding on their carrier atoms; constraining occupancies of related complementary disordered atoms to sum up to one. Isotropic refinements are in fact anisotropic refinements where the ellipsoids describing the atomic thermal motion are constrained to be spheres, reducing the number of atomic thermal displacement parameters from 6 to 1.

The number of degrees of freedom of the refinement can be tuned by varying the number of restraints and constraints. It is necessary to ensure that this number is the most appropriate choice to describe the data, by carrying out a validation procedure in parallel with the refinement.

Cross- and self-validation

Two general kinds of validation methods are self-validation and cross-validation.

In self-validation descriptors are defined, either in reciprocal or in real space, to assess the quality of the refinement procedure. Examples are the R-factor, the real-space R-factor (Bränden & Jones, 1990), stereochemical criteria (Vriend, 1990; Laskowski, MacArthur, Moss and Thornton, 1993), maximum likelihood methods (Bricogne, 1984).

The most popular cross-validation evaluator in protein crystallography is the R free factor (Brünger, 1992). This is to some extent analogous to full statistical cross-validation which determines the power of the model to reproduce the experimental results and to predict unmeasured data. The rationale in cross-validation methods is to split the complete data set into a training set, on which the model is built, and a test set, on which the model is tested; the procedure is repeated by considering, in turn, every data subset as the test set. In crystallography it is not possible to test the model against all the possible data subsets, as this requires unrealistic computing time. The compromise adopted is to evaluate the quality of the refinement on the basis of a randomly selected subset of the data. Use of R free cross-validation in crystallography requires the omission of the reflections in the test set, for instance 10% randomly selected in the reciprocal space, Brünger (1992), from the refinement.

METHOD

The linear Hamilton R factor test

The question as to whether an improvement in R factor due to a decrease in the number of degrees of freedom is significant was first tackled by Hamilton (1964, pp. 157-162; 1965), who considered the problem of linear constrained refinements.. Hamilton defined the R-factor ratio $R=\frac{R1}{R2}$, where R1 and R2 are the r.m.s. R factors referring to the constrained and the unconstrained refinements, respectively. The null hypothesis is that the two refinements do not differ significantly; it is shown that :

$$R_{b,n-m,\alpha} = \sqrt{\frac{b}{n-m} F_{b,n-m,\alpha} + 1} \qquad (1)$$

where $F_{b,n-m,\alpha}$ denotes the F-test analysis of the variance ratio for b constraints and with n-m degrees of freedom; α is the probability of wrongly considering significant the improvement when in fact the second refinement gives no advantage compared to the first one (type I error). Usually refinements proceed from a more to a less constrained model, i.e. from less to more parameters, and the hypothesis that the

releasing of restrictions really improves the model should be tested. Hamilton's analysis refers to unconstrained refinement of m parameters against n data, giving n-m degrees of freedom. Introducing b linear constraints reduces the number of refined parameters leading to a higher number of degrees of freedom (n-(m-b)). The two r.m.s. R-factors, R2 and R1, are the relative estimated standard deviations of the distributions of the weighted Fo-Fc deviates for the two refinements:

$$\text{r.m.s. } R = \sqrt{\frac{\Sigma\, w_i (Fo_i - Fc_i)^2}{\Sigma\, w_i\, Fo_i^2}} \qquad (2)$$

The probability that the observed r.m.s. R-factor ratio R expresses a significant improvement is:

$$P(F_{b,n-m}) = 1 - I_x\left(\frac{n-m}{2}, \frac{b}{2}\right) \qquad (3).$$

I_x is the incomplete beta function (Press *et al.*, 1989) and $x = \frac{n-m}{n-m+bF}$. The behaviour of the function depends on the numerical values of $n1=\frac{n-m}{2}$ and $n2=\frac{b}{2}$, Figure 2. In protein crystallography these are typically large, and $P(F_{b,n-m})$ approaches a step-function centred at the critical point $R = \sqrt{\frac{n-m+b}{n-m}}$

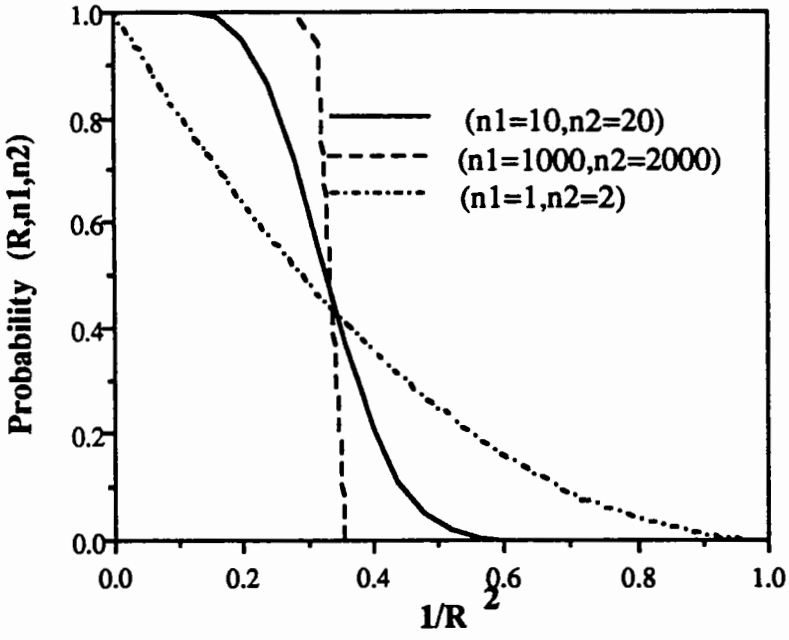


Figure 2. Probability that the improvement expressed by R is significant for different values of n1 and n2.

Extension for restrained refinements

Hamilton limited his analysis to the case of refinements differing by the presence of b linear constraints. The more general case is now examined where the number of observations and the number of parameters change and restraints are applied.. Consider two refinements with different degrees of freedom, Df1 and Df2, and the condition that Df1>Df2 and R1>R2, which means that the first is more constrained and gives a higher r.m.s. R-factor. Is the improvement in r.m.s. R-factor significant or does it merely reflect a reduction of the number of degrees of freedom? The number

of degrees of freedom is defined as the number of observation minus the number of parameters. A linear constraint expresses an exact linear relation between parameters and reduces the number of the refined parameters. A restraint introduces a condition that the system must obey within a certain degree of confidence expressed by a weighting coefficient. A restraint is an additional weighted observational equation. Both the introduction of constraints and of restraints increases the number of degrees of freedom. However some restraints are redundant or are applied only if certain conditions arise (*e.g.* anti-bumping) and these cannot be easily counted.

Therefore, we introduce a *restraints completeness* weighting coefficient w to define the effective number of observational equations:

$$N_{obs}=N_{refl}+w\,N_{restr} \qquad (4)$$

w=0 corresponds to completely unrestrained refinement and w=1 to a refinement where every restraint is treated as a full additional observation.
Let Nr1 and Nr2 be the number of reflections, S1 and S2 the number of restraints and P1 and P2 the number of parameters for the two refinements. The dimensionality (Dim) of the hypothesis is the difference between the number of degrees of freedom :

$$Dim=Df1-Df2=(Nr1+w1\cdot S1-P1)-(Nr2+w2\cdot S2-P2) \qquad (5)$$

Hamilton's linear hypothesis refers to the particular case where Nr1=Nr2, P1=P2, S2=0, w1=1, w2 is indeterminate and Dim=S1.
Since we are considering the case where R1>R2 when Df1>Df2, the condition Dim>0 must hold. This implies that w1 and w2 must satisfy the inequality:

$$w2 < \frac{Nr1 - Nr2 + P2 - P1}{S2} + w1\frac{S1}{S2} \qquad (6)$$

In the two-dimensional space spanned by w1 and w2, a straight line with intercept $\frac{Nr1 - Nr2 + P2 - P1}{S2}$ and slope $\frac{S1}{S2}$ separates the weight-allowed area from the weight-forbidden area, Figure 3.

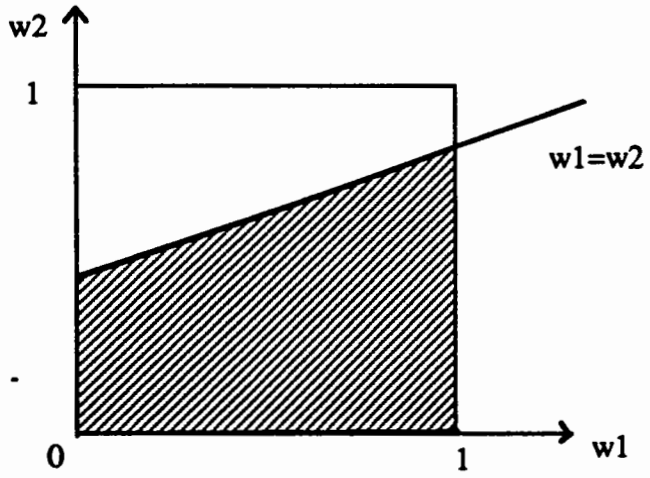


Figure 3. Weights for restraints. The shaded area represents the points (w1,w2) which satisfy the assumption that the dimensionality must be positive.

Since $\frac{S1}{S2} \geq 0$, if $\frac{Nr1 - Nr2 + P2 - P1}{S2} \geq 1$ there are no restrictions on w1 and w2.

If the intercept is between 0 and 1, then the greater the value of $\frac{S1}{S2}$ (i.e. the less restrained is the second refinement compared to the first), the greater the allowed value of w2 for a given w1.

Under these assumptions, the probability that the improvement is significant can be calculated by a modification of Equation 3:

$$P(F_{Dim},Df2) = 1 - I_x\left(\frac{Df2}{2},\frac{Dim}{2}\right) \qquad (7)$$

The critical point is $R = \sqrt{\frac{Df1}{Df2}}$ (8).

APPLICATION AND DISCUSSION

The refinement of xylanase at 1.5 Å (Lamzin, Dauter, Dauter, Bisgard-Frantzen, Halkier & Wilson, to be published) using *SHELXL*-93 (Sheldrick, 1993) is now considered.

	Reflections	Parameters	Restraints	r.m.s. R-factor (%)
Isotropic	34460	7460	6450	20.3
Anisotropic	34460	17770	20610	14.9

Comparison of isotropic and anisotropic refinement for xylanase at 1.5 Å.

Two refinements with isotropic and then anisotropic description of atomic thermal motion gave a ratio $R = \frac{r.m.s.\ Riso}{r.m.s.\ Raniso} = 1.36$.

The probability P(R,w1,w2) that the improvement expressed by R is significant depends on w1 and w2, which are unknown. Figure 3 shows the behaviour of the function P (R,w1,w2) in the hypothetical case when w1=w2. In this case for any value of w1=w2 the probability P (R=1.36) is essentially unity, indicating that the anisotropic model leads to a significant improvement..

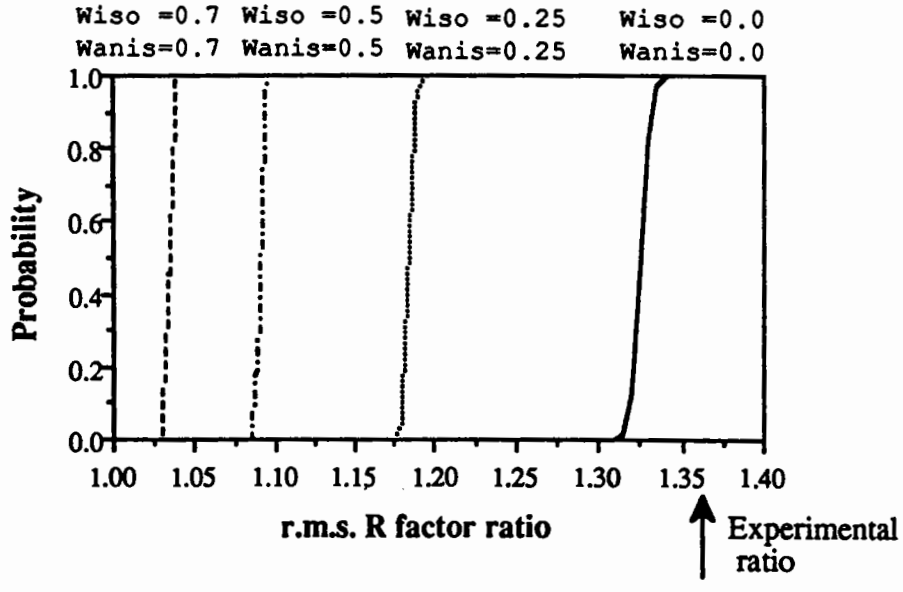


Figure 4. Probability of a significant improvement in R factor for different values of w1=w2 for anisotropic refinement of xylanase.

In the more general case, the probability is expressed as a function of R, w1 and w2. Figure 5(a) shows that for R = 1.36 the improvement is significant for almost any value of w1 and w2. The small region at the bottom of Figure 5(a) where the improvement in the r.m.s. R factor is not significant corresponds to an essentially unrestrained anisotropic model (w2≈0) which is not realistic. Suppose that a poorer ratio R = 1.21 (r.m.s.Raniso=11.6, r.m.s. Raniso=17.0) had been achieved, Figure 5(b). This would be significant provided w1 is not lower than the threshold limit given by the borderline between areas 1 and 2 in the plot. On further decrease of R, Figure 5(c), $P(R)$, becomes more selective and for $R \approx 1$ (i.e. r.m.s. Raniso ≈ r.m.s. Riso) $P(R)$, is high only for weighting schemes such that Dim ≈ 0, where the number of degrees of freedom is equal for the two refinements. Under this condition the anisotropic model would be clearly shown not to be statistically meaningful.

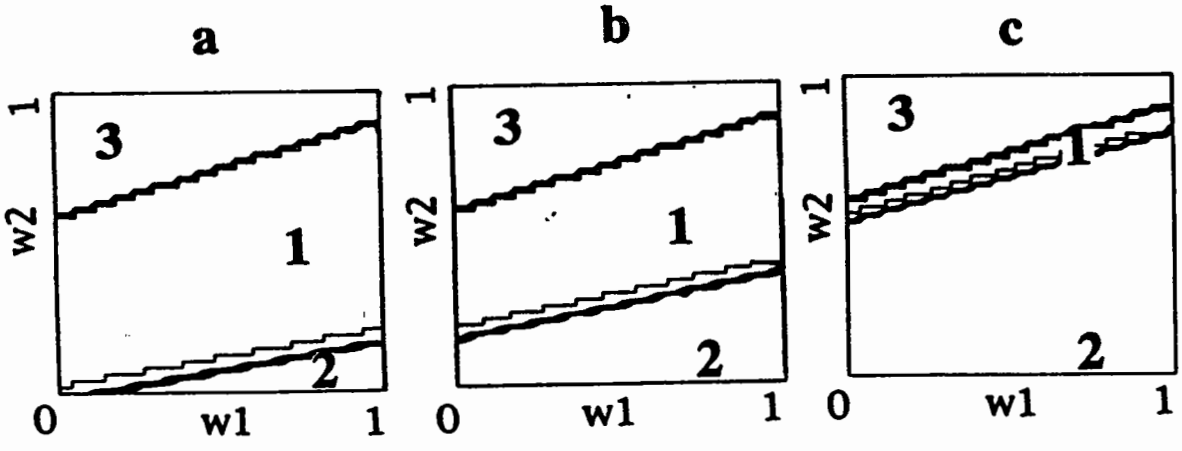


Figure 5. Probability isolines (20%, 40%, 60%, 80%, 100%) of r.m.s. R-factor ratio, R, as a function of weighting coefficients in the case of xylanase: a) for R=1.36 (actual value obtained with anisotropic refinement); b) for a hypothetically poorer improvement R =1.21; c) for R=1.03. Label 1 indicates the region where the improvement is significant, 2 where it is not significant, 3 indicates impossible weighting schemes where inequality (6) does not hold.

SUMMARY

An extension is developed for the self-validation Hamilton test (Hamilton, 1965) for crystallographic refinements. The method is based on the F-test and evaluates the significance of the R-factor ratio between two refinement protocols. The general case of two refinements carried out with different numbers and types of non-linear restraints is examined. The restraints are considered as extra observations weighted by a coefficient expressing their effective number. Robustness of the test in the presence of systematic errors in the estimation of weights and of the resulting non-Gaussian distribution of the residuals is achieved by the introduction of the restraints completeness coefficient w, which damps the instability in the R parameter by

varying the formal number of restraints. There exists a restriction on the weighting coefficients between the two refinements. Examination of the probability that the improvement in the model is significant as a function of (w1,w2) indicates which ranges of weights are allowed for introduction of a new set of parameters. The significance of the improvement obtained by moving from isotropic to anisotropic description of thermal parameters in the refinement of a protein at 1.5 Å resolution is used as an example.

The self-validation procedure based on the Hamilton test overcomes the problems related to omitting data and provides an objective monitor for refinement protocols. An intrinsic drawback in Hamilton's approach to validation is that it is based on a linear hypothesis, assumes Gaussian distribution for the deviates and is designed to work in the absence of systematic errors. Strictly speaking, not all crystallographic restraints are linear but an approximately linear behaviour can be assumed at least at the end of refinement. The weakness of purely statistical methods in the presence of systematic errors was clearly pointed out by Hamilton (1965) and the method is meant to be a guide to the crystallographer rather than a final verdict.

The results presented here are part of the work by A. Bacchi, V.S. Lamzin, and K.S. Wilson: 'A Self-validation Technique for Protein Crystallography: the Extended Hamilton Test'. *Acta Crystallogr. Sect. D*, **D52**, in press.

REFERENCES

Allen, F.H., Kennard, O. and Taylor, R. (1983) *Acc. Chem. Res.* **16**, 146-153.
Bränden, C-I and Jones, T.A. (1990) *Nature* **343**, 687-689.
Bricogne, G. (1974) *Acta Crystallogr.* **A30**, 395-405.
Bricogne, G. (1984) *Acta Crystallogr.* **A40**, 410-445.
Brünger, A. (1992) *Nature*, **355**, 472-474.
Hamilton, W.C. (1964) *Statistics in physical science*,. The Ronald Press Company, New York.
Hamilton, W.C. (1965) *Acta Crystallogr.* **18**, 502-510.
Hirshfeld, F.L. (1976) *Acta Crystallogr.* **A32**, 239-244.
Konnert, J.H. and Hendrickson, W.A. (1980). *Acta Crystallogr.* **A36**, 344-350.
Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) *J. Appl. Crys.* **26**, 283-291.
Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1989) *Numerical Recipes*, Cambridge University Press, Cambridge, New York, Port Chester, Melbourne, Sidney.
Prince, E., Finger, L.W. and Konnert, J.H. (1995) *International Tables of Crystallography, Vol. C*, 609-617. Kluwer Academic Publishers, Dodrecht, Boston, London.
Rogers, D. (1981) *Acta Crystallogr.*, **A37**, 734-741.
Sheldrick, G.M. (1993) *SHELXL*-93, Program for crystal structure refinement, *University of Göttingen*, Germany.
Sussman, J.L., Holbrook, S. R., Church, G.M. and Kim, S.H. (1977) *Acta Crystallogr.* **A33**, 800-804.
Urzhumtsev, A.G. (1991) *Acta Crystallogr.* **A47**, 723-727.
Vriend, G. (1990) *J. Mol. Graph.* **8**, 52-56.
Waser, J. (1974) *Acta Crystallogr.* **A30**, 261-264.

Full Matrix Least Squares

Lynn F. Ten Eyck*
Department of Chemistry and Biochemistry
University of California, San Diego
La Jolla, CA 92093-0654

May 7, 1996

> Knowledge is proud that he has learned so much;
> Wisdom is humble that he knows no more.
>
> William Cowper, *The Task*, bk. 6 (1785)

Refinement of macromolecular structures as a mathematical problem is not different from refinement of small molecule structures. Both are straightforward optimization problems. The difficulties arise because the macromolecular crystallographer rarely has sufficient data to answer questions at the same level of detail as the small molecule crystallographer. Nevertheless, he has a lot of data, and the temptation to over-interpret it is sometimes overwhelming. As a personal note, I began work on refinement of protein structures when I realized that despite having hundreds of thousands of observations I was unable to say with any certainty whether the heme group in deoxyhemoglobin was significantly domed. A quarter of a century later it is still not possible to put a direct measure of accuracy on estimates of the heme geometry in hemoglobin.

Structural questions about macromolecules can be posed on several levels, ranging from "What is the fold?" to "Is one of the bonds in the iron-sulfur cluster significantly different from the others?" The level of detail is widely variable, and structures which are adequate for the former purpose may not be adequate for the latter. Any of us who practice our craft for any length of time will see some beautiful-looking maps, which lead to structures in which we have high confidence - but we cannot put reliable numbers on that confidence. Similarly, we will see maps which are charitably described as obscure, where we can (perhaps) build the chain, but cannot be positive that the density we see is not a phase artifact. Sometimes we see both kinds of density in the same map. This presents us with a real problem. The accuracy of the structures we report is not uniform, but the methods for characterizing this information and reporting it to the users of the coordinates are very poor indeed.

There are a number of common practices in refinement of macromolecular structures which cause serious problems with the accuracy of the final structure. Some of these are omission of weak data, omission of low resolution data,

*Supported by a grant NSF/BIR 9223760 from the National Science Foundation.

improper treatment of solvent, and improper treatment of non- crystallographic symmetry restraints. Kleywegt and Jones [1, 2] discuss some of the problems that arise from these practices.

Many of these practices arise from the confusion of two distinct problems. The first problem is to solve the structure, which means to find the correct model. In this stage of the problem it is often highly appropriate to leave out weak data and concentrate on the strongest signal. It can also be appropriate to alter weights on restraints, relax non- crystallographic symmetry restraints, and generally let the molecule distort in order to fall into the best minimum. Once the model is determined there is the second problem of finding the best values for the parameters of the model. This is a quite different problem and requires different treatment of the data. Much confusion arises because both problems are generally handled by the same software, and superficially appear the same.

The method of analysis presented here is directed at two problems. The first problem is to derive a reliable method of estimating the uncertainty of each individual parameter, which works for all resolutions and for all forms of parameterizing or restraining the model. It is shown below how to ascertain which parameters are determined precisely and which are not, by methods which are not limited by low resolution data. It is also shown how to determine the effect of different ways of parameterizing the problem on the accuracy of the parameters. Practical analysis according to these methods is not complete, but the results are almost certain to be pessimistic. In the words of Ecclesiastes 1:18, *For in much wisdom is much grief: and he that increaseth knowledge increaseth sorrow.*

Another problem which can be addressed by the methods presented in this paper is to determine how the results of the crystallographic experiment can be improved. There are many open questions as to the "best practice" in any experimental field. For example, there are widely varying practices in the use of low resolution data, inclusion of weak reflections in refinement calculations, incorporation of non-crystallographic symmetry restraints, and the tradeoff between completeness and resolution in data collection. There are significantly different conceptual and mathematical descriptions of the models being refined. The mathematical and computational apparatus discussed in this paper provides a rigorous method for analysis of these questions. It appears that it may be possible to use these methods to determine optimal data collection protocols for answering specific questions about a particular structure at higher resolution, and will in general tell what must be done to achieve a specified level of accuracy in a structure determination. Due to space limitations this application will be given very short treatment.

Theoretical discussion of least squares analysis

The theory of least squares analysis of poorly determined systems is well advanced mathematically, but seldom used extensively in practice. The books by Lawson and Hanson [3], and by Golub and Van Loan [4] are highly recommended to the reader. Excellent material is also found in Diamond's discussion of real-space refinement [5]. The following derivations are completely general for all least squares problems, linear and non-linear. To avoid severe notational complexity the specific language of the crystallographic problem is deferred until

necessary.

The general problem of fitting a non-linear model function to a set of observations can be written as a minimization of

$$\Phi(\mathbf{x}) = \frac{1}{2}\sum_{j=1}^{N} w_j^2 \left(f_j(\mathbf{x}) - y_j\right)^2 \tag{1}$$

where $\Phi(\mathbf{x})$ is the sum of squares of residuals, y_j is an observed value, w_j is a weighting factor based on the reliability of y_j, and $f_j(\mathbf{x})$ is the function which calculates the theoretical value of the observable quantity given the parameters $\mathbf{x}$ and the index j which specifies the conditions of the observation. There are a variety of methods for finding the parameters $\mathbf{x}$ which minimize $\Phi(\mathbf{x})$. The commonly used methods for the macromolecular crystallographic problem are simulated annealing [6], conjugate gradients applied directly to the non-linear function itself [7], and conjugate gradients applied to the linear approximation to $\Phi(\mathbf{x})$ [8, 9]. Refinement of parameters in small molecule crystallography is normally done by directly solving successive linear approximations to $\Phi(\mathbf{x})$, a method known as full matrix least squares [10, 11]. All of these methods work, some faster than others. Generally speaking, simulated annealing has the largest radius of convergence, conjugate gradients applied to the non-linear function (especially as modified by Tronrud [12]) is the fastest, and full matrix least squares is the most accurate.

The simplest description of the linear approximation is to expand $\Phi(\mathbf{x})$ as a Taylor series about the minimum point $\phi_0 = \Phi(\mathbf{x}_0)$, where $\mathbf{x_0}$ is the set of parameters which minimize $\Phi(\mathbf{x})$. The expansion is

$$\begin{aligned}\Phi(\mathbf{x}) \;\approx\; & \phi_0 + \left\langle (\mathbf{x}-\mathbf{x}_0) \middle| \left(\frac{\partial\Phi}{\partial x_i}\right)_{\mathbf{x}_0} \right\rangle \\ & + \frac{1}{2}\left\langle (\mathbf{x}-\mathbf{x}_0) \middle| \left(\frac{\partial^2\Phi}{\partial x_i x_j}\right)_{\mathbf{x}_0} \middle| (\mathbf{x}-\mathbf{x}_0) \right\rangle + \ldots\end{aligned} \tag{2}$$

where the Dirac bra-ket notation expresses a column vector as $|x_i\rangle$, a row vector as $\langle x_i|$, and a matrix as $|x_{ij}|$. $\langle x|y\rangle$ is thus the inner product of the vectors x and y. Since the expansion is about a minimum, the gradient at $\mathbf{x}_0$ vanishes for all parameters x_i. Thus we have the approximation that (to second order)

$$\Phi(\mathbf{x}) \approx \phi_0 + \frac{1}{2}\left\langle (\mathbf{x}-\mathbf{x}_0) \middle| \left(\frac{\partial^2\Phi}{\partial x_i x_j}\right)_{\mathbf{x}_0} \middle| (\mathbf{x}-\mathbf{x}_0) \right\rangle \tag{3}$$

and, by differentiation,

$$\left| \left(\frac{\partial\Phi}{\partial x_i}\right)_{\mathbf{x}} \right\rangle \approx \left| \left(\frac{\partial^2\Phi}{\partial x_i x_j}\right)_{\mathbf{x}_0} \middle| (\mathbf{x}-\mathbf{x}_0) \right\rangle . \tag{4}$$

Given the first and second derivatives of $\Phi(\mathbf{x})$, Equation (4) can be solved for the correction to $\mathbf{x}$ which brings it closer to $\mathbf{x}_0$. The approximation is the use of second derivatives evaluated at $\mathbf{x}$ instead of $\mathbf{x}_0$, and the neglect of higher order terms in the Taylor series. Neither condition is a problem for parameter estimates close to $\mathbf{x}_0$. Note that the assumption that the function

can be approximated locally by a quadratic polynomial is equivalent to assuming that the matrix of second derivatives is constant.

An alternative is to expand the residuals in terms of the parameter shifts. In this formulation each weighted observation is expanded in a Taylor series as

$$w_j y_j = w_j f_j(\mathbf{x}) + w_j \left\langle \left(\frac{\partial f_j(\mathbf{x})}{\partial x_i} \right)_{\mathbf{x}} \middle| (\mathbf{x}_0 - \mathbf{x}) \right\rangle \tag{5}$$

where w_j is the weight associated with observation y_j. Writing the system of equations (5) as a matrix equation we have

$$\mathbf{A}(\mathbf{x} - \mathbf{x}_0) = \mathbf{r} \tag{6}$$

where $\mathbf{A}$ has m rows and n columns, $\mathbf{x}$ is a column vector of length n, and $\mathbf{r}$ is a column vector of length m, with elements $w_j(f_j(\mathbf{x}) - y_j)$. $\Phi(\mathbf{x})$ is given by

$$\Phi(\mathbf{x}) = \frac{1}{2}\langle r | r \rangle = \frac{1}{2}\mathbf{r}^T\mathbf{r} = \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \mathbf{A}^T\mathbf{A}(\mathbf{x} - \mathbf{x}_0) \tag{7}$$

where the superscript T denotes the transpose of a matrix or vector. It is well known [3] that the solution to (5) which minimizes $||\mathbf{A}(\mathbf{x} - \mathbf{x}_0) - \mathbf{r}||$ (and hence minimizes $\Phi(\mathbf{x})$) is the solution to the $n \times n$ matrix equation

$$\mathbf{A}^T\mathbf{A}(\mathbf{x} - \mathbf{x}_0) = \mathbf{A}^T\mathbf{r} \tag{8}$$

The equivalence of the two approaches is readily demonstrated by expanding the terms in the two formulations. The elements g_i of $\mathbf{g} = \mathbf{A}^T\mathbf{r}$ and h_{ij} of $\mathbf{H} = \mathbf{A}^T\mathbf{A}$ are given by

$$g_i = \sum_{k=1}^{m} w_k^2 (f_k(\mathbf{x}) - y_k) \left(\frac{\partial f_k(\mathbf{x})}{\partial x_i} \right) \tag{9}$$

$$h_{ij} = \sum_{k=1}^{m} w_k^2 \left(\frac{\partial f_k(\mathbf{x})}{\partial x_i} \right) \left(\frac{\partial f_k(\mathbf{x})}{\partial x_j} \right). \tag{10}$$

Differentiation of Equation (1) gives

$$\left(\frac{\partial \Phi(\mathbf{x})}{\partial x_i} \right) = \sum_{k=1}^{m} w_k^2 (f_k(\mathbf{x}) - y_k) \left(\frac{\partial f_k(\mathbf{x})}{\partial x_i} \right) \tag{11}$$

$$\left(\frac{\partial^2 \Phi(\mathbf{x})}{\partial x_i \partial x_j} \right) = \sum_{k=1}^{m} w_k^2 \left(\frac{\partial f_k(\mathbf{x})}{\partial x_i} \right) \left(\frac{\partial f_k(\mathbf{x})}{\partial x_j} \right) + \sum_{k=1}^{m} w_k^2 (f_k(\mathbf{x}) - y_k) \left(\frac{\partial^2 f_k(\mathbf{x})}{\partial x_i \partial x_j} \right) \tag{12}$$

The second derivative term on the right hand side of Equation (12) which is not found in Equation (10) vanishes as $\mathbf{x} \to \mathbf{x}_0$. Equations (4) and (8) thus converge to the same form.

The matrix equation

$$\mathbf{H}(\mathbf{x} - \mathbf{x}_0) = \mathbf{g} \tag{13}$$

is the set of *normal equations* for the least squares problem. Since a protein refinement can easily have 10^4 parameters, the size of the normal matrices can

become very large. Full matrix least squares is not normally applied to large proteins.

The normal equations can be solved by inverting $\mathbf{H}$,

$$\mathbf{H}^{-1}\mathbf{g} = \mathbf{H}^{-1}\mathbf{H}\,(\mathbf{x} - \mathbf{x}_0) = (\mathbf{x} - \mathbf{x}_0) \tag{14}$$

$\mathbf{S} = \mathbf{H}^{-1}$ is the covariance matrix times the mean square residual, which means that after scaling the elements of $\mathbf{S}$ are

$$s_{ij} = c_{ij}\sigma_i\sigma_j \tag{15}$$

where c_{ij} is the correlation coefficient between parameters i and j, and σ_i is the standard deviation of parameter i. Since the correlation of any parameter with itself is 1, the diagonal elements are the variances of the parameters determined by solving the normal equations. *The inverse of the normal matrix is the source of the detailed accuracy information from traditional small molecule least squares analysis of X-ray diffraction data.*

The matrix of correlation coefficients is often used to detect dependencies between variables in a least squares problem. Values of $|c_{ij}|$ close to 1 indicate dependencies. However, this is limited to the detection of pairwise dependencies. Higher order dependencies do not necessarily have pairwise components. Lawson and Hanson [3, page 72] give a 3×3 example of strongly interdependent variables in which the magnitude of the largest correlation is 0.49.

If there are insufficient observations to explicitly determine all parameters, the matrix $\mathbf{H}$ becomes singular and the inverse matrix is not defined. For crystallography this occurs if the resolution is low, which is a common case for macromolecules. All of the preceding analysis concerning the Taylor series expansions and normal equations is still valid up through Equation (13). Methods for minimizing $\Phi(\mathbf{x})$ which do not depend on inverting the matrix $\mathbf{H}$ (such as simulated annealing or conjugate gradients) will still find a minimum, but in a formal sense the variance of some of the parameters will be infinite. The minimum will not be unique.

Even singular normal equations can be solved by diagonalizing the matrix $\mathbf{H}$. The eigenvalues and eigenvectors of $\mathbf{H}$ are solutions to the matrix equation

$$\mathbf{H}\mathbf{v} = \lambda\mathbf{v} \tag{16}$$

where λ is an eigenvalue of $\mathbf{H}$ and $\mathbf{v}$ is the corresponding eigenvector of $\mathbf{H}$.

For the case in which $\mathbf{H}$ is a normal matrix for a least squares problem, we have the interesting result from Equation (7) that

$$\Phi\,(\mathbf{x}_0 + \mathbf{v}_i) = \frac{1}{2}\mathbf{v}_i^T\mathbf{H}\mathbf{v}_i = \frac{1}{2}\lambda_i\mathbf{v}_i^T\mathbf{v}_i = \frac{1}{2}\lambda_i \tag{17}$$

when $\mathbf{v}_i$ is the i^{th} eigenvector of $\mathbf{H}$ and λ_i is the corresponding eigenvalue. The eigenvectors of $\mathbf{H}$ specify combinations of parameters which are statistically independent of one another, and the eigenvalues are proportional to the reciprocal of the variance of those parameter combinations. Another way of expressing the same idea is that the eigenvectors which correspond to large eigenvalues are directions in which parameter shifts have a large effect on the sum of squares of the residuals, and thus are well determined. Eigenvectors which correspond to small eigenvalues have little effect on the sum of squares of the residuals and

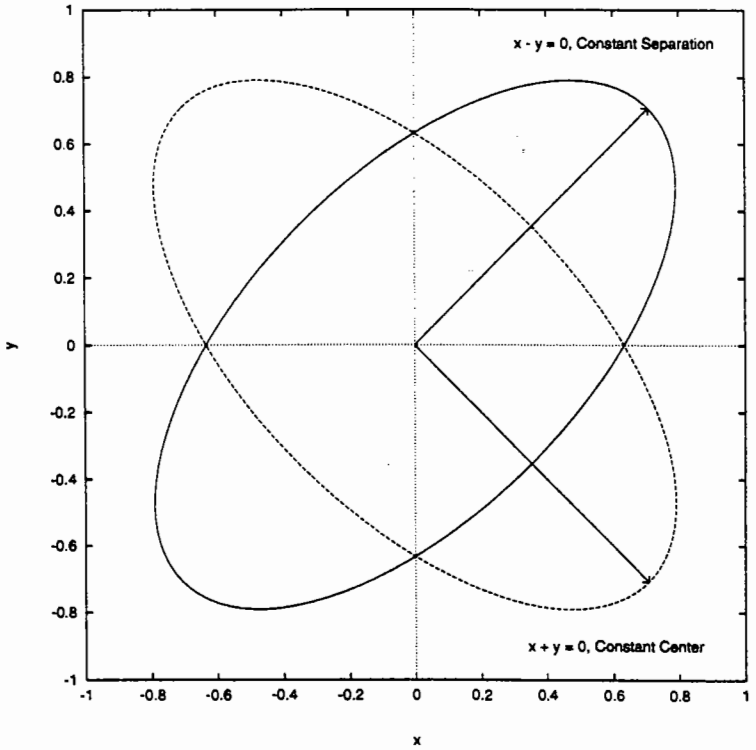


Figure 1: Both ellipses have the same eigenvectors, but the eigenvalues are swapped. The eigenvector in the $(+,+)$ quadrant corresponds to a shift of parameters which preserves the distance between two points. The eigenvector in the $(+,-)$ quadrant corresponds to a shift of parameters which preserves the center of mass of two points. The two ellipses reflect situations in which either the separation is more accurately known than the position, or in which the position is known more accurately than the separation.

thus correspond to poorly determined combinations of parameter shifts. (In fact the reciprocals of the eigenvalues are proportional to the variances of the corresponding combinations of parameters.)

This situation is illustrated for a two-parameter case in Figure 1. The ellipses are contours of constant $\Phi(\mathbf{x})$ in the second order approximation. The principal axes of the ellipses correspond to the variances of the parameters. The short axis of the ellipse gives the direction in which $\Phi(\mathbf{x})$ has the most sharply determined minimum.

Undetermined and Poorly Determined Systems

The inverse of a matrix can be constructed from the eigenvectors and eigenvalues of the matrix. If $\mathbf{V}$ is the orthogonal matrix constructed so that the columns of $\mathbf{V}$ are the eigenvectors $\mathbf{v}$ of $\mathbf{H}$, it is easily shown that

$$\mathbf{H}^{-1} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T \tag{18}$$

where $\mathbf{\Lambda}^{-1}$ is a diagonal matrix containing the reciprocals of the eigenvalues of $\mathbf{H}$. If $\mathbf{H}$ is singular some of the eigenvalues are zero, and $\mathbf{\Lambda}^{-1}$ is not defined.

However, if we define the *pseudo-inverse* of $\mathbf{\Lambda}$ as

$$\mathbf{\Lambda}^{+} = \begin{cases} 1/\lambda_i & \lambda_i > 0 \text{ and } i = j \\ 0 & \lambda_i = 0 \text{ or } i \neq j \end{cases} \tag{19}$$

then

$$\mathbf{\Lambda}^{+}\mathbf{\Lambda} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where the matrix product yields an identity matrix of the same rank as $\mathbf{H}$, with the remainder of the product being 0. The equation corresponding to 18 is $\mathbf{\Lambda}^{+}$ give

$$\mathbf{H}^{+} = \mathbf{V}\mathbf{\Lambda}^{+}\mathbf{V}^{T}. \tag{20}$$

The elements of $\mathbf{H}^{+}$ contain the same correlation information and variance information as the elements of $\mathbf{H}^{-1}$, except that it applies only to the parameter combinations which are in fact still determined by the data. The pseudo-inverse is identical to the inverse if the matrix $\mathbf{H}$ is of full rank.

This apparatus provides a complete mechanism for determining which parameters of a model are actually determined by the least squares procedure. It also gives direct measures of the precision of the determinations of the parameters for those parameters which are actually derived from the data. Preliminary calculations on two small molecules and a protein have shown that even singular crystallographic systems contain a large number of large eigenvalues, and hence many accurately determined parameters.

Separation of restraints from data

There are several different methods for applying restraints, and there are different degrees of approximation that can be used in computing the elements of the normal matrices. It is important that the effects of these different approaches be understood precisely. When the restraints are put into the least squares calculation as additional observations to be fit, the matrix $\mathbf{A}$ of Equation (6) can be partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix} \text{and} \mathbf{A}^{T} = [\mathbf{A}_1^{T}\mathbf{A}_2^{T}] \tag{21}$$

where all of the experimental observational equations are in $\mathbf{A}_1$ and all of the restraint equations are in $\mathbf{A}_2$. If we attach an explicit scale factor K_r to the equations of restraint the matrix $\mathbf{H}$ becomes

$$\begin{aligned} \mathbf{H} &= \mathbf{A}^{T}\mathbf{A} \\ &= \mathbf{A}_1^{T}\mathbf{A}_1 + K_r^2\mathbf{A}_2^{T}\mathbf{A}_2 \\ &= \mathbf{H}_1 + K_r^2\mathbf{H}_2 \end{aligned} \tag{22}$$

This directly separates the contributions of the two portions of the problem to the solution and will at long last clarify the effects of different restraint schemes on the results of a crystal structure refinement.

Construction of $\mathbf{H}$ from Equation (22) has the advantage that the quality of the parameters and the goodness of fit can be studied as a function of K_r to determine the correct relative weight to assign to the restraints. The benefit of this approach over Brünger's [13] is that all of the data can be used while still avoiding overfitting. Brünger's cross-validation approach requires that a

fraction of the data not be used in the refinement so that it can be used as an objective check on the progress of the refinement and on the validity of changes in parameters. In cases which are poorly determined it is not desirable to give up a fraction of the data if it can be avoided. (It should be noted that cross validation is good for testing other things besides K_r, such as the validity of basic changes in the model. It is not yet clear whether the methods being developed here could replace R_{free} for those purposes.)

Data Collection Protocol Analysis

Substitution of crystallographic variables into Equations (7) and (14) gives

$$h_{ij} = \sum_{hkl} w_{hkl} \left(\frac{\partial |\mathbf{F}^c_{hkl}|}{\partial x_i} \right) \left(\frac{\partial |\mathbf{F}^c_{hkl}|}{\partial x_j} \right) \tag{23}$$

which shows that the normal matrix *does not depend directly on the observed data.* The normal matrix depends on the model, the set of observations which are included in the calculation, and the statistical weight assigned to each observation, but does not depend on the values of the observations. The values of the parameters of the model do depend on the data, and this does affect the values of the elements of the normal matrix.

It is thus possible, given a model, to evaluate the effect of different data collection protocols on the accuracy with which the parameters will be determined. This formalism will decisively answer the question as to whether the omission of data observed at less than 2σ harms the accuracy of the model (it does), and settle the wars concerning the inclusion or omission of data inside the 6Å sphere during refinement. It will also tell specifically how the collection of additional data will improve the accuracy of the model, subject to the assumption that the model does not change dramatically in view of the new data. For example, just how good does your data have to get before you can tell if one bond in your iron-sulfur cluster is significantly different from the others? Are you better off getting more data, or improving the accuracy of the data you already have?

References

[1] Gerard J. Kleywegt and T. Alwyn Jones. Where freedom is given, liberties are taken. *Structure*, 3(6):535–540, 1995.

[2] Gerard J. Kleywegt and T. Alwyn Jones. Good model-building and refinement practice. In R. M. Sweet and C. W. Carter Jr., editors, *Macromolecular Refinement*, Methods in Enzymology. Academic Press, Orlando, in press.

[3] Charles L. Lawson and Richard J. Hanson. *Solving Least Squares Problems.* Prentice-Hall, Englewood Cliffs, New Jersey, 1974.

[4] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, Baltimore, 1989.

[5] R. Diamond. A real-space refinement procedure for proteins. *Acta Cryst.*, A27:436–452, 1971.

[6] A.T. Brünger, J. Kuriyan, and M. Karplus. Crystallographic R-factor refinement by molecular dynamics. *Science*, 235:458–460, 1987.

[7] D. E. Tronrud, L. F. Ten Eyck, and B. W. Matthews. An efficient general-purpose least-squares refinement program for macromolecular structures. *Acta Cryst.*, A43:489–501, 1987.

[8] John H. Konnert. A restrained-parameter structure-factor least-squares refinement procedure for large asymmetric units. *Acta Cryst.*, A32:614–617, 1976.

[9] J. H. Konnert and W. A. Hendrickson. A restrained-parameter thermal-factor refinement procedure. *Acta Cryst.*, A36:344–350, 1980.

[10] G. H. Stout and L. H. Jensen. *X-ray Structure Determination: A Practical Guide*. John Wiley and Sons, New York, 1989.

[11] G. M. Sheldrick. *SHELXL-93, a Program for the Refinement of Crystal Structures from Diffraction Data*. Institut fuer Anorg. Chemie, Goettingen, Germany, 1993.

[12] D.E. Tronrud. Conjugate direction minimization - an improved method for the refinement of macromolecules. *Acta Crystallographica*, A48:912–916, 1992.

[13] Axel T. Brünger. Free R-value - a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355(Jan. 30):472–475, 1992.

Least-Squares Refinement of Macromolecules: Estimated Standard Deviations, NCS Restraints and Factors Affecting Convergence

George M. Sheldrick, Institut für anorganische Chemie der Universität Göttingen, Tammannstr. 4, D37077 Göttingen.

1. Introduction

This contribution is based on the experiences of the author in (mis)using a small molecule structure refinement program (SHELXL-93) for macromolecules, and describes some of the features that it is hoped to include in the next release (SHELXL-96 ?).

This program uses an exact structure-factor summation rather than an FFT approximation, and so is extremely slow; on the other hand it is very general, and includes a number of features not usually found in programs written specifically for macromolecules. It is valid for all space groups and types of structure, and includes restrained, rigid group and riding models, restrained anisotropic refinement, anomalous dispersion, CIF output etc. It can handle Laue data, twins, and complicated disorder. For the least-squares refinement either (blocked) full-matrix or conjugate-gradient solution of the sparse normal equations may be employed.

It is to be expected that the use of SHELXL for macromolecules will be restricted to small structures (say not more than 5000 unique atoms) at high resolution (2 Å or better). For further details of macromolecular applications of the program see Sheldrick & Schneider (1996).

2. How to Estimate Esds in Atomic Positions

Given a small protein and data to (almost) atomic resolution, it is indeed possible to obtain *estimated standard deviations* (shortly to be renamed *standard uncertainties*) in atomic positions (and the geometrical parameters derived from them) by similar methods to those used for small molecules.

The structure should first be refined to convergence by conjugate-gradient solution of sparse-matrix normal equations (CGLS). Then one final full-matrix cycle should be performed with zero damping and zero shift multiplier (L.S. 1 and DAMP 0 0) and all restraints switched off. Restraints and Marquardt damping would lead to underestimated esds. All the reflection data should be used, i.e. no threshold should be used to suppress weak data, and no resolution shells should be excluded. If a full-matrix cycle would take longer than a week or require the purchase of extra memory, an adequate compromise is BLOC 1 to define a full-matrix block consisting of all geometrical but no thermal displacement parameters.

SHELXL uses the full covariance matrix and the estimated unit-cell errors to estimate the esds in ALL dependent parameters (for the esd in the angle between two least-squares planes a small approximation is involved).

Fig. 1 shows the distribution of the (three-dimensional) atomic positional esds of the fully occupied carbon atoms as a function of the effective B-value ($B=8\pi^2U_{eq}$, where U_{eq} is one third ot the trace of the orthogonalised U^{ij}-tensor) for a cytochrome c_6 refined anisotropically against data collected to 1.1 A by Frazão et al. (1995), and Fig. 2 shows the corresponding diagram for the oxygen atoms (including the solvent waters). There is a surprisingly good correlation between the esd and B; it is not far from linear, and a quadratic function in B would be even closer. The slope of the curve depends inversely on the atomic number; the points for nitrogen lie in between those for carbon and oxygen. The esds for the bond lengths show similar trends when plotted against the average B-values of the two atoms concerned; Fig. 3 shows the C-C bond length esds for this structure (excluding disordered atoms). The esds in the bond lengths are smaller than those in the (3-dimensional) atomic positions by about the factor of $\sqrt{(2/3)}$ expected for uncorrelated atoms. Preliminary tests on other small proteins refined at very high resolution give similar results to those shown here for cytochrome c_6.

Fig. 2. Oxygen atom positional esds (Angstroms)

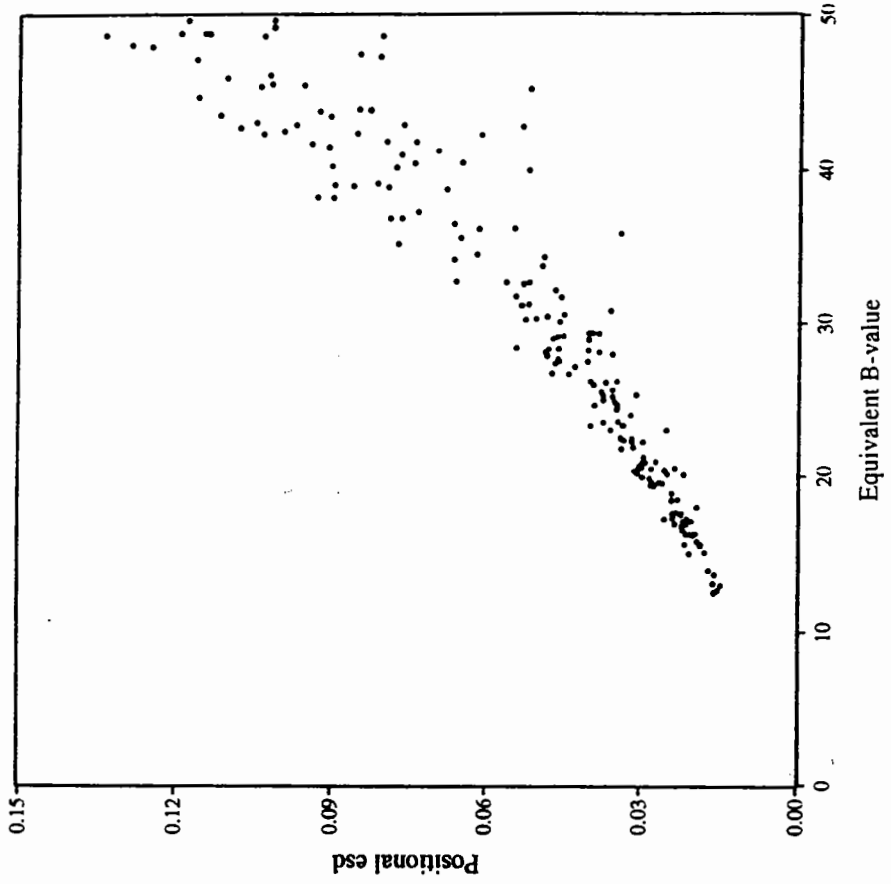


Fig. 1. Carbon atom positional esds (Angstroms)

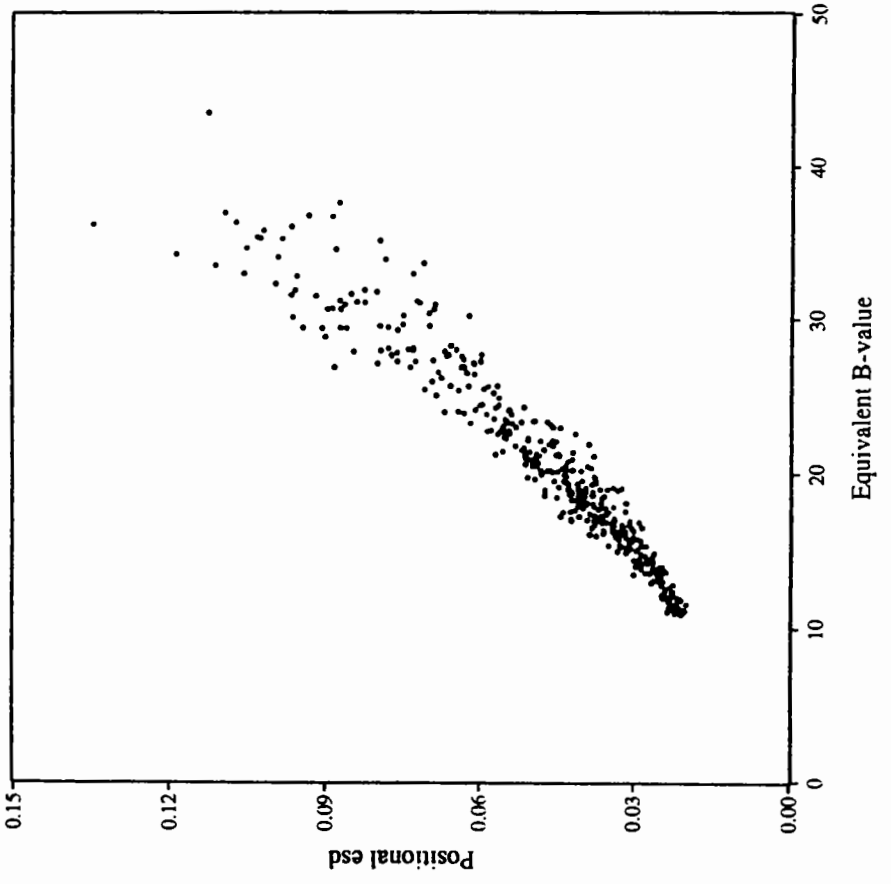


Fig. 3. C-C bond length esds (Angstroms)

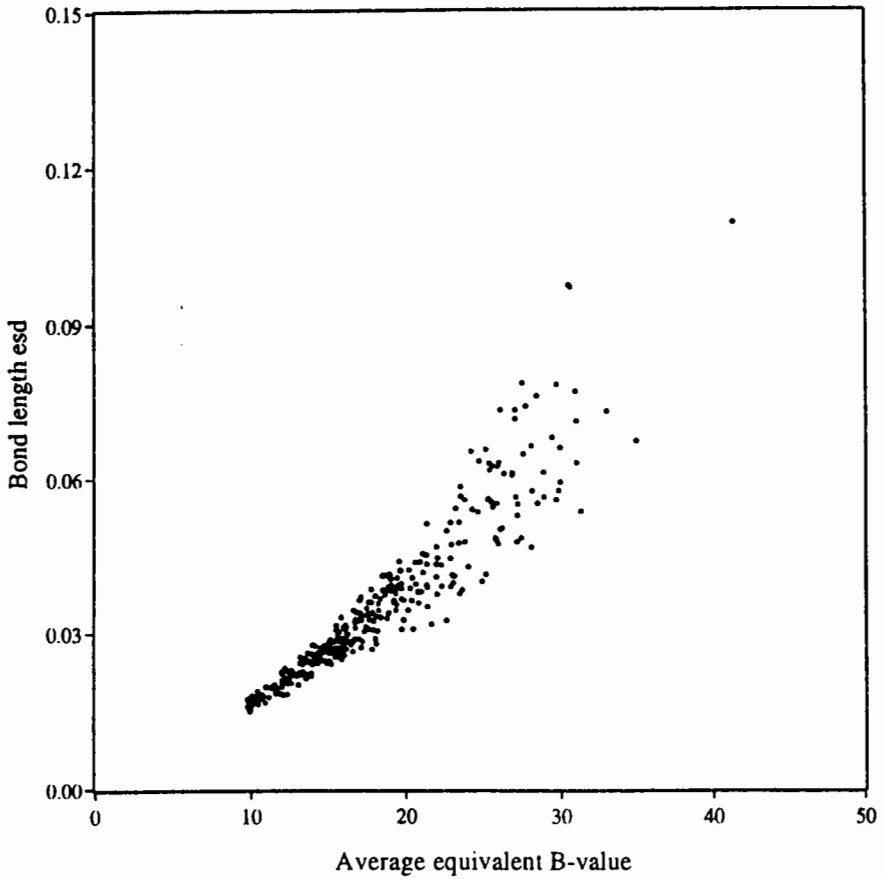
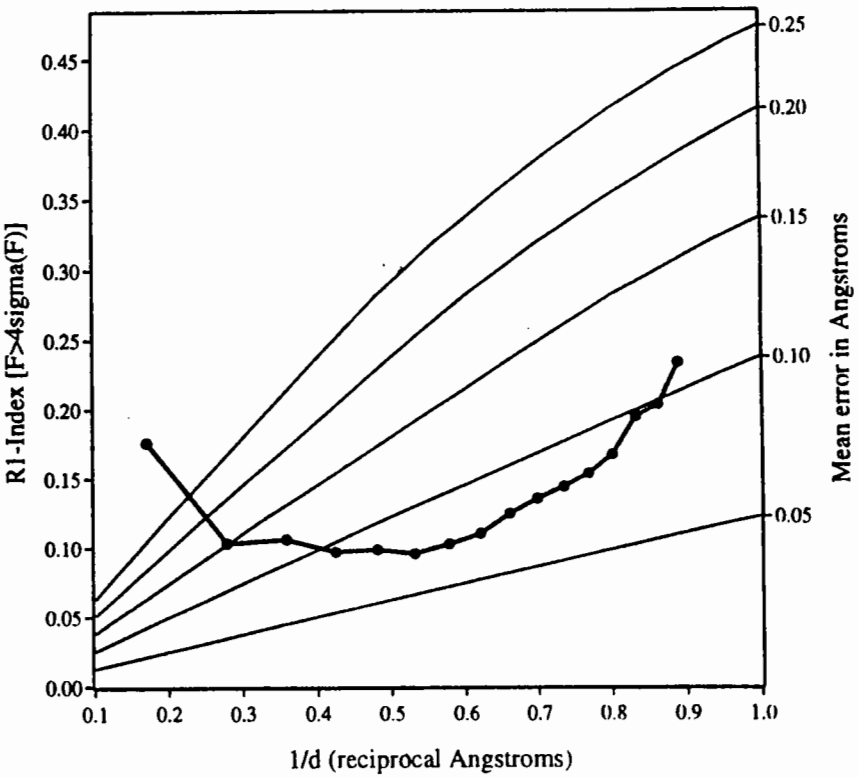


Fig. 4. Luzzati plot



Cruickshank (1949, 1960, 1995) has made some most instructive comments on the treatment of errors in crystallographic least-squares refinement. In 1995 he suggested a formula for the average atomic positional error in terms of the R-index, the resolution, and the completeness of the data. From the preliminary investigations presented here, it looks as though it should be possible to extend Cruickshanck's *Diffraction Precision Indicator* to obtain quite reasonable estimates of individual errors in atomic positions by adding empirical terms in B, Z (the atomic number) and perhaps also the occupancy. As Cruickshank pointed out, the widely used Luzzati (1952) plot was never intended to estimate such errors, although it has often been used for this purpose. The Luzzati plot for the cytochrome c_6 (Fig. 4) does in fact give numbers of roughly the right order of magnitude.

3. Rigid-Group refinement of the initial Model

Full-matrix refinement has other applications in macromolecular refinement, especially for problems with high correlations and a relatively small number of parameters, as for example in the early stages of refinement after solving a structure by molecular replacement.

Molecular replacement may well give us a model consisting of domains that can be refined as rigid groups, linked by hinges that are better described by restraints. The resolution range may be restricted, and fixed or group temperature factors employed. At this stage of the refinement, there are not many parameters, but there may be large correlations between them. A good way to handle these is by full-matrix refinement, usually with Marquardt damping.

4. Anti-Bumping Restraints

Refinement at low (>2Å) resolution can result in unwanted contacts involving solvent molecules and even main-chain atoms. How can one deal with this situation without the need for user intervention ?

SHELXL (now) checks all short non-bonded distances, taking symmetry equivalents into account, before each cycle. If two atoms that are not linked by one, two or three bonds in the connectivity array are too close, an anti-bumping restraint is generated to push them apart.

H•••H anti-bumping restraints may also be generated automatically and used to discourage energetically unfavourable side-chain rotamers.

5. Non-Crystallographic Symmetry (NCS) Constraints and Restraints

NCS is usually applied as a *constraint*. Structure factors are calculated for one repeating unit (monomer) defined using a *mask*, and the contributions from the NCS-related units are then found by applying a *rotation matrix* and *translation vector*. This method is fast but inflexible, and requires a mechanism for finding and possibly refining the mask, matrix and vector.

For small (and many macro-)molecules, it is not unusual to have more than one chemically identical molecule in the asymmetric unit. A very simple and effective restraint is to assume that all bond lengths (1,2-distances) and all 1,3-distances (i.e. distances through one angle) can be restrained so that chemically equivalent distances are equal. In SHELXL this requires one instruction (SAME) per equivalent molecule. The torsion angles are not restrained, and it is assumed that the differences between the molecules are purely conformational. In practice this is an excellent assumption, and since its introduction in SHELXL-93, this restraint has found extensive application.

This procedure has found some application in macromolecules - for example by restraining the equivalent 1,2- and 1,3-distances to be equal in the sugar and phosphate groups of oligonucleotides, or to take advantage of the 4- or 8-fold redundancy of chemically equivalent distances in heme groups - but for proteins the idea of making all copies of each amino-acid equal is less appropriate because some amino-acids may occur often and others only once in typical small proteins. In practice it is better to restrain the 1,2- and 1,3-distances in proteins to standard target values, for example those recommended by Engh & Huber (1991), which are also used in the standard restraints dictionary for SHELXL.

The idea can be carried one bond further for macromolecules, and the corresponding 1,4-distances restrained to be equal. Similarly, the isotropic U-values of the non-crystallographic symmetry related atoms can also be restrained to be equal. SHELXL applies NCS as a *local restraint* rather than a *global constraint*. This method is slower, but is much more flexible, and is much easier to apply than a NCS *constraint* because it does not require a mask, matrix or vector

3-Fold NCS could be specified by the SHELXL instructions:

```
NCSY 1000 N_1001 > OT2_1109
NCSY 2000 N_1001 > OT2_1109
```

where the residues in the three chains are numbered 1001-1109, 2002-2109 and 3001-3109.

Fig. 5 shows the results of applying such NCS restraints to the refinement against 1.7 Å data for a Rei_v immunoglobulin mutant containing two light-chain monomers in the asymmetric unit. The structure exhibits good two-fold non-crystallographic symmetry. It will be seen that the absolute differences in the NCS-related phi- and psi-angles are all small, but that a few of the chi-angles differ for lysine residues (and one arginine) that project into the solvent and so are poorly defined.

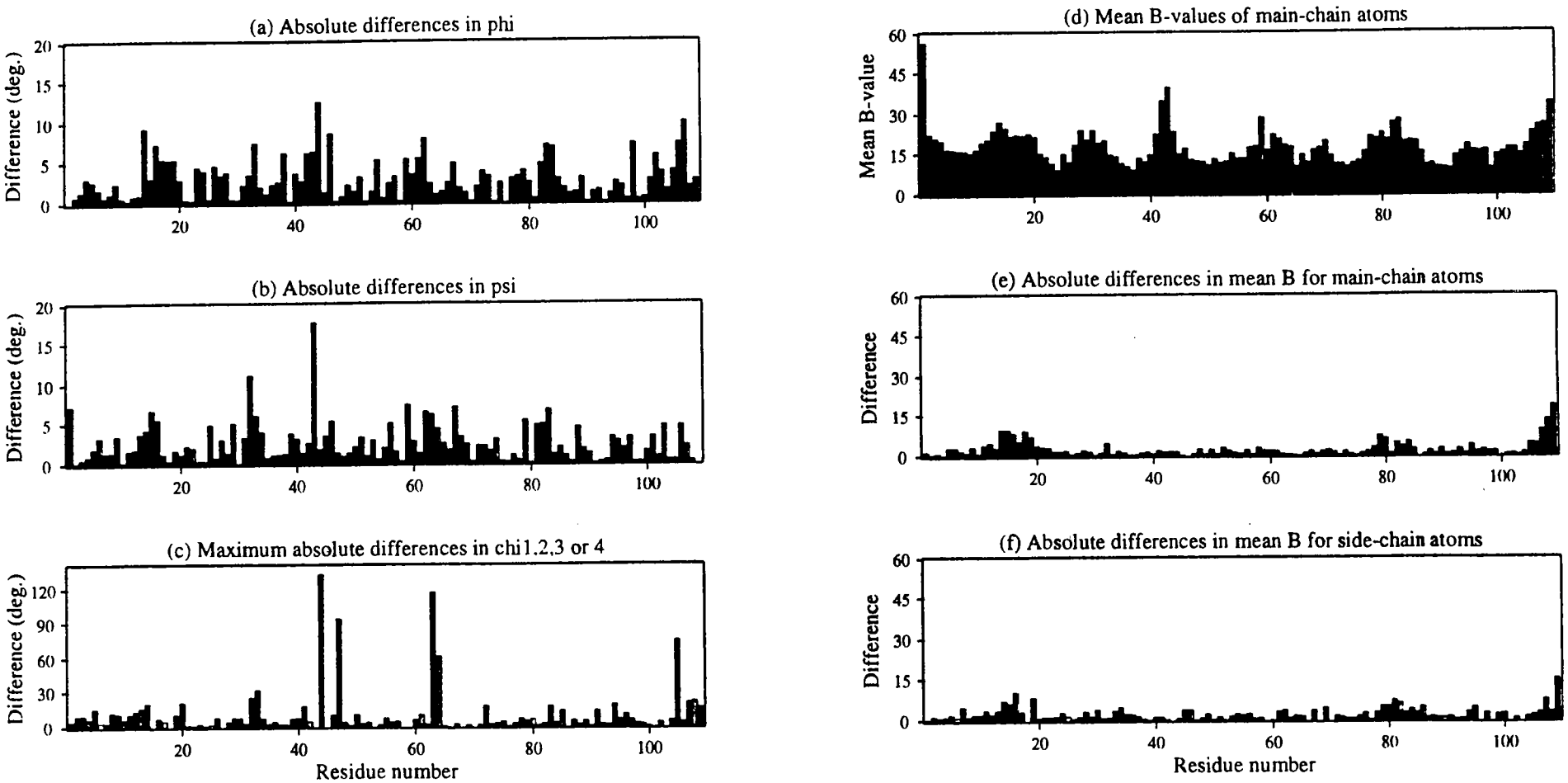


Fig. 5. Comparison of monomers related by twofold non-crystallographic symmetry.

The NCS deviations of the arginine could be eliminated by refining both residues with two-fold disorder, but an investigation of the lysine deviations showed that these arose because the 1,4-distances are the same for the gauche$^+$ and gauche$^-$ conformations, and so for unbranched chains the NCS restraints do not distinguish between these. On the other hand, these different conformations, although poorly defined, are chemically quite plausible.

The comparison of the mean B-values ($B=8\pi^2U$) shows that in general the agreement is excellent, but there are small deviations around residues 16 and 80 caused by contact with other molecules not related by the NCS, and there are also deviations for the C-terminus for which the density is poorly defined. The $3F_o$-$2F_c$ maps after application of the NCS-restraints were of excellent quality.

6. Disorder made simple

The disorder components are included with the same atom names in the same residues but identified by different PART numbers. Atoms in PART 1 may bond to other atoms in PART 1 and also to those in PART 0, but not to those in PART 2 etc. All other instructions are the same as for non-disordered residues. The program works out itself how to apply the restraints, add H-atoms etc.

```
RESI 38 SER
N   3 0.77141 0.92674 0.00625  11.0  0.10936
CA  1 0.78873 0.97402 0.07449  11.0  0.13706
PART  1
CB  1 0.83868 1.04271 0.05517  41.0  0.11889
OG  4 0.89948 1.00271 0.02305  41.0  0.18205
PART  2
CB  1 0.84149 1.03666 0.06538 -41.0  0.14933
OG  4 0.83686 1.10360 0.01026 -41.0  0.17328
PART  0
C   1 0.74143 1.01670 0.10383  11.0  0.08401
O   4 0.70724 1.02319 0.06903  11.0  0.10188
```

The PART numbers used by SHELXL map exactly onto the codes A,B,C... used by the PDB format to flag disorder, so this way of representing disorder should be completely compatible with any program that adheres rigorously (as does SHELXL in its PDB output file) to the Brookhaven PDB rules. In this example, the use of a free variable for the refinement of common occupancies as p and 1-p for the two disorder components should also be noted.

7. Making the most of weak data

By small-molecule standards, ALL macromolecular data are weak ! Throwing away weak data (by imposing a threshold of say $F>4\sigma(F)$) may artificially improve the R-factor, but it wastes valuable experimental information, usually obtained at considerable effort.

Refinement against F^2 enables ALL data to be used (properly weighted). It is not advisable to use all data when refining against F because of the difficulty of deriving $\sigma(F)$ from $\sigma(F^2)$ when F^2 is small or negative.

The conventional index $R = \Sigma|F_o-F_c| \; / \; \Sigma F_o$ is still useful for the comparison of structures determined at similar resolution and data completeness, even when refining against F^2; it is difficult to fudge it by adjusting the weights !

8. How to avoid Local Minima

Refinement techniques that ignore correlations are more likely to appear to stick in *false minima*. For example, if a structure has been solved by molecular replacement, some of the loops may well be displaced from their correct positions. Direct minimisation without taking correlations involving restraints into account may well not be able to move the loops, because (if a well refined search structure has been used) the restraints will all hold well already. The Konnert-Hendrickson (1980) solution of the sparse-matrix normal equations employed in SHELXL takes all

correlations introduced by the restraints into account, and is able to refine all parameters in the same refinement cycle, which also increases the radius of convergence.

Refinement against F^2 for ALL data without the use of a sigma threshold makes the most of the diffraction data, but it may well be advisable to extend the data to higher resolution gradually during the refinement.

The presence or absence of local minima depends very much on the form of the function that is being minimised, and in particular upon the contributions to this function from the restraints. UNIMODAL restraint functions, such as distances, angles, chiral volumes, least-squares planes and $\Delta(U^{ij})$ restraints do not generate additional local minima - indeed they may remove them - whereas MULTIMODAL functions, i.e. functions that themselves have multiple minima, will usually generate extra local minima. Examples of such multimodal functions are torsion angle and hydrogen bond restraints. Fortunately we have a choice - indeed we must make a choice - of which prior information we should include in the refinement in the form of restraints, and which should be kept in reserve to check the validity of the structure. MULTIMODAL functions are clearly better employed as criteria for the correctness of a structure. Very fortunately, torsion angles and hydrogen bonds and other non-bonded contacts are the very functions most useful in PROCHECK (Laskowski, MacArthur, Moss & Thornton, 1993) and other programs for independently verifying the structure !

The author is grateful to all the guinea pigs who were kind enough to test these (and other less successful) ideas, and for their diplomatic suggestions for improvements, and in particular to Isabel Usón, Miene Schäfer, Thomas Schneider and Johan Wouters who provided examples used to illustrate this talk.

References

Cruickshank, D.W.J. (1949). *Acta Cryst.*, **2**, 65-82.

Cruickshank, D.W.J. (1960). *Acta Cryst.*, **31**, 774-777.

Cruickshank, D.W.J. (1995). *Structure Refinement Meeting, York, UK.*

Engh, R.A. & Huber, R. (1991). *Acta Cryst.* **A47**, 392-400.

Frazão, C., Soares, C.M., Carrondo, M.A., Pohl, E., Dauter, Z., Wilson, K.S., Hervás, M., Navarro, J.A., De la Rosa, M.A. & Sheldrick, G.M. (1995). *Structure*, **3**, 1159-1169.

Hendrickson, W.A. & Konnert, J.H. (1980). *Computing in Crystallography*, edited by R. Diamond, S. Ramaseshan & K. Venkatesan, pp. 13.01-13.25. I.U.Cr. and Indian Academy of Science: Bangalore, India.

Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. (1993). *J. Appl. Cryst.*, **26**, 283-291.

Luzzati, P.V. (1952). *Acta Cryst.* **5**, 802-810.

Sheldrick, G.M. & Schneider, T.R. (1996). *Methods in Enzymology*, edited by C.W. Carter, Jr. & R.M. Sweet, in press.

Torsion angle dynamics refinement of the Chaperonin GroEL at 2.8 Å resolution

Paul D. Adams[2], Kerstin Braig[3], Luke M. Rice[1,2], and Axel T. Brünger[1,2]

[1]The Howard Hughes Medical Institute and [2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520.

[3]MRC Laboratory of Molecular Biology, Cambridge, CB2 2QH.

The refinement of a crystal structure consists of fitting an atomic model to the diffraction data. In general, macromolecular structures are not solved at sufficiently high resolution to be multifold over-determined (i.e. many more observations than refined parameters). In fact, macromolecular crystal structures are sometimes refined at resolutions (2.5 Å or less) where the number of observations is approximately equal to the number of refined parameters. The refinement process is therefore prone to two main problems:

- Convergence (is the starting model close enough to the correct answer?)
- Over-fitting (is the model too complex?)

The refinement process can be made more robust by incorporation of chemical information, i.e. restraining bond lengths and angles to values typically observed in high resolution structures [1]. The refinement of a crystallographic structure often suffers from the multiple minima problem; there are many local minima of the target function which the model can easily become trapped in. The use of molecular dynamics methods coupled with simulated annealing techniques greatly increases the chance of finding the global minimum, and it reduces the need for human intervention in the refinement process [2]. While this method increases the convergence of refinement it also increases the danger of overfitting poor or insufficient diffraction data.

The introduction of the free R-value allows an objective assessment of the validity of a refined structure in order to prevent overfitting [3]. However, the free R-value is an empirical measure, and as such provides little information of how to improve the model to prevent overfitting. A new refinement methodology has been recently introduced which extends the radius of convergence of refinement and decreases the likelihood of overfitting. The method uses a torsion angle representation of macromolecules where bond lengths and angles remain fixed during the refinement process [4]. These constraints on the model reduce the number of free parameters by about an order of magnitude. The removal of bond and angle vibrations allows higher temperatures to be employed to extend the searching power, overfitting to be reduced, and refinement to be carried out in less time than conventional simulated annealing methods. Possible solutions to the problems outlined above are:

- Convergence - higher temperatures (torsion angle dynamics)
- Less over-fitting - decreased degrees of freedom (torsion angle dynamics, NCS) and assessed by an objective measure (free R-value)

Torsion angle dynamics

The use of a torsion angle representation for macromolecules in the context of least squares crystallographic refinement was first proposed as long ago as 1971 [5]. However, the new method [4] used here is different in that the torsion angle representation is maintained within a molecular dynamics framework. The intramolecular motion of the molecules in the system is constrained to torsion angles; bond lengths and bond angles remain fixed. This serves to significantly decrease the number of degrees of freedom, by about a factor of ten, which reduces the chance of overfitting the data. In addition, the algorithm allows longer time steps and is stable at higher temperatures, thus permitting more extensive searching of conformational space. The conformation space explored is restrained to sensible macromolecular geometry throughout the refinement making the method more efficient. The application of torsion angle dynamics to the refinement of the GroEL structure used a constant temperature protocol (figure 1), although slow cooling simulated annealing protocols could also have been used. In addition, a purely repulsive nonbonded potential [6] was used to facilitate motion about torsion angles.

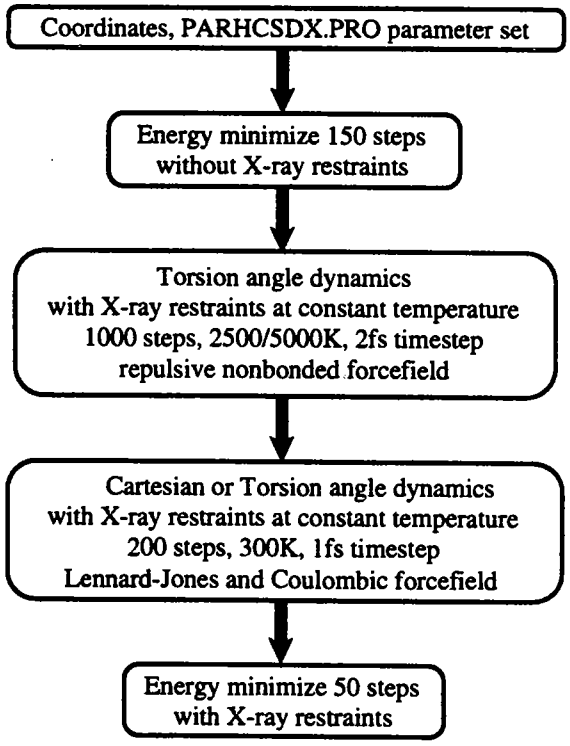


Figure 1: *Refinement protocol for X-PLOR using torsion angle dynamics (see [4] and [12] for further details).*

The Chaperonin GroEL

Little is known about the mechanism by which newly synthesized proteins fold in the cell. In 1973 Anfinsen had shown that polypeptides can fold under *in vitro* conditions spontaneously and reach their native state [7]. However, under non-ideal conditions such as found in the cell, folding of newly synthesized polypeptides is often inefficient due to competing off-pathway-reactions. Recently a group of specialized proteins, molecular chaperones, has been identified as playing an essential role in enabling polypeptides to reach their biologically active forms in a number of cellular compartments [8]. GroEL is the best characterized chaperonin. It is a 720 kDa complex consisting of two heptameric rings of 57 kDa subunits which are stacked back to back. Based on electron microscopic studies it has been suggested that nonnative polypeptides can at least in part be held within the central channel enclosed by the members of each ring [9]. The determination of the native GroEL crystal structure [10] in combination with extensive mutational analysis [11] has provided valuable information that allowed the assignment of functional properties to different regions of the structure. Unfortunately there were problems extending the refinement of the initial structure to all seven protomers, as indicated by high R-values (32.6% R-value and 36.8% free R-value).

Application of Torsion Angle Dynamics in Refinement

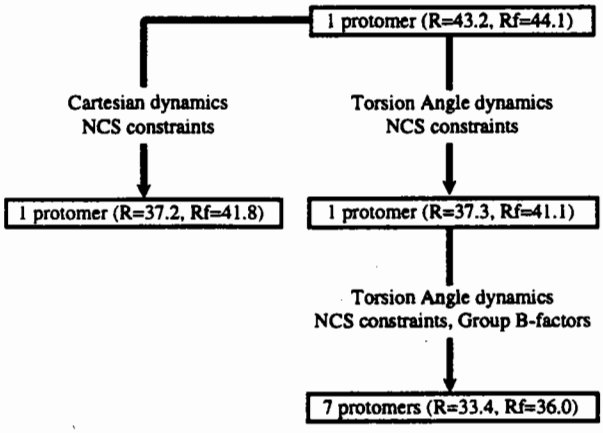


Figure 2: *Refinement of GroEL started using NCS constraints.*

For complete details of the initial structure determination see ref. [10] and for the refinement see ref. [12]. In the final refinement a conservative scheme was used, being careful not to overfit the model to the diffraction data. The refinement started from a partially refined model instead of the previously published final model [10] in order to minimize any possible model bias which may have been introduced. This model comprised residues 6 to 523 which had undergone one round of standard simulated annealing refinement using "strict" NCS constraints, where all protomers in the asymmetric unit were considered identical [13]. The first stages of refinement continued to use strict NCS constraints that impose strict seven-fold symmetry (figure 2).

These constraints were used until the refinement had converged (no significant further change in the R-values was observed), then NCS restraints instead of constraints were used [13]. Here the assumption was made that the protomers are essentially identical but deviations from seven-fold symmetry are possible (figure 3).

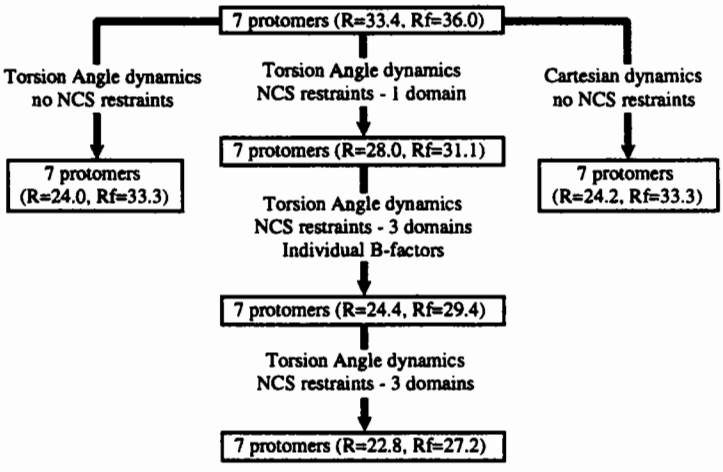


Figure 3: *Refinement of GroEL continued using NCS restraints.*

Estimation of the weight [13, 14] for the NCS restraints was important; too large a weight would prevent conformational variation among protomers, while too small a weight would allow too much variation and lead to overfitting of the model. During the intermediate steps of the refinement, different domains of the structure were given different weights reflecting greater mobility in some regions compared to others, which was suggested by very different average B-factors for each domain (figures 4). Initial weights were choosen to be high in the equatorial and intermediate domains (300 kcal mole^{-1} Å^{-2}) and six-fold lower in the apical domain (50 kcal mole^{-1} Å^{-2}) reflecting the differences in average B-factors for the domains. Several short refinements with progressively lower NCS restraint weights were carried out (figure 5). Decreasing the weights for the three domains resulted in an increase in the free R-value and a decrease in the conventional R-value; the diffraction data was being over-fitted. Therefore these NCS restraint weights for the three distinct topological domains were used until the final refinement cycle.

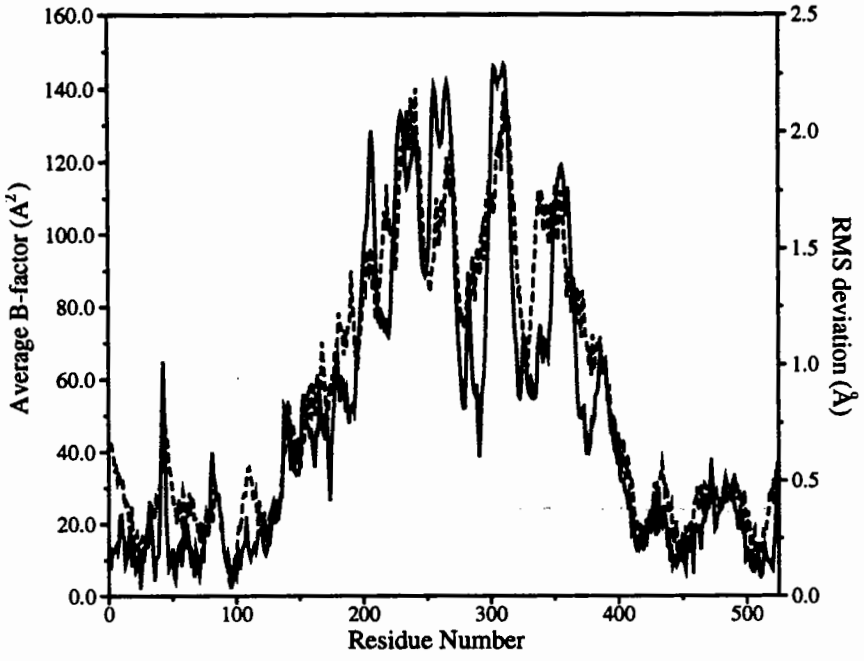


Figure 4: *Average B-factor (solid line) and RMS deviation (dashed line) for all seven protomers in the final model.*

Once the refinement of the model had converged the NCS restraint weights were redetermined. The restraint weights were increased and a short refinement cycle carried out as previously described, the free R-value was again used as an indicator of the most correct model. This indicated that the most simple and conservative restraint weights (300 kcal mole^{-1} Å^{-2} for each domain) gave a minimum of the free R-value. We conclude that the weaker restraints were required during initial refinement in order to escape local minima, but after convergence of refinement the tighter restraints were more appropriate.

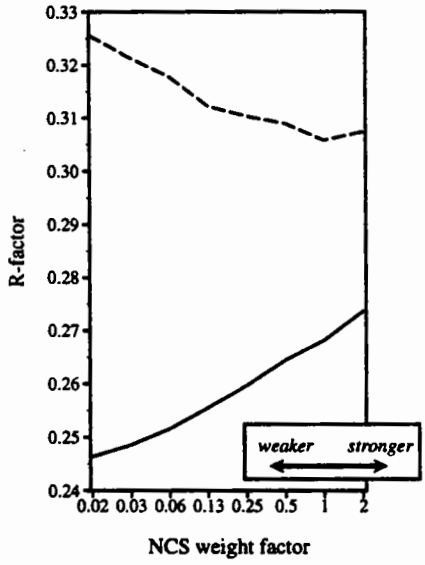


Figure 5: *Effect of NCS restraint weight on R-value (solid line) and free R-value (dashed line) during refinement. Weights were modified and a short cycle of torsion angle refinement performed.*

The improved radius of convergence of the torsion angle dynamics method [4] is demonstrated by comparison to the refinement carried out using the standard simulated annealing protocol. The very first cycle of refinement, using NCS constraints, was repeated using slow-cool simulated annealing and Cartesian molecular dynamics (figure 2). Visual analysis of the result showed that the standard slowcooling method was unable to correct some of the larger differences in the apical domain (figure 6), these errors would have been propagated through the subsequent refinement cycles.

The importance of NCS restraints during the refinement is indicated by the results if these restraints are removed (figure 7). It is clear that removing the restraints allows artificially high rms deviations between protomers. This indicates overfitting of the diffraction data - which can only be detected by the use of the free *R*-value. It is also of interest to note that there is little difference between the results for torsion angle based refinement and the standard simulated annealing method. This is because both methods perform equally well once they are close to the global minimum (the correct solution).

Figure 6: *Comparision of the results of torsion angle dynamics (solid) and standard simulating annealing protocol (dotted) for first refinement cycle using NCS constraints (the rms deviation between the structures is 0.8 Å for C_α atoms and 1.4 Å for all atoms).*

Conformational variability

The refined, atomic restrained B-factors for the structure are unusually high in the apical domains (a maximum of 150 Å^2) (figure 4). The decrease in free *R*-value upon B-factor refinement indicates that these B-values are an appropriate description of the diffraction data. A comparison between the superimposed seven subunits shows a high correlation between B-factors and RMS deviation between protomers and the average protomer structure (figures 4 and 8). The high absolute value of the refined B-factors in the apical domain are therefore a result of the rigid body motions between individual protomers in a ring (and presumably throughout the crystal lattice).

	Torsion NCS	Cartesian NCS	Torsion no NCS	Cartesian no NCS	Final Model NCS
R-value	0.279	0.277	0.240	0.242	0.228
Free *R*-value	0.309	0.305	0.333	0.333	0.272
<ΔNCS> (Å)	0.085	0.089	1.055	1.233	0.024
RMSD (Å) (*vs.* final model)	0.9/1.49	0.93/1.48	1.0/1.65	1.0/1.68	-/-

Figure 7: *Comparision of refinements with and without NCS restraints using either torsion angle dynamics or slow cooling simulated annealing with Cartesian molecular dynamics.*

In mediating the folding of newly synthesized proteins, GroEL has to bind a wide variety of substrates, spanning a spectrum of molecular mass and chemico-physical properties [16]. A protein that has such diverse substrates probably has a high degree of structural flexibility, defying a rigid lock-and-key model for interaction with polpeptide substrate. Although the GroEL complex consists of chemically identical subunits electron microscopic studies have clearly demonstrated that the native complex with either bound GroES and/or substrate is far from symmetric [17]. Electron microscopic studies [17, 18] and mutational analysis [11] have indicated that non-native proteins bind to the internal face of the apical domain. This domain contains regions exposed to the large central cavity (helices H8 and H9) which show high B-values (figures 4 and 8) which may be a result of conformational variability. This variability may be essential for interactions with such a diversity of substrates.

Figure 8: *Superposition of all seven protomers in the asymmetric unit showing the rigid body motion between the three domains (see [12] for details).*

Acknowledgments

We would like to thank A. Horwich, Z. Otwinowski and P. Sigler for discussions during this work. This work was supported by a grant (ASC 93181159) to AT. Brünger from the National Science Foundation. L. Rice is a predoctoral fellow with the Howard Hughes Medical Institute.

References

[1] Hendrickson, W.A. *Meth. Enzymol.* **115**, 252-270 (1985)

[2] Brünger, A.T., Kuriyan, J., and Karplus, M. *Science* **235**, 458-460 (1987)

[3] Brünger, A.T. *Nature* **355**, 472-474 (1992)

[4] Rice, L.M., and Brünger, A.T. *Proteins: Structure, Function, and Genetics* **19**, 277-290 (1994)

[5] Diamond, R. *Acta Cryst.* **A27** 436-452 (1971)

[6] Nilges, M., Clore, G.M., and Gronenborn, A.M. *FEBS Lett.* **229**, 317-324 (1988)

[7] Anfinsen, C.B. *Science* **181** 223-30 (1973)

[8] Ellis, J. *Nature* **328** 378-9 (1987)

[9] Braig, K., Simon, M., Furuya, F., Hainfeld, J.F., and Horwich, A.L. *PNAS* **90**, 3978-82 (1993)

[10] Braig, K., Otwinowski, Z., Hegde, R., Boisvert, D.C., Joachimiak, A., Horwich, A.L., and Sigler, P.B. *Nature* **371**, 578-586 (1994)

[11] Fenton, W.A., Kashi, Y., Furtak, K., and Horwich, A.L. *Nature* **371**, 614-619 (1994)

[12] Braig, K., Adams, P., and Brünger, A.T. *Nature Struct. Biol.* **2**, 1083-1093 (1995)

[13] Weis, W.I., Brünger, A.T., Skehel, J.J., and Wiley, D.C. *J. Mol. Biol.* **212**, 737-761 (1990)

[14] Brünger, A.T. X-PLOR Version 3.1 (Yale University, New Haven, CT, 1992)

[15] Brünger, A.T. *J. Mol. Biol.* **203**, 803-816 (1988)

[16] Viitanen, P.V., Gatenby, A.A., and Lorimer, G.H. *Protein Science* **1**, 363-369 (1992)

[17] Chen, S., Roseman, A.M., Hunter, A.S., Wood, S.P., Burston, S.G., Ranson, N.A., Clarke, A.R., and Saibil, H.R. *Nature* **371**, 261-264 (1994)

[18] Langer, T., Pfeifer, G., Martin, J., Baumeister W., and Hartl, F.U. *EMBO J.* **11**, 4757-4765 (1992)

Real Space Refinement as a tool for model building

T.J.Oldfield
Department of Chemistry,
University of York
York, YO1 5DD,
UK.

Introduction

Recent developments in recombinant DNA techniques, crystallisation protocols, X-ray data collection techniques and devices, and computing have led to a substantial increase in the speed and number of protein structure determinations in modern crystallographic laboratories. However, there still remains a number of key stages in the crystallographic process which limit the rate of structure determination. One of these is fitting electron density maps, either in the initial stages of tracing a chain to a new map, or in the manual rebuilding during refinement.

The talk discussed the refinement techniques available within the model building application of QUANTA96 (MSI), and how these refinement techniques can be used as tools for model building. Since denovo tracing of the first experimental phased maps and model building as part of refinement represent processes that cannot be carried out by conventional refinement methods, the algorithms developed as part of this application represent new approaches to this problem, and are probably not applicable to black box refinement.

This article describes four methods of refinement, as well as their implementation in real space using the torsion angles as variables. The advantages as well as approaches to overcome disadvantages associated with these methods is described.

Overview

Four methods of refinement have been developed that approach the problem of modeling in different ways. These are: gradient torsion angle refinement, Monte Carlo torsion angle refinement, grid search refinement, and geometric refinement as a modelling tool. All of these refinement methods have been developed within real space rather than reciprocal space which provides some advantages and one disadvantage.

In real space it is possible to easily refine local regions of the model, which is not true in reciprocal space. In particular, it is difficult to refine water atoms in reciprocal space because of the lack of restraints. As refinement is carried out in real space it is possible to provide a graphical representation of the progress of refinement to provide feedback to the crystallographer. It is therefore possible to provide tools that have a higher radius of convergence, but also have a risk of producing wrong structure, because the changes can be observed and aborted if obviously going wrong. The user remains in control throughout via the graphical user interface. This does mean that the algorithms developed must be fast, so great deal of attention has been made to provide extremely rapid calculations that can be used on moderately priced workstations. Since refinement is carried out in real space against an existing map, phase information is included in the calculation by default. This can be a disadvantage because phase information determined from the model is used (except for the first maps with experimental phase information), and so the process results in the inclusion of bias from previous refinement. However this is true for any model building into density of the form (n+1)Fo-(n)Fc and least squares minimisation that assumes that the true phase equals that calculated from the current model.

It must be stressed that the algorithms developed for the model building process should, if possible, have different targets, and at least different weighting, than those used in reciprocal space automated refinement programs (such as XPLOR, Prolsq, TNT etc), as no advantage would be obtained if the targets were the same. Real space torsion angle refinement provides these different targets.

Implicit in torsion angle refinement is that the bonds, non-bonds, angles, planes and chiral centres have restraints while the torsions angles can change as they have no explicit restraint. For gradient refinement the torsion angles are allowed to change so as to increase the fit to density, while for Monte Carlo refinement the torsions angles are assigned random values, and for grid refinement the torsion angles are set to all possible values between 0 and 360 degrees. An important exception to this is the omega torsion defined in proteins which is not freely rotatable due to the partial double bond. The omega torsion angle is therefore restrained with a low weight to either trans or cis conformation.

Gradient real space torsion angle refinement.

Bob Diamond first developed torsion angle refinement as a tool for protein crystallographic refinement.(Diamond R.) The refinement algorithm described here here is a completely new implementation of torsion angle refinement that allows anything from single atoms to entire molecules to be refined towards experimental data extremely rapidly. Torsion angle refinement has a problem associated with very high correlation along the polypeptide chain, so the algorithm developed here cycles between the use of full restraints, and single residue only restraints during the refinement. This breaks the correlation of the many torsion angles in a polypeptide chain while maintaining sensible structure.

The advantage of gradient torsion angle refinement over xyzb refinement is that the radius of convergence is much higher. A radius of convergence of around 2A is usually observed with torsion angle real space gradient refinement compared with 0.7Å with xyzb reciprocal space refinement.

Side chain and main chain grid search density fitting.

The principle of this refinement techniques is that if we can define at least one atom that has a correct position, then other atoms connected to this atom by rotatable bonds can be fitted by trying all combinations of rotomers using a grid search. This has been implemented for amino acids where the CA atom is defined as being in a correct position, and the side chain atoms are fitted by grid searching all the chi angles in the side chain with a discrimination of minima of 10 degrees and a precision of 2 degrees. For main chain fitting, two CA atoms are defined as correctly fitted, and the -N-C=O- peptide plane is fitted by rotation about the pseudo CA-CA bond, and the omega angle allowed to vary ±10 degrees. This method as implemented has two advantages for model building single residues. It has an "infinite" radius of convergence as defined by the length of the amino acid side chain, and the actual fitting is extremely quick (< 0.1 seconds/torsion - SGI R4000 indigo). The major problem is that it assumes the CA atom is correct. For amino acids, it is possible to get round this problem by providing a tool that allows the crystallographer to pick up the CA atom and move it in x/y/z space. As the fitting is very fast, the crystallographer sees new fitted conformations as the CA atom is moved so the result is an interactive side chain fitting tool Once the side chain atoms looks right, the new CA position can be accepted. Later stages of model building for

the majority of side chains can often be carried out by just using this tool, followed by regularisation.

The main chain and side chain fitting algorithm has been implemented as a powerful method of generating an all atom model from just a CA trace. Hence it is possible to use this refinement methods to take the CA trace generated by denovo fitting and build an all atom model at a rate of about 8 residues/second. The routine is extremely robust to poor CA atom positioning, and also produce very good results when fitting to the poor density associated with initial maps.

Monte Carlo refinement

Monte Carlo refinement provides a method of refinement when the conformation the main chain atoms is not obvious in a local region. This algorithm has been found to be an extremely powerful method for fitting loops and termini where there is only an indication of the main chain pathway in the density, but it is usually limited to maximum of seven/eight residues before the search becomes impossibly long.

The application generates an poly-alanine segment with a set of random phi psi angles. The poly-alanine segment is then attached at one end to the known structure, and if a loop is to be fitted, checked to see if it will span the loop region. If a terminal is to be searched, then this stage is not required. The application then checks the fit to density for this random polypeptide section, and selects/rejects this on the basis of whether it is a better fit than the worst of the best 10 solutions found to this point in time. Non-bond clashes are avoided by masking the electron density already occupied by known atoms in the neighbour-hood of the search region. The conformations are generated and fitted at a rate of more than 2000/second (SGI R4000 indigo) allowing millions of completely random conformations to be screened in 10's of minutes. The GUI is continually updated to show the best 10 solutions found up to that time so the crystallographer can abort the search at any time they observe a sensible solution to the problem. The range of density fits (by fit value, and colour) for these solutions is also shown so that convergence can be observed. As the search is so rapid, there is no improvement in the convergence of the algorithm from screening based on "predicted" conformations. The time taken to screen a conformation as a function of predicted expectation was found to take longer than the time taken to find another conformation, and this screening also "directed" the search which was not the original aim of the idea. On

acceptance of the loop conformation the application adds the side chain atoms using the real space grid search algorithm previously described.

The algorithm has also been implemented to provide a method for fitting ligands. This ligand fitting application first finds all the sites with significant density and without molecular coordinates. It then fits the ligand here by searching conformational space for the ligand using the Monte Carlo refinement algorithm based on the rotatable bonds in the ligand as well as refining the position and orientation of the ligand. Finally the application refines the ligand using the real space torsion angle refinement algorithm described earlier. This procedure has been to shown to work for ligand such as phosphate ions, polypeptides, polysacchoride to even poor density. It has an extremely high success rate (it has not yet produced a solution judged to be incorrect), and only takes minutes to complete of the refinement automatically with no user intervention.

Refinement of geometry.

The refinement of geometry, usually called regularisation, was included in the discussion of refinement techniques. The regulariser has been implemented to allow extremely fast convergence of the geometric terms describing the molecular structure. It will improve bonds, angles, chiral centres (pro-chiral centres), planes, and optionally, non- bonded contacts. As the routine is very fast it is possible to define a region of connected molecular structure for "active regularisation". As the crystallographer picks and moves any atom in this region, the algorithm maintains the expected geometry of the regularised region. Hence the effect is a pick up a drag facility that allows the crystallographer to just place, for example, CA atoms, while the remaining structure is dragged to compensate for the changes made. This has been found to be a much more natural tool to use than the manual manipulation of "zones" of residues using just the translation and rotation of the fragment being edited. If the non-bonding is not active then it is possible to edit more than 20 residues (SGI R4000 indigo), while the use of non-bonds will reduce this to about 10 residues maximum. This is usually far more than would normally be edited by this facility.

Since parameters can automatically be generated for ligands built within QUANTA96, it is possible to pick up a ligand molecule with several rotatable bonds and just pull it round the active site. The ligand can be pushed into the protein and will flex to take up a complementary

conformation. An obvious extension to this facility is to allow the crystallographer to define a "flexible" region in the protein, and then model the ligand here while the amino acids can move due to non-bonding interactions. A preliminary version of this active site modeling facility has been developed to test whether this facility will prove a useful modeling tool for ligand docking.

Ramachandran restraints

A further optional restraint of specific "Ramachandran" torsion angles can be applied when using the gradient refinement and the geometry refinement. This option allows the specification of minima for the torsion pairs know as phi and psi along the backbone chain of a protein. It is possible to set the minima to correspond to the conformations of alpha helix, beta sheet, and the nearest allowed point on the Ramachandran plot to each phi/psi torsion pair. These restraints are small and will push the conformation of a polypeptide chain very slowly towards the designated conformation once all the other restraints (bonds, angles, chiral centres, planes and non-bonds) are near their minima. The aim was to provide a modeling facility when there is only low resolution experimental data, or no experimental data at all. The use of these restraints are obviously open to abuse, and it is questionable whether derived torsional restraints should ever be used in refinement. Since this application provides a Ramachandran plot of the current molecule, the crystallographer will be trying to manually correct the phi/psi angle pairs during model building. Hence this facility will allow the inclusion of expected torsion angles at much faster rate than can be by manual modeling. Can the Ramachandran angles of a protein be used as an independent measure of the quality of said protein once the crystallographer has observed the Ramachandran plot and acted on this information ?

Summary

The refinement techniques described have been implemented in QUANTA96 as part of the large application known as X-Autofit and X-Build. The completely automated ligand fitting is also provided in QUANTA96 as the application X-Ligand.

Other facilities in this application include:-

1) Map masks (generation, interactive editing, void deletion).

2) Electron density skeletonisation (fast calculation, improved connectivity, smooth lines, auto main/side chain determination fast editing tools, and symmetry).
3) Semi automated CA tracing, Rule based fitting, helix/strand fitting, CA-trace refinement.
4) Automatic CA-trace to all atom model
5) Fuzzy logic sequence assignment
6) Rigid body refinement for molecules or any part of a molecule
7) 3D text editor, User annotation of molecules/maps, automatic annotation of validation errors
8) Symmetry generation at 10,000 atoms/bones points per second (NCS also supported)
9) Pickable Ramachandran and CA conformation plots.
10) Alternate conformations fully supported throughout.
11) Protein validation
12) X-solvate: automate solvent fitting
13) Automated water refinement

References

Diamond R. Acta Cryst. A27 (1971) 436-452
MSI : 9685 Scranton Road, San Diego, CA 92121-3752

Improved Structure Refinement Through Maximum Likelihood

Navraj S. Pannu* **and Randy J. Read†**
*Department of Mathematical Sciences
†Department of Medical Microbiology & Immunology
University of Alberta
Edmonton, Alberta T6G 2H7, Canada

When crystal structures of proteins or small molecules are used to address questions of scientific relevance, the accuracy and precision of the atomic coordinates are crucial. Accordingly, the model is generally improved by refining it to improve agreement with the observed diffraction data. Refinement of crystal structures is conventionally based on least-squares methods, but such procedures are handicapped, since conditions necessary for the use of the least-squares target are not satisfied. We propose that refinement should be based on maximum likelihood, and we have implemented two maximum likelihood targets in the program XPLOR. Preliminary tests with protein structures give dramatic results. Compared to least-squares, maximum likelihood refinement can achieve more than twice the improvement in average phase error. The resulting electron density maps are correspondingly clearer and suffer less from model bias.

Introduction

To obtain the most accurate possible crystal structure, one typically refines the atomic model to optimize its agreement with the observed diffraction data. However, the quality of the resulting model will depend on the validity of the target function that is optimized. We believe that, since the conventional least-squares target is poorly justified in this case, the refinement procedures are unduly handicapped. A maximum likelihood target is much better justified, and we show that it performs significantly better in macromolecular refinement.

The standard macromolecular refinement programs, PROLSQ(Konnert & Hendrickson, 1980), TNT(Tronrud, Ten Eyck & Matthews, 1987), XPLOR(Brünger, Kuriyan & Karplus, 1987) and GROMOS(Fujinaga, Gros & van Gunsteren, 1989), minimize a residual that is the weighted sum of squared deviations between the observed (F_O) and calculated (F_C) structure factor amplitudes, including a relative scale factor k, *i.e.* $\sum w(F_O\text{-}kF_C)^2$. The refinement programs differ primarily in minimization methods. Even though the atomic model is improved, problems arise because such a least-squares residual is poorly justified, especially early in refinement. As Silva and Rossmann (1985) have pointed out, what is minimized (ignoring weights) is the rms deviation between the model electron density and the density computed from Fourier coefficients $F_O\exp(i\alpha_C)$. This deviation can be minimized either by improving the model or by introducing systematic errors that obliterate differences from the model in the $F_O\exp(i\alpha_C)$ map. Since most macromolecular refinements have an unfavorable parameter to observation ratio, the data are typically overfit, which means that such systematic errors must be introduced.

The least-squares refinement target could be considered to arise from the principle of maximum likelihood, if the following assumptions held: the deviation between F_O and kF_C would have to be Gaussian, the mean deviation would have to be zero, and the standard deviation of the Gaussian would have to be independent of the parameters of the atomic model. This is not true, as shown below, because the errors have a (changing) phase component. For this reason, we should return to first principles and apply a maximum likelihood analysis to the problem of protein structure refinement, as we (Read, 1990) and Bricogne (Bricogne, 1991; Bricogne, 1993) have suggested. Garib Murshudov (this volume) is also working on an implementation of maximum likelihood. In another crystallographic context, that of multiple isomorphous replacement, a maximum likelihood treatment has also been applied with good results (Otwinowski, 1991).

Devising a likelihood function

The principle of maximum likelihood formalizes the idea that the quality of a model is judged by its consistency with the observations. To say that a model is consistent with an observation means that, if the model were correct, there would be a reasonably high probability of making an observation with that value. Taking all the relevant observations as a set, then, the probability of making the entire set of observations is an excellent measure of the quality of the model. If we assume that the observations are independent, the joint probability of making the set of observations is the product of the probabilities of making each independent observation. This joint probability is the likelihood function.

$$L = \prod_{hkl} p(F_O; F_C)$$

Since it is more convenient to work with sums than products, one typically works with the logarithm of the likelihood function. As well, the maximization problem can be turned into a minimization problem by multiplying by negative one. Therefore, defining $\Lambda = -\ln(L)$ gives the following:

$$\Lambda = -\sum_{hkl} \ln\left(p(F_O; F_C)\right)$$

In the case of crystallographic refinement, it is not strictly true that the diffraction observations are independent; if they were, direct methods and density modification would not work. There is doubtless much useful information to be gained by working with higher order collections of structure factors (Bricogne, 1993) but, as we will show, useful results are obtained even when independence is assumed.

To apply maximum likelihood, one must start from the probability of making a measurement, given the model, its errors, and the measurement errors. We have shown previously that various sources of random error in the model have equivalent effects on the probability distribution for the true structure factor, whether the errors are in atomic positions or temperature factors or whether there are missing or extra atoms; in each case the distribution of the true structure factor is well approximated by a Gaussian distribution centered on $D\mathbf{F}_C$ (Read, 1990). (D can be considered, roughly, as the fraction of the calculated structure factor that is correct.) In the case of acentric structure factors, which make up the bulk of data for macromolecular structures, the distribution $p_a(\mathbf{F}; \mathbf{F}_C)$ is a two-dimensional Gaussian in the complex plane, while for centric structure factors it is a one-dimensional Gaussian ($p_c(\mathbf{F}; \mathbf{F}_C)$).

$$p_a(\mathbf{F}; \mathbf{F}_C) = \frac{1}{\pi\varepsilon\sigma_\Delta^2} \exp\left(-\frac{(\mathbf{F} - D\mathbf{F}_C)^2}{\varepsilon\sigma_\Delta^2}\right)$$

$$p_c(\mathbf{F}; \mathbf{F}_C) = \frac{1}{\sqrt{2\pi\varepsilon\sigma_\Delta^2}} \exp\left(-\frac{(\mathbf{F} - D\mathbf{F}_C)^2}{2\varepsilon\sigma_\Delta^2}\right)$$

- ε = expected intensity factor
- $\sigma_\Delta^2 = \Sigma_N - D^2\Sigma_P$
- Σ_N = distribution parameter of the Wilson distribution for $\mathbf{F}$ (Wilson, 1949)
- Σ_P = distribution parameter of the Wilson distribution for $\mathbf{F}_C$

The probability of the true structure factor amplitude F, conditional on the calculated amplitude F_C, is obtained by integrating over the unknown phase difference to give the following:

$$p_a(F;F_C) = \frac{2F}{\varepsilon\sigma_\Delta^2}\exp\left(-\frac{F^2+D^2F_C^2}{\varepsilon\sigma_\Delta^2}\right)I_0\left(\frac{2FDF_C}{\varepsilon\sigma_\Delta^2}\right)$$

$$p_c(F;F_C) = \sqrt{\frac{2}{\pi\varepsilon\sigma_\Delta^2}}\exp\left(-\frac{F^2+D^2F_C^2}{2\varepsilon\sigma_\Delta^2}\right)\cosh\left(\frac{FDF_C}{\varepsilon\sigma_\Delta^2}\right)$$

where $\sigma_\Delta^2 = \Sigma_N - D^2\Sigma_P$

The probability distribution required to apply maximum likelihood, however, is the probability of the observed diffraction measurement given the calculated diffraction measurement, as the true value is not known. We have used two methods to approximate this distribution, differing in the level of approximation and in the distribution assumed for the observational error. In the first method (MLF1), the measurement error is assumed to be Gaussian in structure factor amplitudes, and a Gaussian approximation is made for the resultant combined distribution, expressed in terms of structure factor amplitudes. In the second method (MLF2), the measurement error is assumed to be Gaussian in the intensities, and a series representation of the resultant combined distribution is expressed in terms of structure factor amplitudes squared.

MLF1: An amplitude-based likelihood function

If the probability of the measurement error is assumed to be Gaussian in structure factor amplitudes, with standard deviation σ_F, the required probability distribution, $p(F_O;F_C)$, is obtained by convoluting $p(F;F_C)$ by $p(F_O - F)$.

$$p(F_O;F_C) = p(F;F_C) \otimes p(F_O - F)$$

As far as we have been able to determine, there is no analytical solution to this convolution for the important acentric case. (A series representation could be derived similarly to MLF2, as discussed below. We believe that it is better to use MLF2, if one goes to the effort of computing the series representation.) However, a good Gaussian approximation can be obtained using the first two central moments of the distribution. The expected value for the acentric case is given by the following:

$$\langle F_O \rangle = \tfrac{1}{2}\sqrt{\pi\varepsilon\sigma_\Delta^2}\ \Phi\left(-\tfrac{1}{2},1,-\frac{D^2F_C^2}{\varepsilon\sigma_\Delta^2}\right)$$

For the centric case, the expected value is:

$$\langle F_O \rangle = \sqrt{\frac{2\varepsilon\sigma_\Delta^2}{\pi}}\ \Phi\left(-\tfrac{1}{2},\tfrac{1}{2},-\frac{D^2F_C^2}{2\varepsilon\sigma_\Delta^2}\right)$$

In these expressions, $\Phi(a,b,z)$ is Kummer's confluent hypergeometric function, also denoted by ${}_1F_1(a,b,z)$. The variance for both the acentric and centric distributions is given by the following:

$$\sigma_{ML}^2 = \left\langle \left(F_O - \langle F_O \rangle\right)^2 \right\rangle$$

$$= \varepsilon\sigma_\Delta^2 + \sigma_F^2 + D^2F_C^2 - \langle F_O \rangle^2$$

As F_C increases, σ^2_{ML} tends towards $\varepsilon\sigma^2_\Delta + \sigma^2_F$ in the centric case, or $\frac{1}{2}\varepsilon\sigma^2_\Delta + \sigma^2_F$ in the acentric case because, in the limit, only the component of model error parallel to $\mathbf{F}_C$ contributes to the error in the amplitude. When these moments are used to construct a Gaussian approximation, the negative log likelihood function (Λ) is:

$$\Lambda = -\sum_{hkl} \ln\left(p(F_O;F_C)\right) = -\sum_{hkl} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2_{ML}}} \exp\left(-\frac{\left(F_O - \langle F_O\rangle\right)^2}{2\sigma^2_{ML}}\right)\right)$$

$$= \sum_{hkl} \tfrac{1}{2}\ln(2\pi) + \ln(\sigma_{ML}) + \frac{1}{2\sigma^2_{ML}}\left(F_O - \langle F_O\rangle\right)^2$$

The quality of the Gaussian approximation can be judged from a comparison of distributions shown in Figure 1.

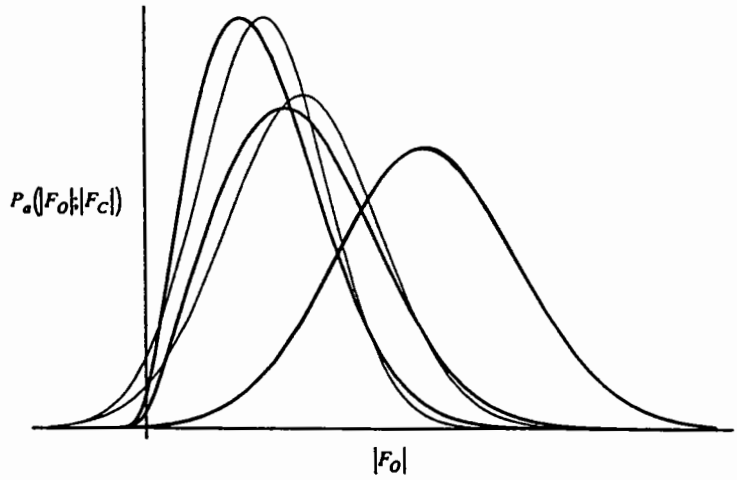


Figure 1. Comparison of the Gaussian approximation to $p(F_O;F_C)$ (thin lines) with the exact form determined by numerical integration (thick lines). Three pairs of curves are shown, corresponding to weak, average and strong reflections with D=0.7. This figure, Figure 2, and some of the mathematical derivations were made with the assistance of the program *Mathematica* (Wolfram, 1991).

If σ_{ML} is assumed to be relatively constant within a cycle of refinement, maximum likelihood refinement can be approximated as a modified least-squares refinement, in which the following target is minimized.

$$\text{WSSQ} = \sum_{hkl} \frac{1}{\sigma^2_{ML}}\left(F_O - \langle F_O\rangle\right)^2$$

This target can readily be implemented in any crystallographic refinement program that uses a least-squares target by weighting each term by $1/\sigma^2_{ML}$, replacing kF_C with $\langle F_O\rangle$ and replacing $\frac{\partial F_C}{\partial p}$ with $\frac{\partial\langle F_O\rangle}{\partial F_C}\frac{\partial F_C}{\partial p}$, where p is any parameter of the model being refined. The required derivative for the acentric case is given by

$$\frac{\partial\langle F_O\rangle}{\partial F_C} = \sqrt{\frac{\pi}{\varepsilon\sigma_\Delta^2}}\,\frac{D^2F_C}{2}\,\Phi\left(\tfrac{1}{2},2,-\frac{D^2F_C^2}{\varepsilon\sigma_\Delta^2}\right)$$

and for the centric case by

$$\frac{\partial\langle F_O\rangle}{\partial F_C} = \sqrt{\frac{2}{\pi\varepsilon\sigma_\Delta^2}}\,D^2F_C\,\Phi\left(\tfrac{1}{2},\tfrac{3}{2},-\frac{D^2F_C^2}{2\varepsilon\sigma_\Delta^2}\right)$$

Note that the F_C term eliminates the singularity in the derivatives that can arise in least-squares refinement on amplitudes (Schwarzenbach *et al.*, 1989).

MLF2: An intensity-based likelihood function

The second method that we use to derive the required probability distribution works in terms of structure factor amplitudes squared ($J = F^2$). Two advantages are attained by working in J instead of F. First, measurement errors frequently lead to a negative net intensity, which is reduced to negative J; when these legitimate observations are transformed to F, one has the choice of omitting them, replacing them with zero, or replacing them with a non-zero Bayesian posterior value (French & Wilson, 1978). By working in terms of J, this problem is avoided. Furthermore, a Gaussian measurement error is better justified in J than in F. In principle, maximum likelihood is insensitive to variable transformations such as from F to F^2 (Edwards, 1992). If MLF2 did not differ from MLF1 in the distribution assumed for the measurement error, the two likelihood functions would differ only in precision of the approximation.

The required probability distribution $\mathrm{p}(J_O;J_C)$ is derived by multiplying $\mathrm{p}(J;J_C)$ by a Gaussian probability for the measurement error with standard deviation σ_j, and integrating over the true structure factor amplitude squared (J).

$$\mathrm{p}(J_O;J_C) = \int_0^\infty \mathrm{p}(J_O;J)\times\mathrm{p}(J;J_C)\,dJ$$

A series representation of $\mathrm{p}(J_O;J_C)$ can be computed. For acentric reflections the distribution is

$$\mathrm{p}_a(J_O;J_C) = \frac{1}{\sqrt{2\pi\varepsilon\sigma_\Delta^2}}\exp\left(-\frac{J_O^2}{2\sigma_j^2}-\frac{D^2J_C}{\varepsilon\sigma_\Delta^2}\right)$$
$$\times\sum_{n=0}^{\infty}\left(\frac{D^2J_C\sigma_j}{\varepsilon^2\sigma_\Delta^4}\right)^n\frac{1}{n!}\exp\left(\frac{\left(\sigma_j^2-J_O\varepsilon\sigma_\Delta^2\right)^2}{4\varepsilon^2\sigma_\Delta^4\sigma_j^2}\right)\mathrm{D}_{-n-1}\left(\frac{\sigma_j^2-J_O\varepsilon\sigma_\Delta^2}{\varepsilon\sigma_\Delta^2\sigma_j}\right)$$

$\mathrm{D}_{-n-1}(x)$ is a parabolic cylinder function. For centric reflections,

$$\mathrm{p}_c(J_O;J_C) = \frac{1}{2\sqrt{\pi\varepsilon\sigma_\Delta^2\sigma_j}}\exp\left(-\frac{J_O^2}{2\sigma_j^2}-\frac{D^2J_C}{2\varepsilon\sigma_\Delta^2}\right)$$
$$\times\sum_{n=0}^{\infty}\left(\frac{D^2J_C\sigma_j}{2\varepsilon^2\sigma_\Delta^4}\right)^n\frac{1}{(2n)!!}\exp\left(\frac{\left(\sigma_j^2-2J_O\varepsilon\sigma_\Delta^2\right)^2}{16\varepsilon^2\sigma_\Delta^4\sigma_j^2}\right)\mathrm{D}_{-n-\frac{1}{2}}\left(\frac{\sigma_j^2-2J_O\varepsilon\sigma_\Delta^2}{2\varepsilon\sigma_\Delta^2\sigma_j}\right)$$

After eliminating terms that are constant within a cycle of refinement, the negative log likelihood for the acentric case is

$$\Lambda = \sum_{hkl} \ln\left(\varepsilon\sigma_\Delta^2\right) + \frac{D^2 J_C}{\varepsilon\sigma_\Delta^2} - \ln\left(\sum_{n=0}^{\infty} \left(\frac{D^2 J_C \sigma_j}{\varepsilon^2 \sigma_\Delta^4} \right)^n \frac{1}{n!} \exp\left(\frac{\left(\sigma_j^2 - J_O \varepsilon\sigma_\Delta^2\right)^2}{4\varepsilon^2 \sigma_\Delta^4 \sigma_j^2} \right) D_{-n-1}\left(\frac{\sigma_j^2 - J_O \varepsilon\sigma_\Delta^2}{\varepsilon\sigma_\Delta^2 \sigma_j} \right) \right)$$

and for centric reflections,

$$\Lambda = \sum_{hkl} \tfrac{1}{2} \ln\left(\varepsilon\sigma_\Delta^2\right) + \frac{D^2 J_C}{2\varepsilon\sigma_\Delta^2}$$
$$- \ln\left(\sum_{n=0}^{\infty} \left(\frac{D^2 J_C \sigma_j}{2\varepsilon^2 \sigma_\Delta^4} \right)^n \frac{1}{(2n)!!} \exp\left(\frac{\left(\sigma_j^2 - 2J_O \varepsilon\sigma_\Delta^2\right)^2}{16\varepsilon^2 \sigma_\Delta^4 \sigma_j^2} \right) D_{-n-\frac{1}{2}}\left(\frac{\sigma_j^2 - 2J_O \varepsilon\sigma_\Delta^2}{2\varepsilon\sigma_\Delta^2 \sigma_j} \right) \right)$$

Derivations and details of implementation of MLF1 and MLF2 can be found in Pannu & Read (submitted).

Some essential differences between least-squares and maximum likelihood refinement can be seen in a comparison (Figure 2) of the derivatives of the target functions, which lead to the atomic shifts in the refinement process.

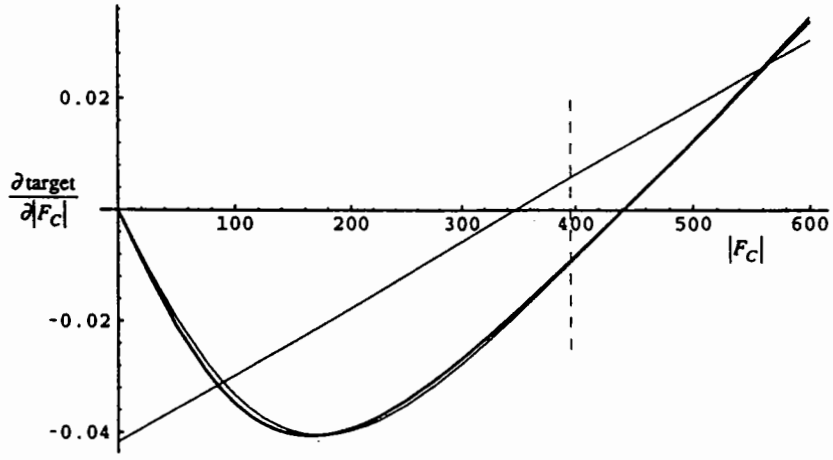


Figure 2. Comparison of the derivatives, with respect to F_C for one reflection, of the refinement targets for least-squares (thin line), MLF1 (thin curve) and MLF2 (thick curve), as a function of F_C. The example (the 2·12·17 reflection of the gTIM test case, discussed below) is chosen to illustrate the degree to which the least-squares and maximum likelihood targets can differ. In XPLOR, the derivative contributes to a force on each atom to move in a direction that will decrease the refinement target. At the start of refinement, F_C is 395.6 (indicated by the dashed vertical line); according to the least-squares target, atoms should move to decrease F_C while, according to the maximum likelihood targets, atoms should move in the opposite direction to increase F_C. Note that, if F_C were zero, the derivative for the maximum likelihood target would also be zero, reflecting the fact that the true phase would be completely uncertain and that a desired direction of shift could not be inferred.

Calibration of structure factor probabilities

The value of the likelihood function depends on the parameters of the atomic model. It also depends on the resolution-dependent parameters D and σ_Δ^2, which characterize the effect of model error on the structure factor probability distributions. (In fact, D and σ_Δ^2 are not independent and can each be computed from the single parameter σ_A (Read, 1990).) In principle, it would be best to optimise the likelihood function by adjusting all parameters simultaneously, including coordinates, B-factors *and* σ_A values. Unfortunately, a problem arises if the σ_A values are refined using the same data against which the model is refined: the poor parameter to observation ratio allows overfitting of the amplitudes, which results in an overestimation of σ_A and hence an underestimation of the errors in the calculated structure factors (Lunin & Urzhumtsev, 1984; Read, 1986). This leads to a positive feedback cycle in which the pressure to overfit becomes stronger. In our first attempt to implement maximum likelihood refinement, this problem was ignored. As the quality of the likelihood function depends strongly on the accuracy of σ_A estimates, the results were unimpressive.

The solution we have adopted is to use cross-validation data (a minority of data omitted from the refinement target) in an active way to provide unbiased estimates of structure factor accuracy. These data are normally used to compute R_{free}, an unbiased measure of refinement progress (Brünger, 1992). The use of cross-validation data to estimate σ_A is complicated, however, by the fact that stable estimates require 500 to 1000 reflections in each resolution shell, especially when the true value is low (Read, 1986). To overcome the problem of instability, we exploit the fact that σ_A varies smoothly with resolution. A simple correction, in which a penalty is applied when a σ_A value lies far from the line connecting its two neighbours, is sufficient (R.J. Read, unpublished).

A better solution would be to refine the σ_A values as parameters in the refinement, but to make allowance for the fact that they are biased estimates, in using them in the likelihood function. Lacking a theoretical basis for the correction for bias, however, this solution cannot yet be applied. We are currently studying the effect of refinement bias on the structure factor distributions, to lay the groundwork for such an improved treatment.

Test Refinements

The two maximum likelihood targets have been implemented in the program XPLOR (Brünger, Kuriyan & Karplus, 1987). Results from runs of the modified XPLOR on two test systems will be discussed here. In each test, the suggested weighting factor (WA) for the diffraction terms in the target, obtained by comparing the gradients from the diffraction and energy terms (Brünger, Karplus & Petsko, 1989), was divided by two.

Streptomyces griseus trypsin.

The crystal structure of *Streptomyces griseus* trypsin (Read & James, 1988) (SGT) was solved originally by molecular replacement, using the structure of bovine trypsin (Chambers & Stroud, 1979) (BT) as a search model. In order to compare the power of the maximum likelihood and least-squares targets in a case where the phase errors are known exactly, we used data calculated from SGT as error-free amplitudes F_O, and a superimposed model of BT as a starting structure. Since these two proteins share about 33% sequence identity, BT provides a relatively poor model that will only be capable of refining into a local minimum.

Data from infinity to 2.8Å resolution (5732 reflections, of which 578 were flagged as cross-validation data) were used for both refinements. (One often omits the low resolution data for least-squares refinement because of the problems caused by disordered solvent, but in this case there is no disordered solvent.) Table I shows the results obtained in the different refinements. While none of the refinements could achieve an accurate model, owing to the inadequacies of the starting model, the maximum likelihood targets gave more than twice as large an improvement in the average phase error. Note that, owing probably to the small number of reflections used in this case, R_{free} provides a weak indication of phase accuracy.

Table 1. Refinement statistics for SGT test case. The starting model (BT superimposed on SGT) was refined against calculated SGT data in two runs of XPLOR, identical except for the target function. In total, 420 cycles of energy minimization refinement were carried out.

	Start	Least-squares	MLF1	MLF2
R-factor	0.515	0.403	0.416	0.422
R_{free}	0.542	0.511	0.525	0.528
Mean phase error	62.2°	60.0°	56.7°	56.5°
Mean cos(phase error)	0.365	0.394	0.436	0.437

Trypanosoma brucei glycosomal triosephosphate isomerase.

At an intermediate stage in the refinement of the glycosomal triosephosphate isomerase (gTIM) from *Trypanosoma brucei* (Wierenga, Noble, Vriend, Nauche & Hol, 1991), data to a resolution of 1.83Å became available to replace the data to 2.4Å resolution that had been used to that point (Wierenga, Kalk & Hol, 1987). We tested the three refinement targets on this intermediate model, using the observed diffraction data (model and data kindly supplied by Dr. R.K. Wierenga). Of the 38812 observed amplitudes, 1014 were flagged randomly as cross-validation data. Because this is a real data set measured from a crystal with disordered solvent, data from infinity to 8Å resolution were omitted in the least-squares refinement, while they were used in both maximum likelihood refinements.

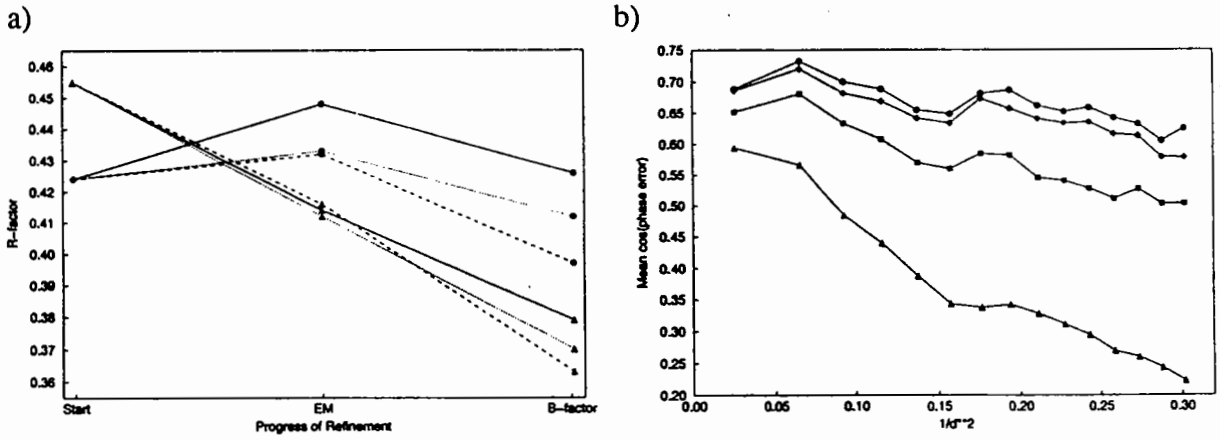


Figure 3. a) R-factors through test refinements of gTIM. The runs were identical except for the target function and the treatment of low resolution data; for the least-squares refinement, data from infinity to 8Å were omitted, while they were included for both maximum likelihood refinements. In each case, 250 cycles of energy minimization (EM) refinement were run, followed by 30 cycles of B-factor refinement. The solid lines indicate R-factors for the least-squares target, the dotted lines indicate R-factors for the MLF1 target, and the dashed lines indicate R-factors for the MLF2 target. R_{free} for least-squares; downward triangles, R_{free} values for the three different target functions are represented by circles, and R values are represented by triangles. b) Phase accuracy after gTIM test refinements. The phase accuracy is computed as the mean cosine of the phase error, which is comparable to the mean figure of merit. Triangles correspond to the starting model, squares to the least-squares model, diamonds to the MLF1 model, and circles to the MLF2 model.

As shown in Figure 3, both maximum likelihood target functions achieved a substantially greater improvement in the model, measured by both R_{free} and phase differences with the final model. As one might expect from the increased precision of the

approximation, the MLF2 target gives significantly better results than MLF1. This improvement is achieved for a modest computational cost. Compared to an equivalent refinement with the least-squares target, the MLF1 target requires about 1% more computer time, while the MLF2 target requires about 10% more computer time.

Conclusions

While the current implementations of maximum likelihood refinement already provide significant benefits, a number of improvements can be foreseen. First, the algorithm for the computation of σ_A does not take into account measurement errors. Either of the likelihood functions derived here, MLF1 or MLF2, can be used to compute σ_A values that take into account measurement errors, and these modified likelihood functions will be implemented in the SIGMAA algorithm. As is clear from the variance term in the Gaussian approximation MLF1, observational error has little influence on the likelihood function unless the model is quite accurate. Nonetheless, it will become significant at the end of refinement and a proper treatment will be important to obtain an optimal final model.

Arbitrary relative weights between diffraction and geometry terms should not be required, in principle, if each is introduced to maximum likelihood through the appropriate probability distributions. However, we have found that some overweighting of the diffraction terms, relative to the theoretical value, is needed to achieve convergence. This may be necessary because the inevitable overfitting of the diffraction amplitudes alters the distribution $p(\mathbf{F};\mathbf{F}_C)$. In various tests, the comparison of gradients has led to weights that are increased by factors between 4 and 50, with higher weights being required for less refined models at lower resolution. Further tests will be required to decide whether these relative weights are optimal.

Finally, the maximum likelihood approach allows one to include, in a sensible way, any combination of information. We believe that considerable scope for improvement exists in the simultaneous refinement of structures, for instance, native with liganded, or native with heavy atom derivatives. In such a refinement, all observations would be fit simultaneously, using models that are restrained to resemble one another to a degree required by the relationships among the measured sets of structure factors.

Acknowledgments

This work might not have been carried out if not for the opportunity provided by Dr. Rik K. Wierenga, who was the host for RJR as a summer visitor to the EMBL, Heidelberg in May 1993, when the first steps to implementation were taken. The manuscript was prepared while RJR was an academic visitor at the MRC Laboratory of Molecular Biology in Cambridge. Financial support was provided by the Alberta Heritage Foundation for Medical Research, the Medical Research Council of Canada and an International Research Scholar award to RJR from the Howard Hughes Medical Institute. This contribution has been adapted from a manuscript submitted for publication in *Acta Crystallographica.*

References

Bricogne, G. (1991). Acta Cryst., A47, 803-829.

Bricogne, G. (1993). Acta Cryst., D49, 37-60.

Brünger, A.T., Karplus, M. & Petsko, G.A. (1989). Acta Cryst., A45, 50-61.

Brünger, A.T. (1992). Nature, 355, 472-474.

Brünger, A.T., Kuriyan, J. & Karplus, M. (1987). Science, 235, 458-460.

Chambers, J.L. & Stroud, R.M. (1979). Acta Cryst., B35, 1861-1874.

Edwards, A.W.F. (1992). Likelihood. Baltimore: The Johns Hopkins University Press.

French, S. & Wilson, K. (1978). Acta Cryst., A34, 517-525.

Fujinaga, M., Gros, P. & van Gunsteren, W.F. (1989). J.Appl.Cryst., 22, 1-8.

Konnert, J.H. & Hendrickson, W.A. (1980). Acta Cryst., A36, 344-350.

Lunin, V.Y. & Urzhumtsev, A.G. (1984). Acta Cryst., A40, 269-277.

Otwinowski, Z. (1991). Isomorphous replacement and anomalous scattering: Proceedings of the CCP4 Study Weekend 25-26 January 1991, edited by W. Wolf, P.R. Evans & A.G.W. Leslie, pp. 80-86. Daresbury, UK: Science and Engineering Research Council.

Read, R.J. (1986). Acta Cryst., A42, 140-149.

Read, R.J. & James, M.N.G. (1988). J.Mol.Biol., 200, 523-551.

Read, R.J. (1990). Acta Cryst., A46, 900-912.

Schwarzenbach, D., Abrahams, S.C., Flack, H.D., Gonschorek, W., Hahn, Th., Huml, K., Marsh, R.E., Prince, E., Robertson, B.E., Rollett, J.S. & Wilson, A.J.C. (1989). Acta Cryst., A45, 63-75.

Silva, A.M. & Rossmann, M.G. (1985). Acta Cryst., B41, 147-157.

Tronrud, D.E., Ten Eyck, L.F. & Matthews, B.W. (1987). Acta Cryst., A43, 489-501.

Wierenga, R.K., Kalk, K.H. & Hol, W.G.J. (1987). J.Mol.Biol., 198, 109-121.

Wierenga, R.K., Noble, M.E.M., Vriend, G., Nauche, S. & Hol, W.G.J. (1991). J.Mol.Biol., 220, 995-1015.

Wilson, A.J.C. (1949). Acta Cryst., 2, 318-321.

Wolfram, S. (1991). Mathematica: A system for doing mathematics by computer. 2nd ed. Reading,MA: Addison-Wesley.

Maximum-Likelihood Refinement of Incomplete Models with BUSTER + TNT.

Gérard Bricogne[1,2] and John Irwin[2]

[1] MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England,
[2] LURE, Bâtiment 209D, Université Paris-Sud, 91405 Orsay, France.

0. Introduction.

The Bayesian viewpoint has long suggested that structure refinement should be carried out by maximising the log-likelihood gain LLG rather than by minimising the conventional least-squares residual [1,2,3] : only the maximum-likelihood (ML) method can take into account the uncertainty of the phases associated to model incompleteness and imperfection by suitably downweighting the corresponding amplitude constraints. It was predicted [3,4] that ML refinement would allow the refinement of an incomplete model by using the structure factor statistics of randomly distributed scatterers to represent the effects of the missing atoms, in such a way that the latter would not be wiped out; and that the final LLG gradient map would then provide indications about the location of these missing atoms. As will be shown below, these predictions have now been confirmed by actual tests.

The section ends with a discussion of the two main concerns at the moment in the fields of structure refinement and validation where Bayesian methods have much to offer, namely (1) getting better reliability indicators for the final results of structure refinement, and (2) ensuring that these indicators are effectively optimised during refinement.

1. Shortcomings of least-squares, and current remedies.

In small-molecule studies, where the data to parameter ratio is huge, the error-covariance matrix gives a wealth of accuracy estimates, which can be cast into more readable form (e.g. TLS analysis of thermal parameters). The Luzzati error model and plot [5] can also be used to estimate final positional accuracy.

With macromolecules, however, the data to parameter ratio is never huge, even with restraints. In these circumstances least-squares (LS) structure refinement can produce overfitting artefacts by moving faster towards agreement with moduli than towards correctness of the phases, because its shift directions assume the current model phases to be error-free constants. R-factors and Luzzati plots then become misleading. Furthermore, when the model is very incomplete, density for the missing part tends to disappear rather than improve during LS refinement.

The current remedies rely on cross-validation (CV) [6] as a powerful device for detecting the onset of overfitting. It is based on the simple notion that overfitting amounts to fitting "noise" rather than "signal" in the data, which causes a loss of predictive power towards data not used in the fit. It must be borne in mind, however, that using CV in this way as a stopping criterion in a LS refinement only guarantees optimality *along the least-squares path*: it does not guarantee that the solution reached is optimal in a global sense. Assessing the accuracy of the results in the absence of an error-covariance matrix is not straightforward; the safest method available at present for estimating r.m.s. coordinate error seems to be a Luzzati plot from cross-validated σ_A values .[7]

2. *Seeking a more radical cure.*

Improvements on the current state of the art (least-squares refinement with cross-validation by R^{free}) seem desirable in two related directions, in both of which the fundamental techniques of structure factor statistics occupy a central position.

Firstly, since the LS path is deflected towards a premature fit to the moduli by excessive confidence in the current phases, it is natural to think of a feed-back mechanism whereby the current estimate of model error would be converted into a representation of the uncertainty on the phases, so that the latter could be used with more caution. Exercising this caution, however, necessarily involves altering weights (variances), which is not allowed within the least-squares method: the latter must therefore be abandoned in favour of the maximum likelihood (ML) method (see a similar argument about the treatment of non-isomorphism in §2.4.1 of [8]).

Secondly, since in the ML method the model now parametrises its own uncertainty, the question arises of choosing an adequate error model. It will be argued that the Luzzati error model is not suited to the heavily-restrained macromolecular setting, and that a new class of statistical models is required.

3. *Maximum Likelihood vs. Least-Squares.*

ML refinement offers an attractive generalisation over LS [1,2,3,4] by allowing the refinement of parameters which modulate the variances of the model structure factors: the latter are no longer handled as values but as probability distributions, in which variances and covariances can represent both model imperfection and model incompleteness. According to the standard protocol outlined in [1] the probability distributions for model structure factors are integrated over the phase to yield predicted distributions of model amplitudes; substituting the observed values of these amplitudes then yields the likelihood Λ of the model. All parameters can then be refined by maximisation of L or of $L = \log \Lambda$. The error covariance matrix is the final Hessian of L, if it can be calculated. It should be recalled that ML estimation is only an approximation to Bayesian estimation, and that the full force of the latter should be invoked whenever the maximum of L is not so pronounced as to dominate over prior probability in the application of Bayes's theorem.

4. *A prototype of ML refinement using BUSTER and TNT.*

To ascertain the impact of taking phase uncertainty into account on the path followed during structure refinement, we have used BUSTER [4] and TNT [9] on a test data set for crambin [10] suffering from both model imperfection and model incompleteness, and compared the results of LS and ML refinements from these data.

Model incompleteness resulted from taking only residues 1-27 (60% of the atoms) as the fragment to be refined; the remaining 40% (residues 28-46) was modelled through a non-uniform distribution for the missing atoms, defined by a mask for that region which had been extensively smoothed then blurred by a B-factor of 250. The expectation values and variances for the structure factor contributions from this pool of random atoms were calculated within BUSTER according to the equations in §2.1.0 of [4].

Model imperfection was introduced by heating fragment 1-27 to 1000°K then regularising it, using XPLOR [11], thereby creating positional errors with an r.m.s. value of about 1.0Å. This imperfection was treated statistically through a Luzzati model parametrised by a refinable "imperfection B factor" B^{impf}, similar to the quantity B^{glo} used in the parametrisation of non-isomorphism in heavy-atom derivatives (see §2.4.1 of [7]). This B^{impf} intervenes in the calculation

of expectation values $\langle F^{impf}(\mathbf{h}) \rangle$ and variance parameters $\sigma_2^{impf}(\mathbf{h})$ for the structure factor contributions from the imperfect fragment according to:

$$\langle F^{impf}(\mathbf{h}) \rangle = D(\mathbf{h}) \times F^{frag}(\mathbf{h}) \quad (1)$$

$$\sigma_2^{impf}(\mathbf{h}) = \left(1 - D(\mathbf{h})^2\right) \times \langle \left|F^{frag}(\mathbf{h})\right|^2 \rangle_{d_h^*} \quad (2)$$

where

$$D(\mathbf{h}) = \exp\left[-\frac{1}{4} B^{impf} (d_h^*)^2 \right] \quad (3).$$

The expectation values for the imperfect fragment and random atoms contributions, and the variances or covariances caused by imperfection and incompleteness, are added and used as arguments of elliptic Rice likelihood functions [12], in combination with any experimental phase information which may be available.

Refinement was carried out against 1.5Å synthetic data calculated from the correct whole crambin structure, without solvent, with 3% r.m.s. noise added. The reference LS refinement was performed using TNT in the conventional way. The ML refinement proceeded as follows. At each cycle BUSTER refined the values of overall scale and B factors and of B^{impf} by maximum-likelihood, and calculated the value, gradient and Hessian of the log-likelihood gain L with respect to the quantities $F^{frag}(\mathbf{h})$. This "osculating LS" approximation to L was passed on to TNT where is was used to generate parameter gradients (AGARWAL command) and curvatures, and to carry out one cycle of positional refinement on the fragment structure.

In these conditions ML refinement clearly outperformed LS refinement, giving a mean-square distance to the correct positions of 0.176 (ML) instead of 0.415 (LS). Examination of histograms of positional errors showed that, apart from a small number of outliers corresponding to model atoms near the boundary with the missing region, *the ML fit is much tighter than the LS fit.*

Visualisation of the time course of the refinement showed, as anticipated, that not only the end point but *the entire path of the refinement is altered* by switching from LS to ML. This may be understood by noting that the contribution to structure factor variances from model imperfection, given by eq. (2) above, increases sharply with resolution, so that high-resolution contributions to the gradient maps are filtered out in the early stages then gradually switched on as refinement proceeds. This feature leads to considerable *increases in the radius of convergence* of the refinement.

Furthermore the ML method produced a final LLG gradient map displaying highly significant, correct connected features for the missing part (40%) of the molecule, while the final LS difference map showed no such features (see Figs. 1 and 2). This *enhances the possibilities of bootstrapping* from an otherwise unpromising molecular replacement starting point to a complete structure. Essentially the same behaviour was observed at 2.0Å resolution, and with experimental rather than calculated data.

Other prototypes for ML structure refinement have been built and tested by Read [13] (using XPLOR and an intensity-based LLG) and by Morshudov [14] (using PROLSQ [15] and the Rice LLG). The BUSTER+TNT prototype has the advantage of being able to use external phase information by means of the elliptic Rice function [12], as well as prior information about non-uniformity in the distribution of the missing atoms in incomplete models. It also allows the ML refinement of an incomplete model to be carried out in conjunction with phase permutation or phase refinement for those strong amplitudes which are most poorly phased by that model, i.e. have the largest renormalised $|E|$'s . The latter feature establishes a seamless continuity between the middle game of structure determination and the end game.

5. *Limits of the Luzzati model.*

In the test calculations reported above, examination of partially converged models during or after refinement at lower resolution leads to the obvious conclusion that *questions of accuracy concerning the results of macromolecular refinement at medium resolution are fundamentally different from the same questions posed and studied for small molecules at high or very high resolution.* In the latter case it is reasonable to treat the model errors on the positions of different atoms as statistically independent and thus to use Luzzati's treatment for the errors they induce on the structure factors. Macromolecular refinement, on the other hand, is so heavily restrained that the model positional errors at any stage are highly correlated. This affects such crucial quantities as the effective number of degrees of freedom in the error statistics, and the magnitude of the uncertainty along each of these degrees of freedom. The Luzzati model is then inappropriate as a means of relating positional error statistics to structure factor statistics, and hence as a means of constructing a good likelihood function for ML refinement.

6. *A new class of error models for macromolecular structures.*

In a macromolecular refinement, model positional errors will be correlated through "regular perturbations" of a restrained macromolecular structure, i.e. perturbations compatible with the restraints which propagate positional errors between atoms or groups of atoms. New error models are required for deriving the structure factor statistics associated to random regular perturbations.

This may be illustrated by a simple physical analogy, for the physical aspects of which the reader is referred to [16]. The assumption of statistically independent random perturbations of atomic positions underlies not only the Luzzati model in structure factor statistics, but also the Einstein model of thermal motion in crystals and the Debye model of thermal effects on scattering. What is now needed in the field of structure factor statistics is the equivalent of the Born & von Kármán lattice-dynamical model of thermal motion, and of the use of these lattice normal modes in the parametrisation of anisotropic B factors and of thermal diffuse scattering.

An attractive possibility – if computer limitations can be ignored – would be to use the softest lattice 'normal modes' with wave vector $\mathbf{q}=\mathbf{0}$ from the Hessian matrix of the restraint function and parametrise the joint positional uncertainty model in terms of the variances of normal coordinates along these modes. This correlated positional error model could then be converted into a parametrised joint probability distribution of complex structure factors, then of amplitudes, which would yield the best likelihood function for refining both the structural model parameters and the mean-square normal coordinates describing the errors. At the end of the refinement, this error model would embody the description of the accuracy of the refinement results.

7. *Maximum-likelihood refinement for non-macromolecular problems.*

The two main sources of bias in macromolecular LS refinement results, namely the low observation-to-parameter ratio and the inadequate treatment of phase uncertainty, are also present in other fields of crystallography, in particular in Rietveld refinements of powder structures [17] and in multipole refinements of accurate electron densities [18-21]. In the powder case the notion of phase can be generalised to that of a hyperphase,[2] the loss of hyperphase information comprising both that which results from the overlap of different Bragg reflexions and from the ordinary loss of phase for these Bragg reflexions. In this instance, hyperphase-mediated bias is even more pernicious than the phase-mediated bias considered above and is the likely cause of numerous recently diagnosed pathologies in test Rietveld LS refinements. The probability distributions and likelihood functions for powder data derived in [2] will enable the incorporation of hyperphase uncertainty into the refinement and yield a maximum-likelihood Rietveld method which can be expected to cure the observed biases of the current LS method.

8. *Validation and error models.*

The use of cross-validation in the choice of refinable model parameters and in the validation of refinement results [7] has so far been based on the conventional crystallographic R-factor, which is not a particularly optimal criterion from the statistical point of view. In particular, concern has arisen about possible dangers of its use in the presence of non-crystallographic symmetries, since data belonging to the test set may happen to be strongly correlated to data which are being fitted, thus creating misleadingly low values for the free R-factor. The problem is clearly that the R-factor definition makes no reference to any predictable variability in statistical dispersion from one data item to another, nor to expected patterns of correlation in this dispersion.

The Bayesian viewpoint gives an unequivocal answer to this dilemma. Retaining the idea of cross-validation as a measure of the predictive power of a statistical model towards yet unseen data (already present in the scheme proposed in §8.1 of [22]) it leads naturally to suggesting that the free R-factor be replaced by the *free log-likelihood gain* L^{free} calculated over the same test data set. This viewpoint is none other than that formulated in [1] and [4] and does require that the predictions from the fit of the actively used data be couched in terms of a conditional probability distribution for the test data, from which the free LLG (e.g. from the model at the preceding cycle) can be calculated by the standard procedure.

Since the strong correlations between amplitudes created by non-crystallographic symmetries can be taken into account in the calculation of likelihoods [3], the use of L^{free} should be immune to the problems encountered by R^{free} in this case. In less problematic cases L^{free} can still be expected to perform better, in view of the Neyman-Pearson optimality property [1], provided the likelihood functions used are capable of correctly representing the state of knowledge (or uncertainty) prevailing at each stage. In the refinement context this adds to the urgency of the developments outlined in §7.

9. *Conclusion.*

It has been shown that maximum-likelihood structure refinement, long advocated by the first author, is greatly superior to conventional least-squares refinement by virtue of its ability to deal correctly with the phase uncertainties introduced by model imperfection and incompleteness. The end results are more accurate, the radius of convergence is increased, and the final log-likelihood gradient map gives useful indications as to the location of missing atoms.

The increase in radius of convergence may rapidly overturn the present reliance on simulated annealing [11] as a means of getting out of local least-squares minima: the automatic "blurring" of the LLG gradient maps in the early stages of the refinement will largely suppress such spurious minima. It is thus conceivable that simulated annealing might be dispensed with altogether in the future, any possible bifurcation being handled through phase permutation techniques.

The optimal performance of ML refinement will depend crucially on the design and implementation of better statistical error models in real space as the basis for better likelihood functions in structure factor space. Much remains to be done in this area, as well as in making better use of off-diagonal interactions during the likelihood-maximisation process itself.

Acknowledgements.

We wish to thank Dale Tronrud and Lynn Ten Eyck for advice on the use of TNT and for making minor changes at our request. We are grateful to Angela Loh and to Digital Equipment Corp. for the loan of generous amounts of state-of-the-art computer equipment.

This research was supported in part by a Tage Erlander Guest Professorship from the Swedish Natural Sciences Research Council (NFR) in 1992-93, and by an International Research Scholars award from the Howard Hughes Medical Institute since 1993.

Figure 1. The log-likelihood gradient map at the end of LS refinement. The missing structure is drawn for reference. There is very little reliable information to help complete the refined partial structure.

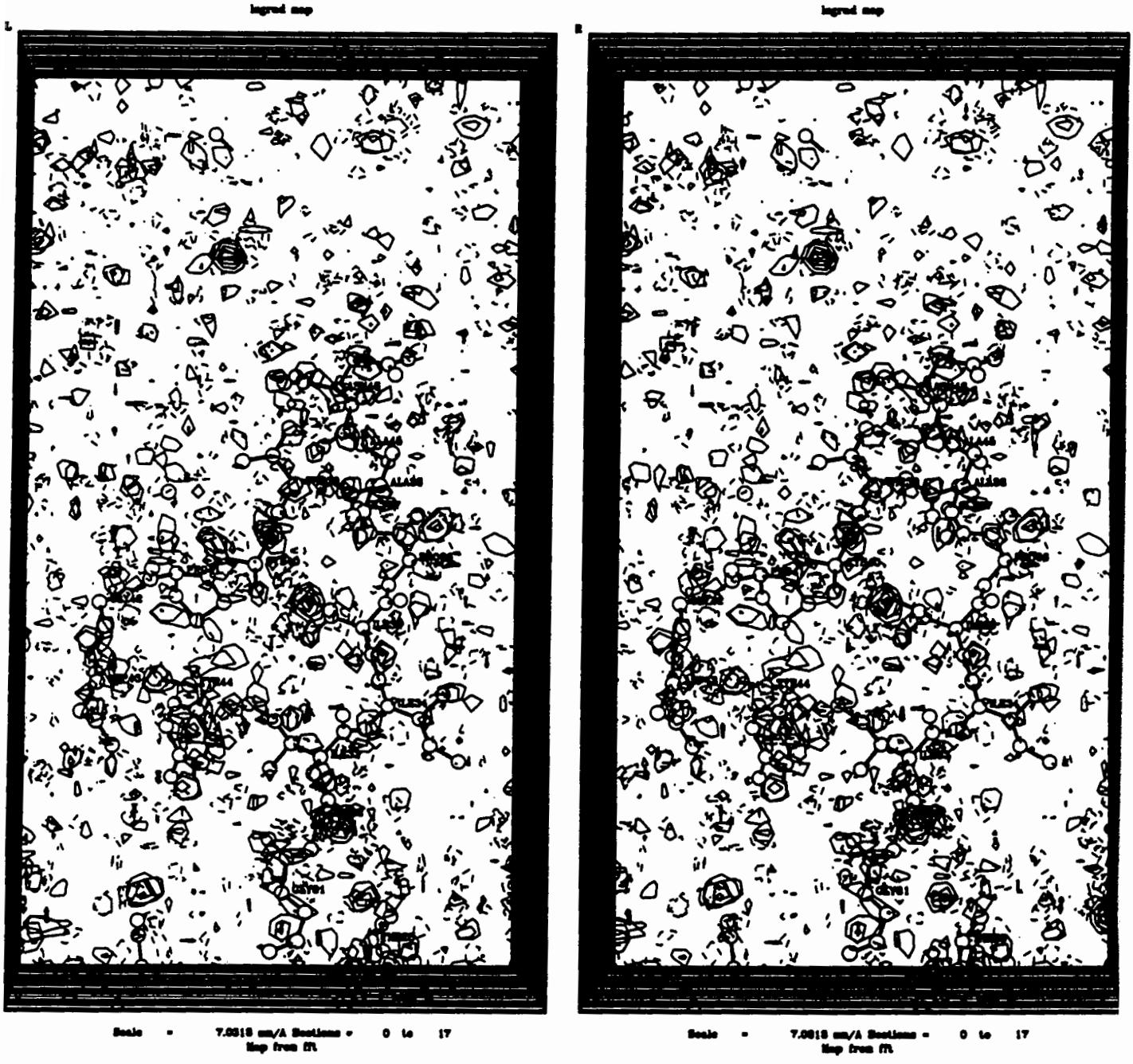
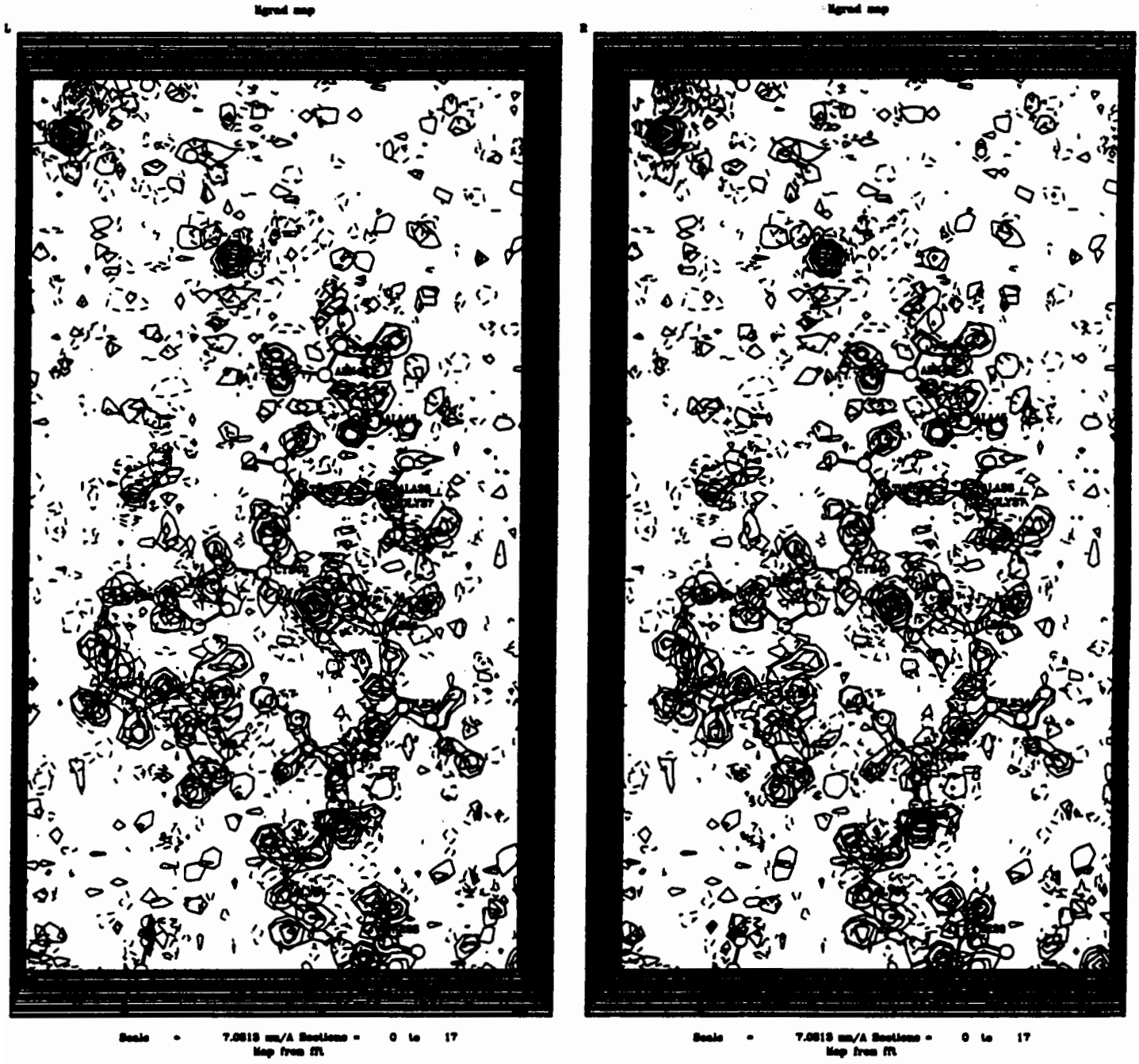


Figure 2. The log-likelihood gradient map at the end of ML refinement. There is considerably more reliable information to help complete the partial structure after it has been refined.



References.

[1] G. Bricogne, *Acta Cryst.* A**44**, 517-545 (1988).

[2] G. Bricogne, *Acta Cryst.* A**47**, 803-829 (1991).

[3] G. Bricogne, In *The Molecular Replacement Method. Proceedings of the CCP4 Study Weekend 31 January – 1st February 1992*, edited by W. Wolf, E.J. Dodson and S. Gover, 62-75, Daresbury Laboratory, Warrington (1992).

[4] G. Bricogne, *Acta Cryst.* D**49**, 37-60 (1993).

[5] V. Luzzati, *Acta Cryst.* **5**, 802-810 (1952).

[6] A.T. Brünger, *Acta Cryst.* D**49**, 24-36 (1993).

[7] A.T. Brünger, *Methods in Enzymology* **276** (in the press).

[8] E. de La Fortelle and G. Bricogne, *Methods in Enzymology* **276** (in the press).

[9] D.E. Tronrud, L.F. Ten Eyck, and B.W. Matthews, *Acta Cryst.* A**43**, 489-501 (1987).

[10] W.A. Hendrickson and M.M. Teeter, *Nature* **290**, 107-109.

[11] A.T. Brünger, J. Kuriyan, and M. Karplus, *Science* **235**, 458-460 (1987).

[12] G. Bricogne, *Methods in Enzymology* **276** (in the press).

[13] R.J. Read, (1996), To appear in *Macromolecular Refinement*, edited by M. Moore and E.J. Dodson, Daresbury Laboratory, Warrington (1996).

[14] G. Morshudov, To appear in *Macromolecular Refinement*, edited by M. Moore and E.J. Dodson, Daresbury Laboratory, Warrington (1996).

[15] J.H. Konnert and W.A. Hendrickson, *Acta Cryst.* A**36**, 344-349 (1980).

[16] B.T.M. Willis and A.W. Pryor, *Thermal Vibrations in Crystallography*, Cambridge University Press, Cambridge (1975).

[17] H.M. Rietveld, *Acta Cryst.* **22**, 151-152 (1967) ; *J. Appl. Cryst.* **2**, 65-71 (1969).

[18] F.L. Hirshfeld, *Acta Cryst.* B**27**, 769-781 (1971).

[19] N.K. Hansen and P. Coppens, *Acta Cryst.* A**34**, 909-921 (1978).

[20] M.A. Spackman and R.F. Stewart, In *Methods and Applications in Crystallographic Computing*, edited by S.R. Hall and T. Ashida, 302-320, Clarendon Press, Oxford (1984).

[21] B.M. Craven, H.P. Weber, and X.M. He, *The POP Refinement Procedure.* Technical Report, Dept. of Crystallogr., University of Pittsburgh (1987).

[22] G. Bricogne, *Acta Cryst.* A**40**, 410-445 (1984).

Application of Maximum likelihood Methods for Macromolecular refinement

Garib N. Murshudov, Eleanor J. Dodson
Chemistry Department, University of York, Heslington, York, U.K.

Alexei A. Vagin
UCMB-ULB, Free University of Brussels, avenue Paul Heger cp160/16 - P2 1050 Brussels, Belgium

Notations

$|Fo|$ and $|Fc|$ - experimental and calculated amplitudes of structure factors

$|Eo|$ and $|Ec|$ - experimental and calculated normalised amplitudes of structure factors

$Fc = |F_c|e^{i\phi_c} = (A_c, B_c)$

s - vector of position of reciprocal space point $|s| = 2\sin\theta/\lambda$

σ_e - experimental uncertainties of structure factor amplitudes

$\sigma_{e;n}$ - experimental uncertainties of normalised amplitudes

f_j - atomic scattering factor

$\Sigma_N = \sum_{j=1}^{Nall} f_j^2$

$\Sigma_c = \sum_{j=1}^{Npresent} f_j^2$

$\Sigma_q = \sum_{j=1}^{Nabsent} f_j^2$

N_{all} - total number of atoms

$N_{present}$ - number of atoms used in the present model

N_{absent} - number of atoms not included in the present model

ε - multiplicity of the scattering plane

$\Sigma = \varepsilon(\Sigma_c(1 - D^2) + \Sigma_q)$

Δr - average coordinate error

$D =< \cos \Delta rs >$

$F_{wc} = \sum_{j=1}^{N_{part}} D_j F_j^c = |F_{wc}|e^{i\phi_{wc}}$ - weighted sum of partial calculated structure factors

$\sigma_A = \frac{\Sigma_c}{\Sigma_N} D$

m - figure of merit

m should be equal to $< \cos \Delta\phi >$ where $\Delta\phi$ is the phase error between the current ϕ_c and the true value of ϕ

$m = \frac{I_1(X)}{I_0(X)}$ for acentric, $\tanh(X)$ for centric reflections
where $X = \frac{2\sigma_A|Eo||Ec|}{2\sigma_{e|n}+\Sigma}$ for acentric
and $\frac{\sigma_A|Eo||Ec|}{\sigma_{e|n}+\Sigma}$ for centric reflections

1 Introduction

The aim of this talk is to review the advantages of the maximum likelihood refinement method over least-squares minimisation for macromolecules. As Read (1990,1996), and Bricogne (1992,1996) have given a comprehensive description of the theory of maximum likelihood we will give only a brief summary of the basic differences between the least-squares and maximum likelihood approach.

Amplitude based least-squares (LSQF) minimises the following quantity against the atom parameters and the overall scale.

$$\sum w(|Fo| - k|Fc|)^2 \tag{1}$$

The assumption behind this is that the conditional distribution of amplitudes of structure factors with respect to the atom coordinates is approximately Gaussian. Although LSQF have been applied successfully for refinement of crystal structures for many years it has some disadvantages for macromolecules where the parameters are ill determined, and the errors are often large. We will mention only three of them.

1) Determination of useful weights for the observations. Many programs now in use use a unit weighted least-square residual (w=1) which implies that all reflections have been measured with equal accuracy and the contribution of coordinate errors to each reflection is the same. David Smith (1996) shows that choosing a two line weighting scheme depending on resolution improves the refinement behaviour. But this only partially solves the problem: these weights do not include any information of the experimental uncertainty of reflections and they must be chosen by trial and error methods.

2) Determination of the overall scale factor between Fo and Fc. This is an essential first step before calculating the least squares derivatives. Most existing programs use a form of Wilson scaling

$$k = k_0 e^{-B_0 s^2} \tag{2}$$

which assumes that atoms are evenly distributed over the whole unit cell and that therefore $< Fo >$ falls off smoothly with resolution. But for macromolecules this is not true, there is a clear distinction between buried and surface protein regions, and between these and the solvent, which is reflected in the distribution of $< Fo >$.

3) Adding experimental phase information (eg: MIR/MAD/NCS) is not straightforward.

Following on from Luzzati's (1952) distribution Srinavasan and Ramachandran (1965) shows that when there is no phase information the conditional distribution

of amplitudes of structure factors are better expressed as a Rice distribution where each observation is weighted by function of σ_A. Estimates of these are deduced from the agreement between $|Fo|$ and $|Fc|$. If we use their result and add experimental uncertainties of structure amplitudes to the Σ-s we obtain the following expression for the log of the maximum likelihood (LLK_h) (A similar way of adding experimental uncertainties has been suggested by Bricogne and Gilmore 1990):

$$LLK_h = \begin{cases} \frac{|F_o|^2+D^2|F_c|^2}{2\sigma_e^2+\Sigma} - \log I_0(\frac{2|F_o|D|F_c|}{2\sigma_e^2+\Sigma}) + \log(2\sigma_e^2+\Sigma) & \text{for acentric reflections} \\ \frac{|F_o|^2+D^2|F_c|^2}{2(\sigma_e^2+\Sigma)} - \log\cosh(\frac{|F_o|D|F_p|}{\sigma_e^2+\Sigma}) + \frac{1}{2}\log(\sigma_e^2+\Sigma) & \text{for centric reflections} \end{cases} \quad (3)$$

where $\Sigma = \varepsilon(\Sigma_c(1-D^2)+\Sigma_q)$

The version of this equation for normalised structure factors is:

$$LLK_h = \begin{cases} \frac{|E_o|^2+\sigma_A^2|E_c|^2}{2\sigma_{e;n}^2+\varepsilon(1-\sigma_A^2)} - \log I_0(\frac{2|E_o|\sigma_A|E_c|}{2\sigma_{e;n}^2+\varepsilon(1-\sigma_A^2)}) + \log(2\sigma_{e;n}^2+\varepsilon(1-\sigma_A^2)) & \text{acentric} \\ \frac{|E_o|^2+\sigma_A^2|E_c|^2}{2(\sigma_{e;n}^2+\varepsilon(1-\sigma_A^2)} - \log\cosh(\frac{|E_o|\sigma_A|E_c|}{\sigma_{e;n}^2+\varepsilon(1-\sigma_A^2)}) + \frac{1}{2}\log(\sigma_{e;n}^2+\varepsilon(1-\sigma_A^2)) & \text{centric} \end{cases} \quad (4)$$

Macromolecular crystallographers should already be familiar with these equations. Read's program (1986) SIGMAA uses them to generate less biased coefficients for maps calculated using phases from partial Fc-s. It is essential to get a reasonable estimate of σ_A as a function of resolution. SIGMAA does this by using equation (4) in reciprocal space resolution shells. Each of these shells needs to include several hundred reflections to give a reliable estimate. Another way of estimating σ_A could be by fitting it to some function of resolution as is done for the scale factor (see below).

2 Implementation within REFMAC

Maximum likelihood refinement (MLKF) has been implemented in the program REFMAC. At each cycle the program performs two steps. First it estimates the overall parameters of likelihood (σ_A-s). Secondly it uses these parameters to build the likelihood function and refine the atomic parameters.

1) For estimation of overall parameters of likelihood REFMAC uses an idea suggested by Dale Tronrud (1995) to find the overall scale between $< Fo >$ and $< Fc >$. He uses a two Gaussian approximation for the scale factor which is based on the assumption that the contributions of solvent and protein parts of the crystal to the structure factors are negatively correlated and the scale could be better expressed as:

$$k = k_0 e^{-B_0 s^2}(1 - k_1 e^{-B_1 s^2}) \quad (5)$$

Typically B_1 is large, and only modifies the scale at resolution below 6A. The same idea can be used to estimate σ_A and we can write:

$$\sigma_A = \sigma_{A;0} e^{-B_0 s^2}(1 - \sigma_{A;1} e^{-B_1 s^2}) \quad (6)$$

(There is a high degree of correlation between the variables in these equations and to get a satisfactory solution the program uses singular value decomposition to solve the linear equations (Press et al, 1986). This method is very similar to that described by Ten Eyk (1996).)

This representation of σ_A means that only a few hundred reflections are enough to estimate it and that one can use only the "free" reflections (Brunger 1993) for this. (In one of our test cases we estimated σ_A satisfactorily using only 200 reflections). Using the "free" reflections means that this error estimate is less biased towards the existing model. The program calculates the scale factor using all working reflections but it should be noted that in the case of maximum likelihood refinement the scale factor is only used for R-value calculations which allow users to follow the progress of the refinement using this familiar statistic. The program also reports the figure of merit m which should approximate the $< \cos \Delta\phi >$.

2) After estimation of the overall parameters the program performs one cycle of atomic parameters refinement. To do this it calculates the gradient and the diagonal terms of the second derivative matrix. For the gradient calculation it uses the technique of convoluting the atomic density with the difference density as suggested by Agarwal (1978) and Lifshcitz (Agarwal et al, 1980). In the LSQF case we calculate the difference map with coefficients:

$$w(|Fo| - k|Fc|)e^{\iota\phi_c} \tag{7}$$

In MLKF refinement we calculate the map with coefficients

$$[(m|Fo| - D|Fc|)e^{\iota\phi_c})/\Sigma] \tag{8}$$

Read (1986) shows that map with coefficients (8) is less biased towards the incorrect parameters than a map calculated with coefficients (7).

At the end of a cycle the program also writes out map coefficients for SIGMAA style $(m|Fo| - D|Fc|)$ and $(2m|Fo| - D|Fc|)$ maps taking care to restore missing data. Tronrud (1996) and Cowtan (1996) show that absent reflections cause unpredictable noise in map calculations which sometimes may lead to errors in interpretation. Assuming that absent reflections are best approximated by setting them equal to their calculated value (or in the case of maximum likelihood setting $m|Fo| = D|Fc|$) then the difference contribution is zero, and the $2m|Fo| - D|Fc|$ contribution is $D|Fc|$.

These coefficients are:

$$FWT = \begin{cases} (2m|Fo| - D|Fc|)e^{\iota\phi_c} & \text{if reflection was included in refinement} \\ D|Fc|e^{\iota\phi_c} & \text{otherwise} \end{cases}$$

$$DELFWT = \begin{cases} (m|Fo| - D|Fc|)e^{\iota\phi_c} & \text{if reflection was included in refinement} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

But one should be careful when restoring absent reflections in this way. This type of map coefficients will reduce noise but may introduce bias. We believe the best way of dealing with absent reflections is measuring them.

3 Examples of application.

In each case described here the MKLF refinement was carried to convergence from an existing model without any rebuilding. Results were compared to maps and phases generated from the final coordinates provided by the authors. The examples discussed are listed in the Table 1.

Table 1: Examples

	BA2	Cytochrome c'	Insulin
Space group	$C222_1$	$P6_522$	$P2_1$
Cell dimensions	52.5 77.7 238.2 90 90 90	54.5 54.5 181.0 90 90 120	53.9 64.8 48.9 90 109.81 90
Resolution(Å)	2.2	2.0	1.9
Number of residues	483	125	318
Method of solution	MIR good model	MR low homology	MR high homology
Completeness of the model	85% of protein atoms no waters	25% homology $\propto$ 10 residues were misfitted	all protein atoms and 20% of waters
Final R-value/R-free	11.9/20.9	16.7/NA	18.4/25.2

3.1 Bacterial Chimaeric α-amylase (BA2). Beginning refinement from an excellent model.

The structure of BA2 was solved by Brzozowski et al (1996) by the MIR method. The initial and final coordinates and the data were provided kindly by Dr Lawson. 5% of the reflections were reserved for FreeR estimation before any refinement was carried out. The initial model was built so carefully that any program could refine it satisfactorily. Even in this case however maximum likelihood refinement gave a lower phase error than the least-square methods (Table 2). Various map correlations to the final Fc map were calculated. As can be seen from Table 2 the map with coefficients calculated using SIGMAA has a higher correlation with this map than the unweighted $2|Fo| - |Fc|$ map. After REFMAC there was more improvement of the map than after LSQF followed by calculation of SIGMAA coefficients. The reason for this is partly because phase errors are less than for the LSQF refined model and partly because m and σ_A are estimated using the free R reflections which gives a more realistic estimate of the coordinate error. The overall m before and after REFMAC are very close to the real $< \cos \Delta\phi >$. After LSQF refinement the m calculated by SIGMAA from all the data is overestimated and hence the maps are more biased towards the existing model than that generated from the REFMAC coefficients.

Table 2: BA2 refinement results. MIR Model

	R-value	R-free	$< \Delta\phi >$	$< cos\Delta\phi >$	m	mapcorrelation
Initial	0.49	0.47	56.4	0.44	0.46	?
LSQF	0.29	0.41	41.3	0.63	-	0.76
SIGMAA	-	-	-	-	0.79	0.81
ML	0.30	0.38	37.4	0.67	0.66	0.83

3.2 Cytochrome c'. Preparing to rebuild from a Molecular Replacement solution

This starting model was based on a molecular replacement solution which had been subjected to initial LSQF refinement using all reflections. The model used had only 25% homology to this form of Cytochrome c'. (Subsequently the structure has been fully refined, and the coordinates have been deposited at Brookhaven (1cgn.pdb). All comparisons use these coordinates as the final set. (Baker et al 1995).) Ten residues had out of register errors and another 10 residues were completely misfitted. In these cases where an extensive rebuilding is necessary, the problem of map bias is very serious. For the subsequent maximum likelihood refinement 5% of reflections were assigned as "free"which were used for estimation of the overall parameters of likelihood. Again after LSQF refinement the map calculated using coefficients from SIGMAA correlated with the final Fc model map better than a $2|Fo| - |Fc|$ map (Table 3). REFMAC was able to refine the LSQF model further and the phase error was reduced by 6 degrees. The map correlation coefficient is 5% higher than that for the map calculated by SIGMAA coefficients. The "free R-value" increased towards a more realistic value during refinement and the maps became less biased. After 30 cycles the m was close to the real $< \cos \Delta\phi >$ but still overestimated, showing the importance of assigning "free" reflections at the beginning of refinement. The behaviour of m and the real $< \cos \Delta\phi >$ vs resolution (Figure 1) shows that during refinement the phases for the high resolution shells were improved most. The $2Fo - Fc$ map is seriously biased and noisy and it would be possible to build an incorrect model through the density for a water molecule, with the chain perpendicular to its true direction. The SIGMAA map is better but still there is a break in the main chain and the electron density could be interpreted wrongly. The map after REFMAC shows less ambiguous connectivity and density for side chains and water molecules has appeared.

3.3 Cross-linked Insulin. End stages of refinement when LSQF has apparently converged

The solution for this cross-linked insulin was found by molecular replacement using a model with 95% homology. Data for this case were provided by Dr David Edwards.

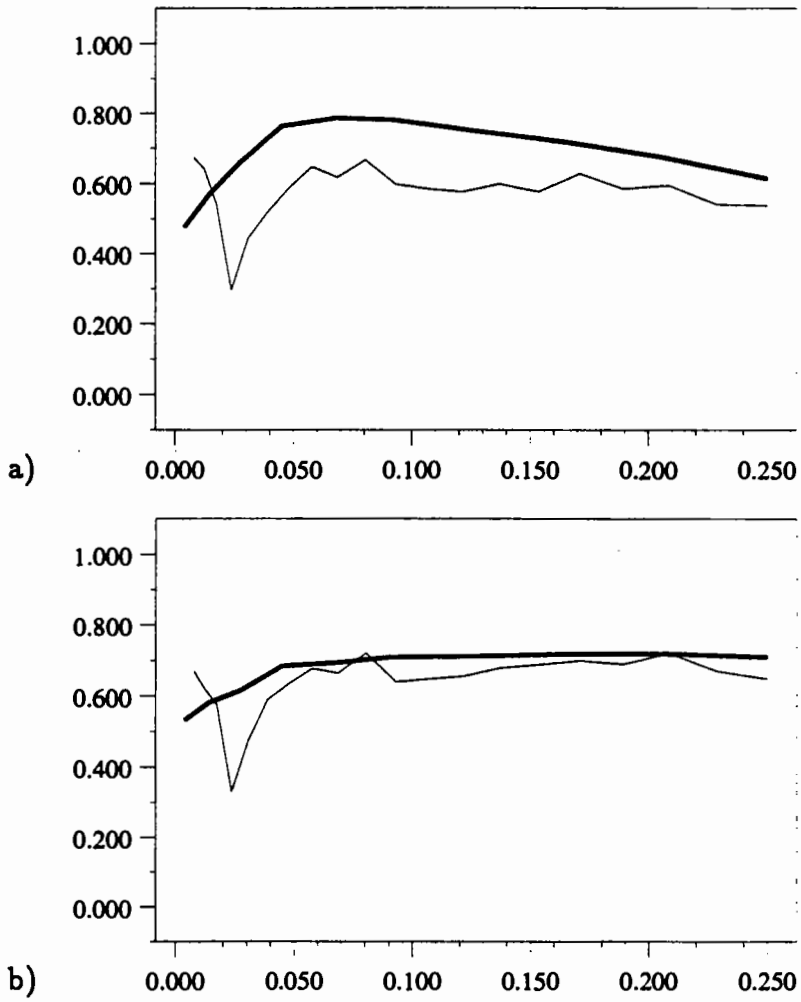


Figure 1: Cytochrome C': Behaviour of m and $< \cos \Delta\phi >$ where $< \Delta\phi >$ is the r.m.s phase difference from the final model a) before and b) after MLKF refinement. Bold lines show estimated m, thin lines show $< \cos \Delta\phi >$. The fluctuation in $< cos\Delta\phi >$ at low resolution is probably due to the small number of reflections in these bins; it is not a general property for the proteins studied.

Least-square refinement and rebuilding cycles had given an R-value of 24% and free R-value 34%. Maximum likelihood refinement using REFMAC reduced the R-value by 4% but the free R-value dropped even more - by 6% (Table 4). It is interesting to note that geometric parameters such as r.m.s. deviation of bond lengths from ideality also improved. The plot of R-value vs resolution (Figure 3) shows that there is more improvement at high resolution than at low resolution. The reason for this is that the fit of the high resolution structure factors depends on the accurate position of atoms but the low resolution data fit depends on large movement of structure or a more complete description of the model. To improve low resolution one needs to rebuild the model or add new features such as waters. After rebuilding and further refinement by REFMAC the R-value was reduced to 18% and free R-value to 25% (Edwards, Personal communication). There are two Zn atoms in this insulin structure. It is well known that the position of heavy atoms are more accurately determined than that of lighter

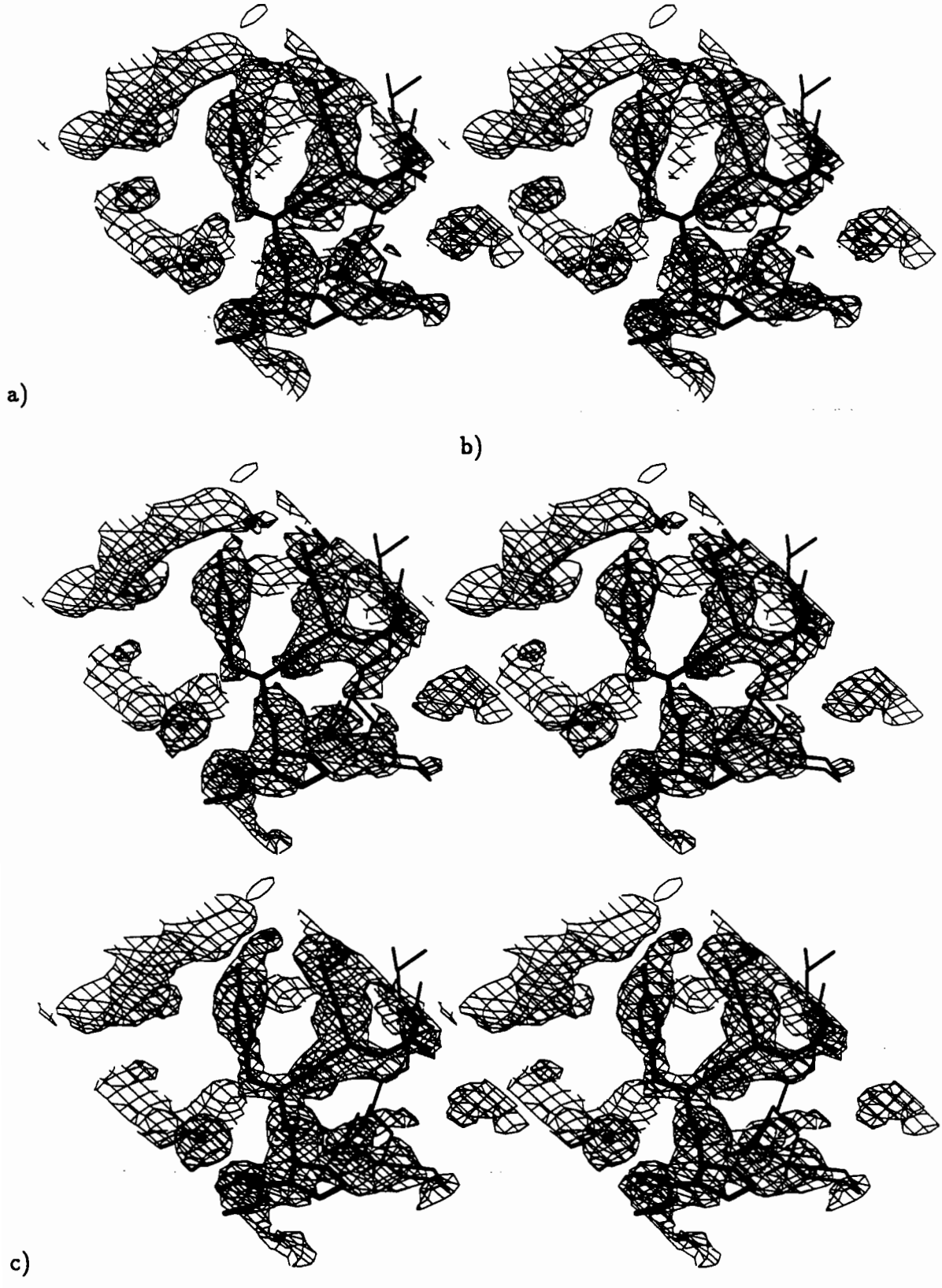


Figure 2: Cytochrome C' Electron densities calculated after LSQF and MLKF. Bold lines show final coordinates, thin lines show coordinates included in refinement. This loop was completely misplaced. a) $2|Fo|-|Fc|$ map b) SIGMAA map and c) REFMAC map.

Table 3: Cytochrome c' refinement results

	R-value	R-free	$< \Delta\phi >$	$< \cos \Delta\phi >$	$< m >$	mapc
LSQF	34.8	NA	44.4	0.58	-	0.66
SIGMAA	-	-	-	-	0.72	0.73
ML	31.2	37.1	38.0	0.66	0.70	0.79

atoms. This may cause problems during refinement owing to the assumption that all atoms have same expected positional error and error of B-values. An approach for dealing with this problem will be discussed in the next section.

Table 4: Cross Linked Insulin.

	R-factor	R-free	r.m.s.d(Å)	r.m.s.a(o)
LSQF	24.3	33.8	0.021	4.1
ML	20.0	28.1	0.018	2.6

r.m.s.d - r.m.s. deviation of bond distances from ideal ones
r.m.s.a - r.m.s. deviation of bond angles from ideal

4 Implemented but untested features of REFMAC

1. Inclusion of phase information known prior to refinement. The likelihood function can be generalised as following:

$$LLK_h = \begin{cases} \frac{|Fo|^2+D^2|Fc|^2}{2\sigma_e^2+\Sigma} - \log \int_0^{2\pi} P(\phi) e^{\frac{2|Fo|D|Fc|}{2\sigma_e^2+\Sigma}\cos(\phi-\phi_c)} d\phi + \log(2\sigma_e^2+\Sigma) & \text{acentric} \\ \frac{|Fo|^2+D^2|Fc|^2}{2(\sigma_e^2+\Sigma}) - \log \sum_{l=0}^{1} P(\phi_l) e^{\frac{|Fo|D|Fc|}{\sigma_e^2+\Sigma}\cos(\phi-\phi_l)} d\phi + \frac{1}{2}\log(\sigma_e^2+\Sigma) & \text{centric} \end{cases} \quad (10)$$

where $\Sigma = \varepsilon(\Sigma_c(1-D^2)+\Sigma_q)$, ϕ_0 and $\phi_1 = \phi_0 \bmod \pi$ are two possible phases of centric reflections

It is easy to write a version of this equation for normalised structure factors.

It means now one can use experimental phases and figure of merits where they are available. If $m = 1.0$ (ie,the phases ϕ are assumed to be known exactly), then the Rice distribution transforms to a Gaussian distribution for actual structure factors.

2. Sub-dividing the Fc contributions.

In some cases there is an advantage in breaking the Fc into several components and assigning different likelihood functions to each component. The equations then become (for acentric reflections only.):

$$LLK_h = \frac{|Fo|^2+|F_{wc}|^2}{2\sigma_e^2+\Sigma_{wc}} - \log I_0\left(\frac{2|Fo||F_{wc}|}{2\sigma_e^2+\Sigma_{wc}}\right) + \log(2\sigma_e^2+\Sigma_{wc}) \quad (11)$$

where $F_{wc} = \sum_{j=1}^{Npart} D_j F_j^c$ and $\Sigma_{wc} = \varepsilon \sum_{j=1}^{Npart} \Sigma_j(1-D_j^2)$. Summations are over all possible partial structures.

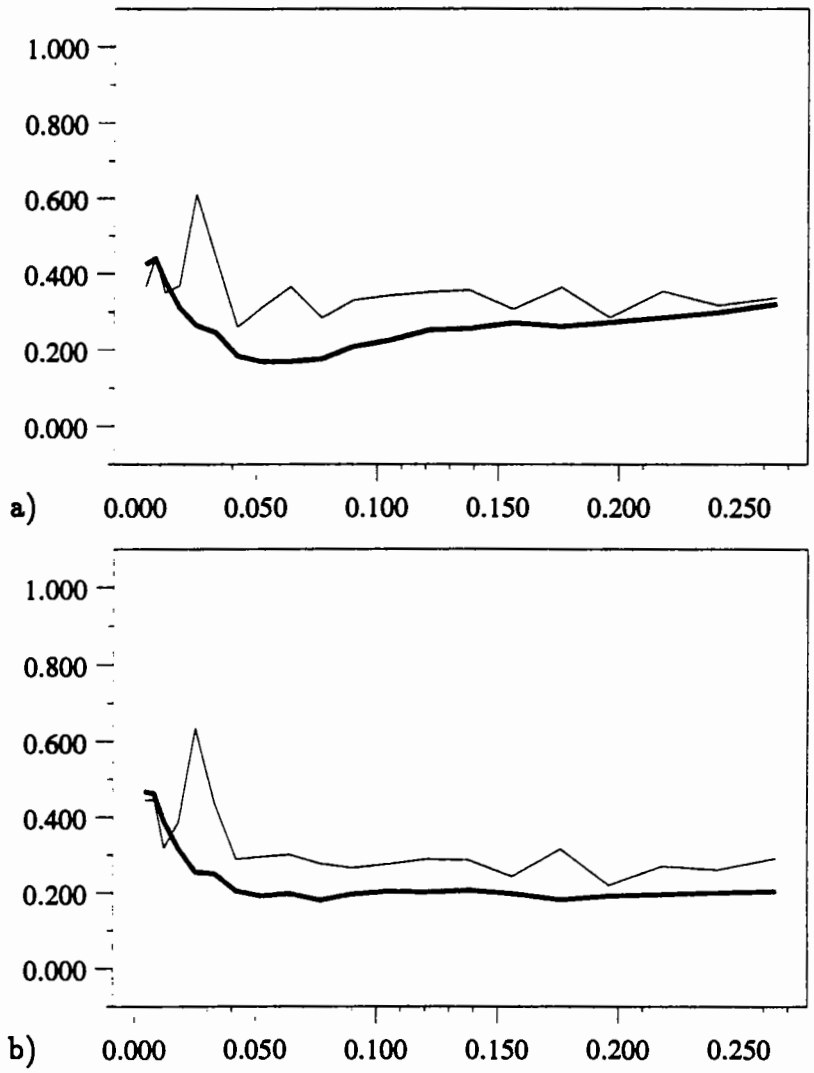


Figure 3: Insulin: Behaviour of R-value and R-free vs resolution a) before and b) after REFMAC. Bold lines show R-value, thin lines show R-free. Note the improvement in the R-value at high resolution.

To write the same equation when some prior distribution for phases is available is straight forward.

The program will scale all available structure factors together and will find all parameters of likelihood. In particular this is advantageous when some part of the Fc is derived from a metal cluster. The error associated with such a heavy atom is much smaller than that of the protein atoms. Another way to exploit this flexibility is to calculate the partial contribution to Fc from by Fourier-transforming density where it is not possible to place atom sites accurately. For example this could done for solvent shells or for part of the structure that cannot yet be built into a experimentally phased map. Another way of using of this expression of likelihood could be to weight domains with markedly different average B values independently.

Using the above features and programs available in CCP4 (1994) it is possible to design different type of refinement or phase improvement procedures.

3. Refinement of the cell parameters. These must be refined using the fit to the geometric restraints. The crystallographic coordinates of the atoms are not sensitive to small changes in cell dimensions, but such errors mean that the conversion of the fractional coordinates to orthogonal coordinates is not reliable, and this leads to error in the deduced geometric characteristics of the molecule. More tests of this refinement have been done.

Acknowledgements

We thank the people in the Protein structure group of the Chemistry Department of University of York for testing the program, also Dr Phil Evans, Dr Misha Isupov and Dr Kostya Polyakov for useful comments. This work has been supported by EU grant BIO2CT-92-0524, and in part by the MRC grant G9413078MB.

References

Agarwal, R. C. (1978) *Acta Cryst.* **A34** 791-809

Agarwal, R.C., Lifchitz, A. and Dodson, E. (1980) in the *Refinement of Protein structures*. Proceedings of Daresbury Study Weekend. 36-39

Baker, E.N., Anderson, B.F., Dobbs, A.J. and Dodson, E.J. (1995) *Acta Cryst.*, **D51** 282-289

Bricogne, G. (1992) in the *Molecular replacement*. Proceedings of Daresbury weekend. 62-75

Bricogne, G. and Irwin, J. (1996) this proceedings

Bricogne, G. and Gilmore C.J. (1990) *Acta Cryst.* **A46** 284-297

Brunger, A.T. (1993) *Acta Cryst. D49* 24-36

Brzozowski, A.M., Lawson, D.M., Frandsen, H.B., Svendsen, A., Borchert, T., Dauter, Z and Dodson, G.G. (1996) submitted to the *Structure*

Collaborative Computational Project, Number 4 (1994) *Acta Cryst.* **D50** 760-763

Cowtan, K (1996) This proceedings

Hendrickson, W.A. and Lattmann (1970) *Acta Cryst.* **B26** 136-143

Konnert, J.H. and Hendrickson, W.A. (1980) *Acta Cryst.* **A36** 344-350

Luzzati, V. (1952) *Acta Cryst.* **5** 802-810

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986) Numerical recipes. The art of Scientific computing. Cambridge University press.

Read, R.J. (1986) *Acta Cryst.* **A42** 140-149

Read, R.J. (1990) *Acta Cryst.* **A46** 900-912

Read, R.J. (1996) This proceedings

Ten Eyk, L. (1996) This proceedings

Tronrud, D. (1995) *Methods in Enzymology* (in press)

Tronrud, D. (1996) This proceedings

Smith, D. (1996) This proceedings

Srinavasan, R. and Ramachandran, G.N. (1965) *Acta Cryst.* **19**, 1008-1014

X-RAY ANALYSIS OF DOMAIN MOTIONS IN PROTEIN CRYSTALS

D. S. MOSS, I. J. TICKLE, O. THEIS and A. WOSTRACK
Department of Crystallography, Birkbeck College,
Malet Street, London, WC1E 7HX, England

1 Introduction

The resolution of most X-ray structure analyses of macromolecules precludes the detailed anisotropic analysis of individual atomic displacements such as are routinely carried out in the structure determinations of small molecules. However, macromolecules contain groups of atoms which move approximately as rigid bodies and by using this prior knowledge the anisotropic displacements of these groups can be determined. The domains and folding motifs of proteins are obvious candidates for rigid-body treatment. The main-chain atoms of α-helices in proteins and the bases, ribose and phosphate moieties in nucleic acids may also be regarded as rigid to a first approximation. The planar groups of protein side chains such as phenyl or imidazole rings are examples of putative rigid bodies if data of sufficient resolution are available.

The use of rigid groups in the analysis of atomic mean-square displacements was pioneered by Schomaker and Trueblood (1968) who used mean-square translation (T), libration (L) and screw-rotation (S) tensors to describe the rigid body motions of groups of atoms in small molecules. This TLS model has been used in a number of macromolecular refinements to produce estimates of mean-square rigid body displacements. Holbrook and co-workers have applied the TLS model to a dodecamer of DNA (Holbrook and Kim, 1984; Holbrook *et al.*, 1985). Applications of the TLS model to the proteins ribonuclease A and papain have been made by Howlin *et al.* (1989) and Harris *et al.* (1992), respectively.

Errors in a TLS analysis may arise from two distinct causes. Firstly the groups of atoms to which the model is applied never constitute a strictly rigid unit. This will

generally cause a systematic error in the TLS parameters although some non-rigid-body motions are also compatible with the model. Such cases imply that the tensor components cannot be interpreted solely in terms of rigid body displacements. The second, and more important cause in practice, is the paucity of data. This causes random errors in the tensor components which are larger in the case of smaller rigid bodies whose displacements have a smaller effect on the diffraction pattern.

The derivation of error estimates for individual parameters from least-squares analysis of X-ray data is notoriously difficult in the case of macromolecules. The difficulties arise primarily from the fact that the usual structure factor model poorly represents the less well ordered parts of the macromolecule. Further problems arise from the restraints or energy terms which have to be used in the refinement of such structures. These cause difficulties in estimating the number of degrees of freedom which should be used in scaling the standard deviations obtained from the inverse of a least-squares normal matrix. Errors in TLS analyses are discussed by Butler *et al.* (1994).

This paper considers the estimation of TLS parameters from least-squares refinement and illustrates the method by reference to TLS refinement of the rigid body displacements of the two domains of the eye-lens protein γB-crystallin.

2 Method

The structure of γB-crystallin (previously called γII-crystallin) had been determined from X-ray diffraction data collected at the Daresbury synchrotron using photographic film (Wistow *et al.*, 1983). The refined co-ordinates resulting from this work (Najmudin *et al.*, 1993) were the starting point of the work of the TLS refinement. The crystal data for the protein is given in table 1.

γB-crystallin is composed of two globular domains related by a pseudo dyad. For each of these domains, twenty TLS parameters were obtained by least-squares refinement using the program *RESTRAIN* (Driessen *et al.*, 1989). Initial values of the TLS parameters were set to zero except for the diagonal elements of the **T** tensor which were set to 0.1 Å^2 and the diagonal elements of the **L** tensor which were set to 1 deg^2. Atoms which were not part of the rigid groups (all side groups and main-chain atoms which are in external loops) were given isotropic temperature factors as determined by conventional refinement. Eight cycles of refinement were carried out at which point the shifts in the TLS parameters became small compared

Table 1: Crystal data of γB-crystallin

Data resolution (Å)	1.47
Space group	$P4_12_12$
Molecules per asymmetric unit	1
Number of residues	175
Number of non-hydrogen protein atoms	1474
Number of solvent molecules	234
PDB code*	1gcs

*Code associated with the co-ordinates deposited in the Protein Data Bank

with their standard deviations. The final conventional R-factor was 19.5 %. A full account of the TLS refinement will be published elsewhere.

Tensor components for each of the domains were then transformed to a co-ordinate system with the origin on the centre of reaction (Schomaker and Trueblood, 1968) and axes parallel to the eigenvectors of the **L** tensor. These transformations were carried out using the program *TLSANL* (Howlin *et al.*, 1993).

It is very important to obtain estimates of the precision of TLS parameters. The estimated standard deviations (esd) were calculated for the TLS parameters of the γB-crystallin from the inverses of the 20 × 20 normal matrices constructed for each rigid body. The values were scaled by the ratio of the least-squares residual to the number of degrees of freedom. The latter was taken as the number of squared terms minus the number of parameters.

3 Results

The values of the **TLS** tensors and their esd's are shown in table 2. All the **T** parameters are highly significant and show that that the translational motion of the domains is approximately isotropic.

The libration tensors of the domains show much greater anisotropy and a greater range of values than the translation tensors. Analysing the principal axes of the tensors shows that both the N and C terminal domains possess a dominant axis of libration. Each axis passes through the hydrophobic interface between the two domains. Figure 1 shows these axes.

The precision of the screw-rotation **S** tensor is much less than either **T** and **L**. In this refinement most components of the screw rotation tensor were significant.

Table 2: Rigid-body displacements of N-terminal and C-terminal domains of γB crystallin (values in brackets are standard deviations)

N-terminal domain
Number of atoms: 300

Segment	Start	End	Atoms
1	5	199	MNCH
2	245	685	MNCH

Mean centre	(Å)	5.430	17.127	36.739
Centre of reaction	(Å)	6.146	19.199	35.613

T	Å^2	0.1061 (0.0023)	0.1200 (0.0021)	0.1065 (0.0019)	−0.0101 (0.0015)	−0.0006 (0.0016)	−0.0014 (0.0017)	. .	. .
L	deg^2	0.63 (0.08)	1.91 (0.10)	3.91 (0.13)	−0.11 (0.08)	−0.29 (0.08)	0.11 (0.07)	. .	. .
S	Å×deg	−0.018 (0.008)	0.007 (0.010)	−0.071 (0.007)	−0.006 (0.009)	0.034 (0.008)	−0.061 (0.008)	0.134 (0.013)	−0.068 (0.013)

C-terminal domain
Number of atoms: 276

Segment	Start	End	Atoms
1	723	942	MNCH
2	1016	1341	MNCH
3	1388	1434	MNCH

Mean centre	(Å)	13.852	16.605	14.885
Centre of reaction	(Å)	13.249	16.963	17.619

T	Å^2	0.1292 (0.0024)	0.0992 (0.0022)	0.1161 (0.0019)	0.0001 (0.0015)	0.0151 (0.0016)	0.0155 (0.0017)	. .	. .
L	deg^2	0.73 (0.09)	0.74 (0.09)	3.10 (0.12)	−0.26 (0.08)	0.39 (0.07)	0.44 (0.07)	. .	. .
S	Å×deg	0.008 (0.008)	0.027 (0.010)	−0.002 (0.008)	0.112 (0.009)	0.008 (0.007)	−0.015 (0.008)	0.013 (0.013)	−0.040 (0.013)

However, at poorer resolutions or in smaller rigid groups the components of the **S** tensor may not be significant.

4 Conclusions

Looking at the L_{33} values in table 2 shows that there is one major direction of libration through each of the two domains and this is approximately parallel to the **z** axis. Figure 1 shows the principal axes of libration of each domain as determined from *TLSANL*. The midpoint of each axis in figure 1 is the centre of reaction, which to a first approximation can be thought of as the point about which each domain is librating.

These axes pass through the interdomain interface and suggest that this interface is a strong determinant of domain association. This model of torsional motion about this interface is in contrast to a hinge-bending model where the connecting peptide would play a role in holding the domain together.

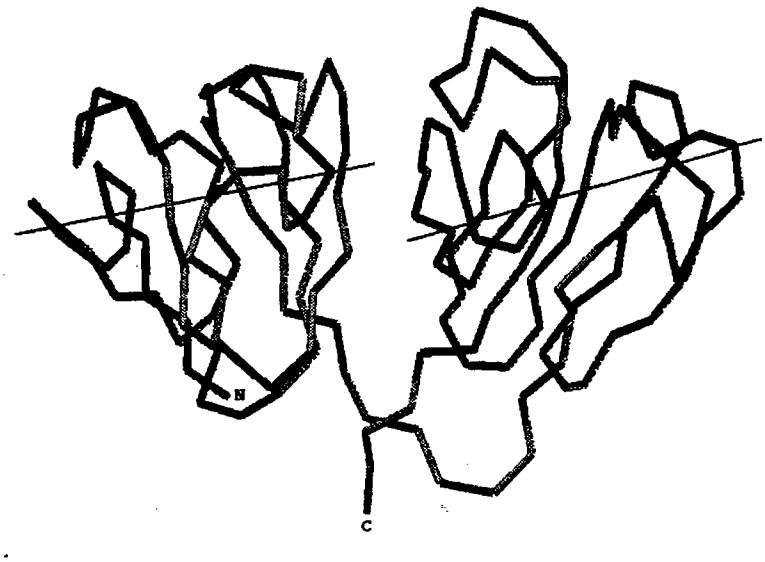


Figure 1: Principal axes of libration of each of the domains of γB-crystallin. The centres of each line are at the centre of reaction of the displacements.

Appendix: Excerpts from the RESTRAIN manual

Physical background:

In this option the atomic displacements of atomic groups are refined using the approximation that the groups have, either partly or wholly 'rigid body' displacements. The atomic groups (TLS groups) may be whole molecules, units of secondary structure (e.g. alpha helices) or they may be pseudo-rigid side groups such as phenyl rings, imidazole, carboxylate, guanidinium or amide groups. When units of secondary structure are chosen, there is an option to include main chain atoms only. For small groups (i.e. < 20 atoms) data at high resolution (e.g. 1.5A) may be required for success. It should also be remembered that the TLS model assumes harmonic displacements and this may not be valid for side groups on the surface of a macromolecule.

The TLS option allows refinement of only 20 parameters for each TLS group instead of six for each anisotropic atom (see section 2.3.7). The rigid groups in proteins which may be suitable are the aromatic rings of PHE, TYR and TRP, the propellors of ASP/ASN, GLU/GLN and ARG, ligands, the secondary structure elements, the domains or the entire molecule.

Usage:

All information for the TLS refinement is contained in the Brookhaven atom file. The steering file does not contain any information. Isotropic refinement can be selected concurrently with TLS refinement (ISO=.T.). The ADDISO option (default is .T.) allows the group TLS and individual isotropic parameters to be refined simultaneously, taking account of their correlation.

Each TLS group is defined in one TLS record in the atom file. One TLS record spans several physical lines. The record starts with the TLS identifier, 'TLS', followed by an integer number of contiguous segments, NSEGM. These segments contain the atoms of the TLS group. The following NSEGM lines contain the specification of atoms in each segment of the associated TLS group.

Every line contains two atom identifiers indicating the start and finish of the segment followed by the identifiers of atoms to be selected from this segment for inclusion in the TLS group. An atom identifier is interpreted as a character string, not as an

integer. Atom names have to conform to the Brookhaven convention. All atom codes found in the Brookhaven atom files can be used. Additionally, four group codes can also be specified: 'SDCH', 'MNCH', 'ALL' and 'NOT'. 'MNCH' will select all mainchain atoms (' N ', ' CA ', ' C ' and ' O '), 'SDCH' selects all non-mainchain atoms, 'ALL' selects all atoms and 'NOT' negates the selection of atom types on the line. The order of atom specifiers is not important. If no atom specifier is given, the default is 'ALL'. The next four lines of the TLS record contain: centre of origin for calculations with this TLS group, T tensor, L tensor and S tensor. The centre of origin for a ring is usually a C-beta atom; for larger groups such as domains it is usually the centre of gravity. The S tensor line ends the TLS record. The position and order of TLS records in the Brookhaven atom file are not important, but it is convenient to have them collected together at the top of the atom file.

The layout of the TLS record is:

RECORD		UNIT
TLS	NSEGM	
	SEGMENT	
	. . . . .	
	ORIGIN	A
	T11 T22 T33 T23 T13 T12	A^2
	L11 L22 L33 L23 L13 L12	degr^2
	S23 S31 S12 S32 S13 S21 S22 S11	A*degr

The format is free, that is items separated by one or more spaces. If items are left blank they default to zero values. Note that this is different from the remainder of the coordinate file, which is in fixed format.

Here the Tij means an element (i,j) of tensor T. Since X-ray data allow the calculation of only eight of nine S tensor elements, the usual constraint of setting the trace of S to zero is adopted in RESTRAIN. This means that the elements S11 and S22 of RESTRAIN are in fact (S11 - S33) and (S33 - S22) of the S tensor as defined by the equation U = T + A L A' + A S + S'A' (Johnson and Levy, 1974).

An example of the TLS record specifying a TLS group consisting of two mainchain segments, atoms 1 to 68 and 129 to 300, respectively, is:

```
TLS          2
               1   68 MNCH
             129   30 MNCH
           2.572  33.400   3.315
            .112    .165    .131   -.052   -.003   -.003
           1.877   2.165   3.471   4.562   6.152   7.313
            .366   -.382    .147   -.981    .185    .118    .132    .140
```

Alternatively, the atoms may be specified by their residue and atom labels; for details see section 3.1.1 below under "XTRDIS".

Warning and error messages:

Where TLS tensors result in U tensor that is not positive-definite, a warning message is printed out stating the atom name, number and U tensor.

If the L tensor elements are large (>20 degr^2) and an atom is far away from the centre of origin for the calculation of the TLS tensors (>20 A), then the observed and calculated structure factor amplitudes can be different by several orders of magnitude. This is a consequence of the numerical instability in calculation of derivatives of the TLS tensors with respect to positional coordinates (on some machines it may also result in an over-flow floating point error). These problems usually appear at the beginning of the TLS refinement of large groups if the user does not set the initial L small enough and origin of the rigid group sufficiently close to the centre of gravity. RESTRAIN checks for such an error in two ways. First, it prints a warning message if the selected origin is more than 10A away from the gravity centre. Second, it prints a warning message if more than 30% of elements of U tensors for individual atoms had to be reset to an arbitrary interval [0, ULIMH].

Please note that since the TLS records are not according to the Brookhaven specifications, there may be problems with other programs reading these coordinates.

Note that TLS calculations, like all anisotropic calculations, cannot take advantage of space-group specific subroutines. The general space-group subroutine must be used.

References

Butler, S. A., Harris, G. W., Moss, D. S., Gorinsky, B. A., Adams, M. J. and Gover, S. (1994). Error estimates in the determination of rigid body displacements in protein crystals. *J. Chem. Crystallogr.*, **24**, 1–3

Driessen, H., Haneef, M. I. J., Harris, G. W., Howlin, B., Khan, G. and Moss, D. S. (1989). *RESTRAIN*: restrained structure-factor least-squares refinement program for macromolecular structures. *J. Appl. Crystallogr.*, **22**, 510–516

Harris, G. W., Pickersgill, R. W., Howlin, B. and Moss, D. S. (1992). The segmented anisotropic refinement of monoclinic papain by the application of the rigid-body TLS model and comparison to bovine ribonuclease A. *Acta Crystallogr.*, **B48**, 67–75

Holbrook, S. R., Dickerson, R. E. and Kim, S. H. (1985). Anisotropic thermal-parameter refinement of the DNA dodecamer CGCGAATTCGCG by the segmented rigid-body method. *Acta Crystallogr.*, **B41**, 255–262

Holbrook, S. R. and Kim, S. H. (1984). Local mobility of nucleic acids as determined from crystallographic data. I. RNA and *B* form DNA. *J. Mol. Biol.*, **173**, 361–388

Howlin, B., Butler, S. A., Moss, D. S., Harris, G. W. and Driessen, H. P. C. (1993). *TLSANL*: TLS parameter-analysis program for segmented anisotropic refinement of macromolecular structures. *J. Appl. Crystallogr.*, **26**, 622–624

Howlin, B., Moss, D. S. and Harris, G. W. (1989). Segmented anisotropic refinement of bovine ribonuclease A by the application of the rigid-body TLS model. *Acta Crystallogr.*, **A45**, 851–861

Najmudin, S., Nalini, V., Driessen, H. P. C., Slingsby, C., Blundell, T. L., Moss, D. S. and Lindley, P. F. (1993). Structure of the bovine eye lens protein γB(γII)-crystallin at 1.47 Å. *Acta Crystallogr.*, **D49**, 223–233

Schomaker, V. and Trueblood, K. N. (1968). On the rigid-body motion of molecules in crystals. *Acta Crystallogr.*, **B24**, 63–76

Wistow, G., Turnell, B., Summers, L., Slingsby, C., Moss, D., Miller, L., Lindley, P. and Blundell, T. (1983). X-ray analysis of the eye lens protein γ-II crystallin at 1.9 Å resolution. *J. Mol. Biol.*, **170**, 175–202

Group anisotropic thermal parameter refinement of the light-harvesting complex from purple bacteria *Rhodopseudomonas acidophila*

Miroslav Z. Papiz* and Steve M. Prince#.

* CCLRC Daresbury Laboratory, Daresbury, Warrington WA4 4AD, UK, # Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, UK.

Introduction

There are several reasons why it may be preferable to generalise thermal parameter refinement to include anisotropic thermal motions. Some of these reasons are technical ones that aim to achieve the best possible refinement of all parameters by removing the bias introduced into refinement by thermal decay, sometimes anisotropic, of the structure amplitudes. Another reason, often overlooked, is the effect that dynamic processes have on the way that a structure performs its biological role. Isotropic thermal parameters are often examined for clues to the regions with greatest conformational flexibility. For example, this kind of information can give some insight into which atomic groups are most mobile within an enzyme's active site. Anisotropic thermal parameters can add richness to this information: not only can they give magnitudes to co-ordinate displacements, but also the directions in which these groups are moving. Thermal energy is important in facilitating biological action because it allows large movements of substrates, products and protein groups, and because intermediate reaction states can be driven through a potential barrier by specific thermal vibrations. The slightly different structures, so called conformational substrates[1,2], formed by motions which occur on the pico- and subpico-second timescales have an important role in the functioning of macromolecules, since the reactions from different conformational substrates may have very different rates.

Photosynthetic complexes of purple bacteria are striking examples of biological systems which are strongly influenced by thermal motions. Excited states have been found to be strongly coupled into vibronic or phonon states. These states play an important role in the relaxation of energy between excited energy states, the transfer of energy within an excitonic band of closely interacting pigments and the transfer of energy over relatively long distances, 20 to 40 Å. Most of our understanding of electronic properties of molecules rests on the Born-Oppenheimer approximation, the assumption that slow nuclear motions do not effect fast electron motions within electronic excited states. Coherent nuclear motions have been observed that are on the pico-second timescale, this is comparable to the timescale of energy transfer within and between these complexes[3,4]. Within such systems atomic motions can not be ignored; the energies that couple excited states between pairs of pigments are in the range 10 cm^{-1} to 400 cm^{-1}, well within the energy interval for phonon modes observed in protein structures.

Thermal motions between neighbouring atoms are well known to be correlated. Groups as small as 2 or 3 atoms or domains of tens of kD can have correlated motions of some kind. It seems therefore that parameterizing a full atomic anisotropic thermal parameter refinement may greatly overdetermine the problem. That this is so, is very fortunate as few protein crystallographic datasets attain those resolutions (ca. 1.2 - 0.8 Å) that are needed to achieve the required data to parameter ratio for atomic anisotropic thermal parameter refinement. The overdeterminancy has been demonstrated by the use of a small number of normal modes of vibration to obtain isotropic and anisotropic thermal parameters. A striking result of this work is that molecular motions described by 892 isotropic thermal parameters, refined in the conventional way, could be largely reproduced by only 19 thermal parameters[5]. This work shows that it is important to realise that a large part of anisotropic motion may not arise from local molecular motion.

The TLS model

The **TLS** method[6] is more suited for the refinement of anisotropic thermal parameters of protein structures. It is more in tune with the philosophy that motions are correlated, and because any number of atoms can be defined by the same set of thermal parameters the number of additional parameters to be refined is not large. The method can be used in the refinement of nearly all structures as the observation to parameter ratio is sufficiently small to allow data in the range of 2.5 to 1.5 Å to be used. The factor which determines the actual amount of data required, as will be the case for any thermal parameter refinement, is the size of the thermal effect that is being extracted from the data. For example the motions of domains, helices or large co-factors can be determined with less data than those of small rigid groups such as planar carboxylates, amides and guanidinium moieties of side chains such as Asp, Glu, Asn, Gln and Arg.

The **TLS** method, as originally described by Shomaker and Trueblood[6], has been implemented within the program *RESTRAIN* for the refinement of thermal parameters of protein structures[7,8]. The atomic form factor may be written as:

$$\mathbf{f(x, h)} = f_h \exp\left(2\pi i \mathbf{h}^{\mathbf{T}}\mathbf{x} + 2\pi^2\, \mathbf{h}^{\mathbf{T}} U\, \mathbf{h}\right) \qquad 1$$

The atomic co-ordinate **x** gives the mean position of the atom and the tensor U describes the mean-square displacements from that mean. U can be written as the usual anisotropic temperature factor elements u_{ij}, or if the atom is part of a rigid group of moving atoms then it can be expressed as;

$$U = T + A\,L\,A^T + A\,S + S^T A^T \qquad 2$$

T is the symmetric *translation tensor* that applies a translation to the whole group of atoms, L is the symmetric *libration tensor* that rotates the group about some centre of action (usually the centre of mass) and S is the asymmetric *screw tensor* which accounts for the correlation between the T and L tensors. The diagonal elements of the S tensor represent a screw like motion, the off diagonal elements represent rotations about axes which need not pass through the centre of mass. The A matrix is a function of the atomic co-ordinates **r** relative to the centre of libration. As can be seen from equation 2 the contribution of T to mean-square displacement is the same for all atoms in the group, but the contribution of L and S depends on the atomic distance from the centre of libration. Once the ***TLS*** tensors have been refined they can be decomposed into individual atomic anisotropic thermal tensor elements u_{ij} by the program *TLSANL*. In this form the structural implication of thermal parameters can be more readily visualised and quantified. In total for each rigid group of atoms only 20 additional parameters need be refined. If there are more than 3 atoms in the group there will be fewer parameters to refine than in an individual anisotropic thermal parameter refinement. The implementation of **TLS** refinement in *RESTRAIN* optionally allows the refinement of rigid group and isotropic thermal. parameters for the same atoms. The isotropic thermal parameters can be thought of as taking into account of additional local thermal motions. This is important if the rigid group is large and has many contacts which may locally modify the overall group **TLS** motion. For those atoms not part of a rigid group only individual isotropic u factors were refined. This approach resulted in only a small increase in the total number of refined parameters.

The use of non-crystallographic symmetry (NCS)

The structure of light-harvesting complex, from purple bacteria *Rhodopseudomonas acidophila*[9], is nonameric oligomer comprising 18 peptides, 27 bacteriochlorophyll *a* (Bchl *a*) molecules and 18 carotenoids molecules: although only 9 carotenoids are well ordered in the crystal structure. The crystal space group is R32, a = 120.4 Å and c = 296.3 Å with a ninefold

rotational axis superimposed on the crystallographic threefold axis. The consequence of this is that the asymmetric unit comprises only of a third of the complex and the NCS molecular copies are related by exact rotations of 40° and 80° about the *c* axis. *RESTRAIN* can not apply NCS restraints while refining **TLS** parameters. There are therefore, three independent sets of parameters which are refined for each molecule. The refinement, at a resolution of 2.5 Å, may be cross-checked using this redundancy. Generally, it is only at this resolution and better that we would start to trust thermal parameters. It would seem sensible to be wary of anisotropic refinement at this intermediate resolution and to build in checks on the accuracy of the refinement.

The **TLS** tensors can be decomposed into individual atomic anisotropic thermal tensors, u_{ij} that define ellipsoids of vibration. These are best understood by looking at the principal axes of their ellipsoids. These can be written as;

$$\hat{p}_i = \lambda_i \hat{e}_i; \qquad i = 1,3 \qquad 3$$

Where λ_i are the eigenvalues and $\hat{e}_i$ the eigenvectors of the tensor u. The first is a magnitude assigned to the principal axis vector and the second its direction. The angular error when superimposing equivalent principal axis vectors on related NCS rigid groups, can be written as:

$$\langle \cos(\varepsilon) \rangle = \frac{\sum \hat{p}_i \cdot \hat{p}_j}{\sum \lambda_i \lambda_j} \qquad 4$$

and the R-factor arising from superimposing these vectors as;

$$\mathbf{R}p = \left[\frac{\sum \left| \hat{p}_i - \hat{p}_j \right|^2}{\sum \left(\lambda_i \lambda_j \right)^2} \right]^{\frac{1}{2}} \qquad 5$$

The summations are carried out between pairs of vectors on different NCS molecules and over all principal vectors and atoms within the rigid group of atoms. Both will be sensitive to angular errors but **R***p* should also be effected by errors in magnitude.

Finally a measure of the degree of anisotropy can be obtained by calculating a quantity called ellipticity. The usual definition of this is the ratio of the largest and smallest λ's. However here we are using a slightly different expression, namely :

$$E = \left\langle \left(\frac{\lambda_l^2 + \lambda_m^2}{2\lambda_n^2} \right)^{\frac{1}{2}} \right\rangle \quad ; \quad \lambda_l \geq \lambda_m \geq \lambda_n \qquad 6$$

The quantity *E* has the property that for isotropic vibrations it is equal to one and for anisotropic vibrations it is greater than one. *E* will tend to be larger for oblate vibrations than for prolate vibrations.

Choice of TLS groups

One of the pronounced features of the data is a rapid decay of the diffraction pattern, in the crystallographic *ab* plane, to around 2.5 Å, while diffraction along the *c* axis is observed to hold up to at least 2.1 Å. The *ab* plane corresponds the membrane plane and is where in the crystal the detergent micelle surrounds the complex. Although this anisotropy does not greatly effect the

refinement it nevertheless is probably largely responsible for the higher than expected crystallographic R-factor of 21 %. Large anisotropy has the effect of apparently foreshortening bond lengths.

The groups were chosen initially on size as it was considered likely that the resolution of the data would probably prevent sensible **TLS** refinement of the smaller rigid groups of side chains. The largest rigid groups are the backbone atoms of the transmembrane helices of the α and β chains. The α chain has sufficiently large surface lying regions to define the chain in three pieces, segment 1-11, transmembrane helix 12-36 and segment 37-46. The β chain has shorter surface regions and so was given one rigid group comprising the helix running from position 7 to 36. Bchl *a* molecules have conjugated regions within the bacteriochlorin macrocycle. Although there are some torsion angles within this region it can be mostly considered as rigid group; the substitutions of phytyl, ethyl and acetyl groups were not included in the group. The carotenoid (rhodopin glucoside) molecules comprise a glucoside moiety that lies at the membrane surface and which is partially disordered with at least two conformers, and a rhodopin moiety that is highly conjugated. Only the rhodopin part was included in the rigid group. The remaining atoms were refined with individual isotropic thermal parameters.

Refinement models

Three models were considered: model 1, the control, the thermal parameters were refined isotropically; model 2 defined each NCS repeating unit as a rigid group, in this model every atom was **TLS** refined; model 3 defined **TLS** tensors for, 6 transmembrane helices and 9 surface segment chains, 9 Bchl *a* molecules and 3 carotenoids molecules. All models were refined with atomic and overall isotropic thermal parameters. For those atoms for which **TLS** thermal parameters were also refined the calculation did take into account the correlation between the isotropic and **TLS** thermal parameters. The aim of model 2 was to estimate the effect of lattice vibrations and static disorder on the refinement. Model 3 is also looking at some of the internal modes of vibration, although it is impossible to entirely separate these from the lattice modes.

Model	#1	#2	#3	Xplor[#]
R_{cryst}	20.9	20.1	18.4	22.7
R_{free}	27.0	26.1	25.5	25.3
N_{par}	11,845	11,905	12,301	3,953
Δr	0.45	0.41	0.28	0.40
;Δr^*	0.16	0.15	0.12	0.07

Table 1. R_{cryst} is the crystallographic r-factor; R_{free} the free R-factor; N_{par} is the number of refined parameters; Δr and Δr^* are the coordinate errors derived by Read[12] and Cruickshank[13] respectively. The refinement was to a target bond length geometry of 0.02 Å. The data were measured to have an Rsym of 5 % with 25,877 reflections; 98% complete at a resolution of 2.5 Å. # : for comparison isotropic refinement with Xplor[14]. This was refined with constrained NCS applied, hence the lower R_{free}, and using fewer parameters leading to a lower Δr^*: an overall anisotropic thermal parameter correction was applied during refinement (u_{11}=-0.0550, u_{22}=-0.0550, u_{33}=0.1099, u_{12}=-0.1718, u_{13}=0.0000, u_{23}=0.0000).

A significant improvement in refinement was obtained in model 2 relative to model 1.This was obtained with just 3 **TLS** tensors (60 additional parameters). Model 3 which included 24 **TLS** tensors produced a further improvement with an increase in the number of refined parameters of only 3.5 %.

Agreement on NCS

Molecular redundancy was used to check the consistency of **TLS** refinement for model 3, as outlined in an earlier section. *TLSANL* was used to transform **TLS** tensors into u tensors and their eigenvalues and eigenvectors. These were used to calculate the information summarised in table 2. The statistics were calculated from the principal axes data before and after applying NCS transformations to the principal axes vectors. These numbers should be compared for signs of correlation between thermal parameters of different NCS related molecules.

	B850a	B850b	B800	RGlu	alpha	beta
Ellipticity	1.72	1.46	1.67	2.61	1.65	1.79
U_{equ} /Å	0.46	0.49	0.55	0.69	0.49	0.66
ε^+/deg	23.4	31.3	29.9	25.4	24.5	13.1
ε /deg	39.4	45.1	44.2	50.7	37.7	35.1
R_p^+;/%	17.0	24.0	22.0	19.0	17.0	9.0
Rp;/%	29.0	33.0	32.0	36.0	25.0	26.0

Table 2. U_{equ} is the equivalent isotropic displacement factor; ε is angular error and Rp the R-factor due vector error , ε, Rp are before and ε^+,R_p^+ after transforming with NCS operators. B850a, B850b are Bchl *a's* liganded to the α and β chains respectively, B800 the Bchl *a's* absorbing at 800 nm, RGlu is rhodopin glucoside.

The statistics seem to indicate that the parameters arising from **TLS** refinement correlate reasonably well, even at the resolution of 2.5 Å. There is one warning concerning the meaningfulness of these statistics. When an atom has some of its principal axes equal in magnitude then their directions are ill defined; in other words for an isotropic atom the principal axes of an ellipsoid can be defined to point in any direction. This would tend to reduce the correlation between axes and it fits in with the observation that there is better correlation for those parts of the structure with larger ellipticity.

Thermal motion analysis

There are large voids within the unit cell (73 % of the volume) containing water, detergent.and possibly lipid. Each complex can be imagined to be a cylinder of diameter 70 Å and height 45 Å; the cylindrical belt representing the transmembrane part of the structure. Extensive contacts, mediated by solvent molecules, are made in the *c* axis direction at the cytoplasmic surfaces. In the *ab* plane there are only tenuous contacts at the edges of the periplasmic surfaces. It maybe that a large part of the anisotropy, observed in the *ab* plane, originates from static disorder although the detergent belt maybe sufficiently rigid to transmit vibrations between complexes. What ever their origins these lattice effects are only of crystallographic interest and tell us nothing of the internal modes of motion.

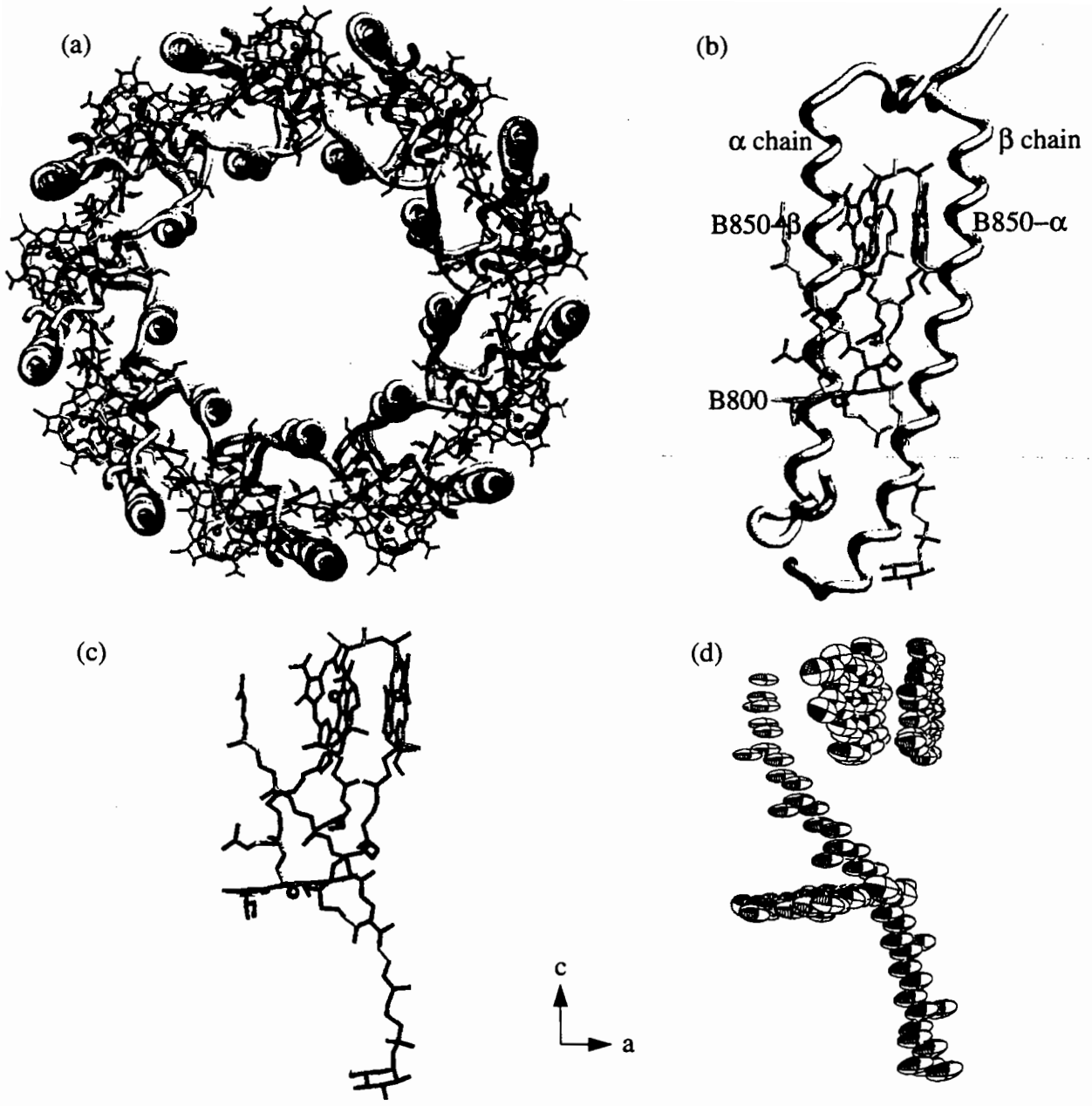


Figure 1 : (a) Light–harvesting complex viewed down the crystallographic c axis which is the crystallographic threefold and NCS ninefold axes; (b) The NCS building unit, viewed along the membrane direction (ab plane), comprising two peptides, 3 bacteriochlorophylls and one carotenoid molecule, there are three of these in the asymmetric unit; (c) The chromaphores of the NCS unit and (d) the equivalent ORTEP[10] display showing thermal ellipsoids for the TLS refined chromaphore atoms (a)–(c) were produced by the program SETOR.[11]

The thermal ellipsoids (Fig 1d) and the improvements observed in the refinement of models 2 and 3 indicate that there are local molecular anisotropic motions superimposed on crystallographically related lattice motions. These two effects are difficult to separate, however it is clear that the lattice motions are almost entirely in the membrane plane and that the anisotropic thermal motions in the *c* direction are mostly detemined by local molecular motions. It is therefore interesting to look at the angular displacement, from the *c* axis, of the principal axes that are approximately in the direction of the *c* axis. This is also generally the direction of the smallest thermal motion and gives some indication of the systematic trends in thermal motion arising from mainly internal anisotropic displacements.

In Fig 2b-2g are displayed the angular displacements ϕ of the minor principal axes from the crystallographic *c* axis. Because there are three copies of each **TLS** group it is possible to calculate the standard deviations of these angles: it should be noted that because the *c* axis is also the NCS axis the application of this symmetry leaves the angle unchanged.

Protein anisotropy

The mean isotropic thermal displacement parameter U_{equ} as a function of residue number (Fig 2a), indicates a featureless flat distribution of thermal parameters with a slight hint of an increase at the ends of the transmembrane helices. The ellipticity E however shows some interesting trends: the surface helices of the α chain have the least anisotropy while the transmembrane helices are significantly higher in anisotropy. The distribution of E is different for the two chains, whereas the α chain transmembrane helices have only slight increases of anisotropy at the helix ends, the β chain transmembrane helices start with high ellipticities of around 2.3, at the N terminus of the helix, and pass through a minimum of around 1.6 near residue 28 then rise slightly to 1.7 at the C terminus of the helix. The minimum is found near to the Bchl *a* liganding residues α-His31 and β-His30 and shifted a little towards the intertwined Bchl *a* phytyl chains. The distribution of angles with residue number also show significant differences (Fig 2b,2c), whereas for the α chain transmembrane helix the minor axis remains close to the *c* axis, starting at 14° at the N terminus and finishing at 4° at the C terminus, the β chain transmembrane helix remains at around 25° for a large length of the helix then rapidly drops down to 7° near to the C terminus. For both helices the minimum angle is on the liganding His residues α-His31 and β-His 30. Another interesting feature can be seen for the β chain helix which shows a 5° variation of this angle on the outside compared to the inside surface of the helix. This can be seen as a periodic variation every 3 to 4 residues. An interesting observation is that when isotropic thermal parameters, refined conventionally, are plotted against residue number they resemble the ellipticity curve in figure 2a rather than U_{equ} which illustrates the poor modelling of these parameters by conventional isotropic B refinement.

Rhodopin Glucoside

The carotenoid molecule exhibits the same variation as the transmembrane helices. Beginning at an angle of 40° near the cytoplasmic membrane surface and then half way along its length rapidly falling off to 10° at the B850 pigments. The anisotropy of the atoms close to the cytoplasmic surface is the largest of all the atoms in the structure as judged from the ellipticity. This is due as much to the minor principal axes being smaller than average as it is to the maximum principal axes being larger. Large anisotropy is also observed as a disorder of the glucoside head groups. For this reason they were not included in the **TLS** refinement. The spikes in the angular distribution (Fig 2g) correspond to the positions of the methyl groups decorating the carotenoid chain. This perhaps indicates the kind of small detail that is observed.

Bacteriochlorophylls

The size of ϕ for the bacteriochlorophylls liganded to the α chain (B850-α) is on average less than 20° and is similar in magnitude to that observed for the α chain helix. The standard deviations for the three kinds of Bchl *a* are around 7° and for the B850-α ϕ angles are only twice the σ. The ϕangles of B850-β are around 30° and again have similar angular magnitude to the β chain that the chromaphore is liganded to. The B800 molecules have the largest variations of 10° to 48°. The large angles are clustered amongst groups of atoms. The first cluster corresponds to the acetyl and methyl groups on the same cycle that point into the detergent micelle, the second to the cycle connecting to the phytyl chain and the third to part of the double cycle including the carbonyl. All of the Bchl *a* molecules have puckered conformations with torsional angles deviating by 5° to 8° from planarity. It maybe that there is a greater degree of

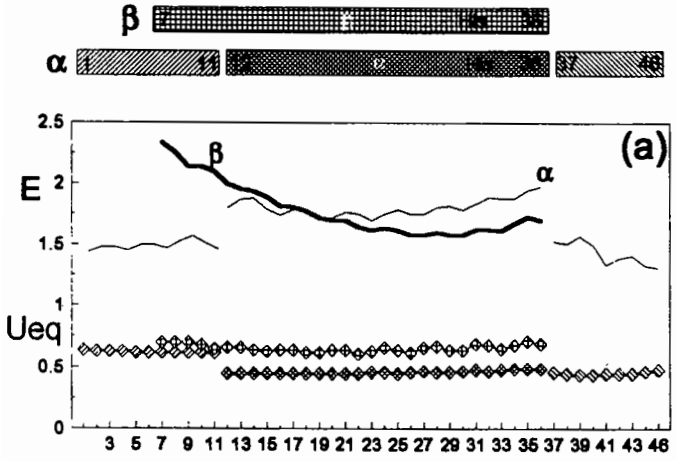
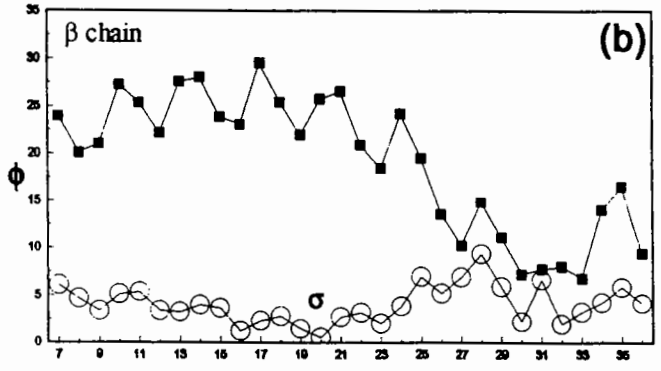
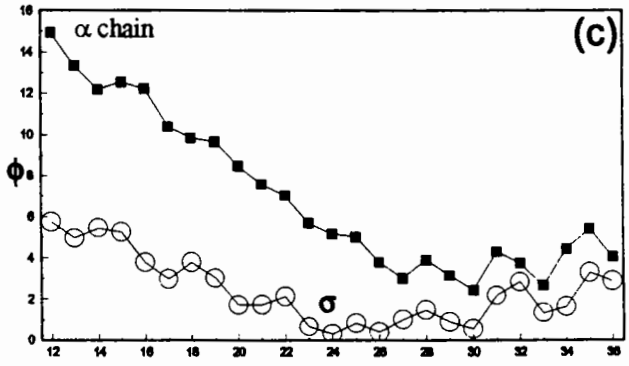
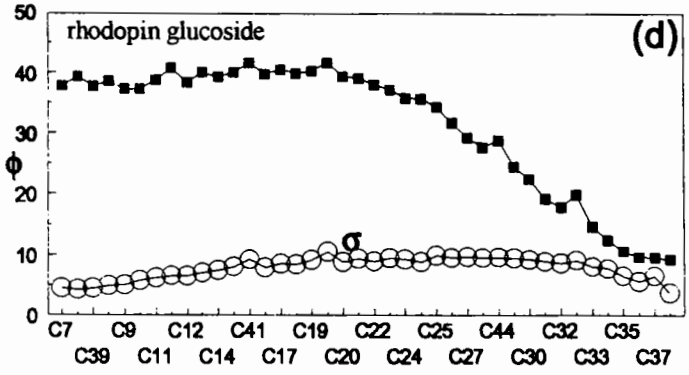
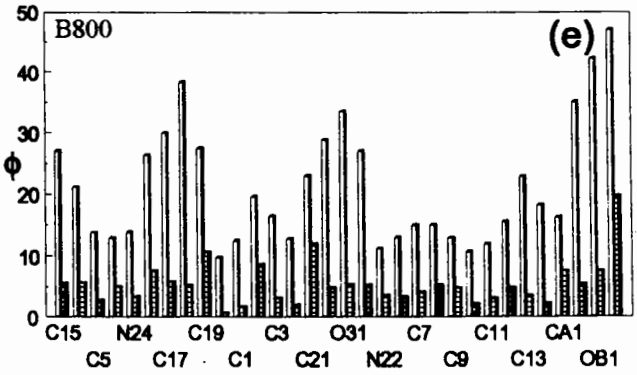


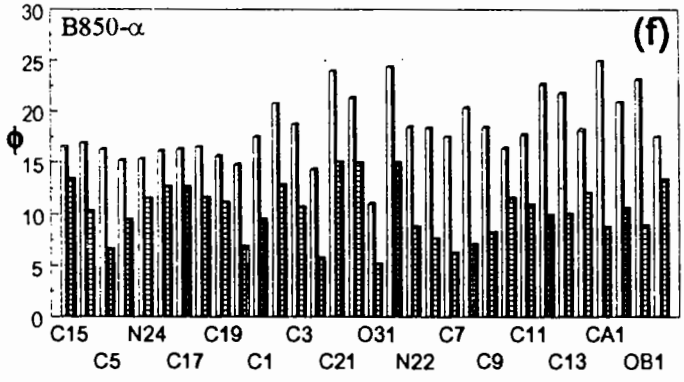
Figure 2. (a) The ellipticity E and equivalent isotropic displacement factor *Ueq* of the α and β chain Cα atoms. Mean angular displacement <ϕ> of the smallest principal axes of the α (c) and β (b) chain Cα atoms to the crystalloagraphic c axis with respect to residue number. (d-g) <ϕ> vs. atom name for rhodopin glucoside. For (d) names are approximately correlated to the distance across the membrane. (e-g) names are not neighbours in a straight forward way, in these cases <ϕ> is depicted by light bars. Standard deviations of <ϕ>, calculated from NCS redundant copies, are shown as open cicles in (b-d) and as darker bars in (e-g).

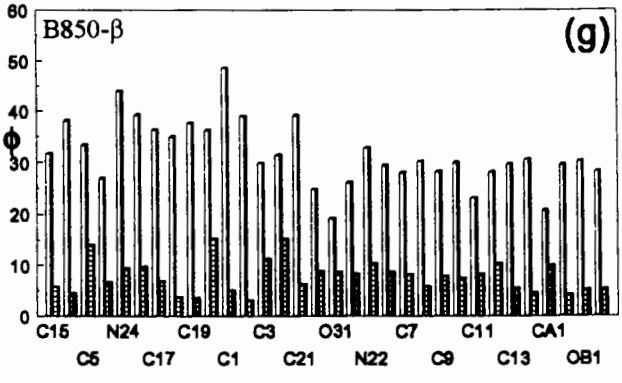












flexibility in the B800 than in the two B850's. It is possible to see why this maybe so as the B850's are in close van der Waals contact with one another, forming a rigid ring structure whereas B800's are situated between β chains and partially within the micelle.

Summary

The refinement of LH2 improved when **TLS** parameters were included as judged by R_{crys}, R_{free}, co-ordinate errors[12,13] and the relative noise levels observed in the 2Fo-Fc maps. The improvement in refinement did not result in any major reassignment of electron density: the structure at the final stage of isotropic thermal parameter refinement was well defined and molecular redundancy produced a map with electron density resembling that of one at a higher resolution.

The benefit of **TLS** refinement is to provide additional information about the motion of various parts of the structure. This refinement has intrinsically more detail about motion than conventional B factor refinement. The quality of anisotropic parameters is good, even at 2.5 Å resolution, although higher resolution would be required to refine smaller **TLS** groups, such as planar side chains of His, Phe and Trp. This kind of analysis throws up interesting questions concerning the role that modes of vibration have in facilitating biological function. In the case of LH2 it is interesting that the largest principal axes of the transmembrane helix backbone atoms seem to be optimally arranged to apply the biggest modes of vibration into the B850 rings.

Anisotropic information is available even at lower resolutions providing the anisotropy is large enough and the groups being refined have sufficient numbers of atoms. At the kind of resolution to which most structures are refined (ca. 1.8 Å) it would be expected that smaller rigid groups such as those of side chains could be refined and that it need not be the case that anisotropic refinement should only be considered at atomic resolution.

References

[1] Frauenfelder, H., Sligar, S.G. and Wolynes, P.G. Science, 254 (1991), 1598

[2] Ansari, A., Berendzen, J., Browne, S.F., Frauenfelder, H., Iben, I.E.T., Sauke, T.B., Shyamsunder E. and Young R.D. Proc. Natl. Acad. Sci. US, 82 (1985), 5000

[3] Vos, M.H., Rappaport, F., Lambry, J-C., Breton J. and Martin, J-L. Nature,363, (1993), 320

[4] Chachisvilis, M., Pullertis, T., Jones, M.R., Hunter, C.N. and Sundström, V. Chem. Phys. Lett., 224 (1994), 345

[5] Diamond, R Acta Cryst., A46 (1990), 425

[6] Schomaker, V. and Trueblood, K. N. Acta Cryst, B24 (1968), 63

[7] Driessen, H., Haneef, M.I.J., Harris G. W., Howlin, B., Khan, G. and Moss D.S. J. Appl. Cryst., 22 (1989), 510

[8] Howlin, B., Moss, D. S. and Harris G.W. Acta Cryst, A45 (1989), 851

[9] McDermott, G., Prince, S. M., Freer, A. A., Hawthornthwaite-Lawless, A. M., Papiz, M. Z., Cogdell, R. J. and Isaacs, N. W. Nature, 374 (1995), 517

[10] Johnson, C. K. ORTEP-II (1976) Report ORNL-5138 Oak Ridge National Laboratory, Tennessee, USA

[11] Evans, S. V. J. Mol. Graphics 11 (1993), 134

[12] Read, R. J. (1986), Acta Cryst, A42, 140

[13] Cruickshank, D. W. J. Proceedings of CCP4 Study Weekend (this issue)

[14] Brunger, A. T. X-PLOR Manual (Yale University., New Haven, CO, 1990).

Is Refinement from a Random Start Possible? - given diffraction data to medium resolution -

Piet Gros, Dept. of Crystal and Structural Chemistry,
Bijvoet Centre for Biomolecular Research,
University Utrecht, Padualaan 8, 3584 CH Utrecht, The Netherlands
(e-mail: gros@chem.ruu.nl)

Introduction

The last step in a protein crystal-structure determination is refinement of the structure, where the *x*,*y*,*z*-coordinates and typically an isotropic *B*-factor of the atoms are optimized. The target function used in these optimizations relies both on the measured diffraction data and on geometrical data of protein structures, like observed bond lengths, bond angles, dihedral angles, *etc.* (see e.g. Hendrickson, 1985). The geometrical restraints are needed, because the diffraction data alone (when the resolution is less than atomic, $d > 1.2$ Å) doesn't overdetermine the system sufficiently. Unfortunately this target function, that combines structural restraints and diffraction-amplitude restraints, is complex and has an immense number of local minima. The best search methods developed so far still have a limited convergence radius (e.g. 1.7 Å rms for backbone atoms as reported by Rice & Brünger, 1994). Thus, a reasonably good starting model must be obtained based on phase information from experimental methods, such as isomorphous replacement, molecular replacement or multiple-anomalous dispersion. In our research we address the question whether (and how) the refinement methodology can be applied to a random starting model. Here, we discuss some of the general aspects of such a method of phasing by *ab initio* modelling.

Ab initio phasing by Direct Methods (rev. Woolfson, 1987) is done routinely nowadays to solve crystal structures of less than ca. 300 atoms. For *ab initio* structure determination of proteins various approaches are being developed. Probabilistic theories and optimization techniques from Direct Methods are extended and applied to phasing of protein data sets. (Bricogne, 1993; DeTitta *et al.* 1994; Sheldrick & Gould, 1995). However, these methods rely on diffraction data to atomic resolution ($d \leq 1.2$ Å), which is rarely the case when a structure needs to be solved. Other approaches, more related to the method presented here, start from a real-space model consisting of a limited number of spheres (Subbiah, 1992; Lunin *et al.*, 1995). These methods so far have yielded very low-resolution phases. Thus, reliable phasing in an *ab initio* manner (*i.e.* from a single diffraction data set) of protein data to ca. 2 to 3 Å resolution has remained an elusive goal so far.

A fundamental question arises in the application of structure optimization to *ab initio* structure determination, as we propose. Is the desired protein-crystal structure fully defined by the geometrical protein-structure restraints and the diffraction data to medium (2 to 3 Å) resolution; or, reformulated, does the crystal structure correspond to the global minimum of the target function used? If this is indeed the case, the phase problem is reduced to finding the *minimum minimorum* of the target function. We assume this to be true, and attempt to solve the given search problem.

In principle, one can start an *ab initio* structure optimization with a polypeptide chain of the correct sequence placed in the cell with a random conformation (or any other conformation unrelated to the answer). Given our hypothesis the answer is fully defined by the geometrical and the diffraction (with $d \leq 3$ Å) information. However, the answer will be extremely hard to find. For example, wrong (e.g. reverse) tracing of the model through electron density cannot be corrected by gradient-driven procedures, that are used commonly in refinements like energy minimization or simulated annealing. Perhaps, new methods from artificial intelligence may be designed to deal with these manifold diverse and complex situations. The other extreme in structure optimization from a random start is using a model consisting of loose atoms. In this case the search may be expected to be very efficient. However, the geometrical restraints are not defined anymore. Thus, the desired protein-crystal structure cannot be found, because it is not sufficiently defined.

Our approach of *ab initio* modelling (called AIM[1]) is to use loose and equal atoms, thus facilitating efficient searching, and to redefine geometrical information. In (conventional) structure refinement the geometrical information is defined for a given topology of a protein molecule. This implies that each atom in a protein structure is uniquely identified. In our model all atoms are of identical type, and thus the conventional restraints cannot be applied. Therefore, the geometrical information must be redefined, such that it becomes applicable to loose and equal atoms. We attempt to redefine as much information as possible, since this will be required to define the desired answer (*i.e.* the true crystal structure) sufficiently. However, the information must be defined in such a way to minimize potential search barriers in the optimization process.

In practice, a number of problems appear in this *ab initio* structure optimization, that are not present or not critical when refining a near-correct protein crystal structure. At low (infinite to ca. 8 Å) resolution the contribution of the bulk-solvent region is large. Since, optimization starting from a random model implies starting at infinitely low resolution, the bulk-solvent contribution must be properly taken into account. A related problem is data completeness at low resolution. To model a structure at low resolution, the data at low resolution must be present. A third problem in our optimization process concerns scaling of the structure amplitudes. (The structure-amplitude restraint used in our optimization is the commonly used restraint: $E = \Sigma\ (F_{obs} - k\ F_{model})^2$, which requires scaling by a linear scale factor k of the model derived amplitudes F_{model} with respect to the observed amplitudes F_{obs}.) The observed amplitudes are measured from a crystal diffracting to high resolution, at least beyond 3 Å resolution. The fall-off with resolution of the observed data corresponds to an overall B-factor that is (much) lower than the B-factors obtained for the low-resolution models in the optimization process. A straightforward linear scaling is not appropriate in this case. Thus, even though sufficient geometrical information may be defined, the optimization of a random starting model will only succeed when all (or most) general aspects of the diffraction data are taken into account.

The Model

The model used in the AIM optimization process consists of all expected non-hydrogen scatterers in the asymmetric unit without the atoms filling the bulk-solvent region, which is modelled by a continuum model. The use of all expected atoms is in contrast to the procedures

---

[1] A full account of the method and the force field used will be given elsewhere (Gros, Brünger, van Gunsteren & Kroon, in preparation).

described by Subbiah (1992) and Lunin *et al.* (1995). In these methods only a few spheres are used to model a protein, thus the low-resolution aspects can be modelled by only a few parameters. However, phase extension from the few initially phased reflections has not yet been demonstrated for these approaches. We use all atoms, thus no limitation in resolution exists *a priori*. Furthermore, the use of an atomic model allows direct usage of the observed atomic distributions in known protein structures. The disadvantage is primarily two-fold: *i.* the high parameter-to-observation ratio; and, *ii.* the large amount of cpu cost. The very high parameter-to-observation ratio may lead to overfitting of the data. Therefore, precaution must be taken not to overinterpret the resulting atomic configuration or "structure". This means, that given an (intermediate) solution with phases that are only valid at low resolution, we must not interpret atoms individually, but must interpret the globular features of the distribution corresponding to the appropriate resolution of this solution. This approach of using all atoms allows in principle for a gradual shaping of the atomic configuration from low to high resolution.

Since all atoms in our model are of equal type, all atoms have identical scattering; we choose carbon form factors. The atomic temperature factors are estimated from the atomic configuration obtained in the optimization (see section "Atomic *B*-factors"). Therefore, the parameters of the atomic model, which are being optimized, are the *x,y,z*-coordinates of the individual atoms.

Besides the atomic model, a bulk-solvent model is used for the bulk-solvent contribution. In between structure optimization cycles a limited number of cycles of solvent flattening are performed (Wang, 1985). We use the implementation as developed by Roberts & Brünger (1995). The resulting estimate of the solvent contribution $\boldsymbol{F}_{solvent}$ is added to $\boldsymbol{F}_{calc}$ yielding the complete model $\boldsymbol{F}_{model}$.

Geometrical Information

In our procedure we consider a protein structure to consist of equal and loose atoms, thus the molecule is modelled by a fluid of atoms. In analogy to simulations of simple fluids, global features of atomic configurations are described by two aspects: the atomic density and the radial distribution of atoms.

The atomic density of a protein is easily calculated from the average number of atoms per volume; the average volume for a (non-hydrogen) protein atom is ca. 16 $Å^3$. Thus, the simplest of models might consist of hard spheres with ca. 1.4 Å radii. However, this assumes a closest packing of spheres, which is obviously incorrect.

Radial distribution histograms of non-hydrogen atoms are calculated for a few proteins, see Figure 1. The average profile is shown in Figure 1a. Two individual cases are given in Figures 1b and 1c. Clearly, the bond distance (*i* to *i*+1) and bond-angle distance (*i* to *i*+2) are the most prominent features in this distribution. Beyond 3 Å the various contributions (*i* to *i*+3, *i* to *i*+4, etc.) yield overlapping peaks in the histogram. These contributions are separable by considering radial distributions of atoms connected by a fixed number of bonds. From the information in these histograms we have derived interatomic distance restraints. Furthermore, additional restraints are required to restrain atomic density. For example, the number of atoms within bonding distance of any given atom must be restrained to maximally 3. The restraints derived in this way are applicable to identical atoms, because the structure analysis was performed for all (non-hydrogen) atoms in protein structures irrespective of the atom types. The derivation of these restraints, or interaction potential functions, will be described elsewhere (Gros, Brünger, van Gunsteren & Kroon, in preparation).

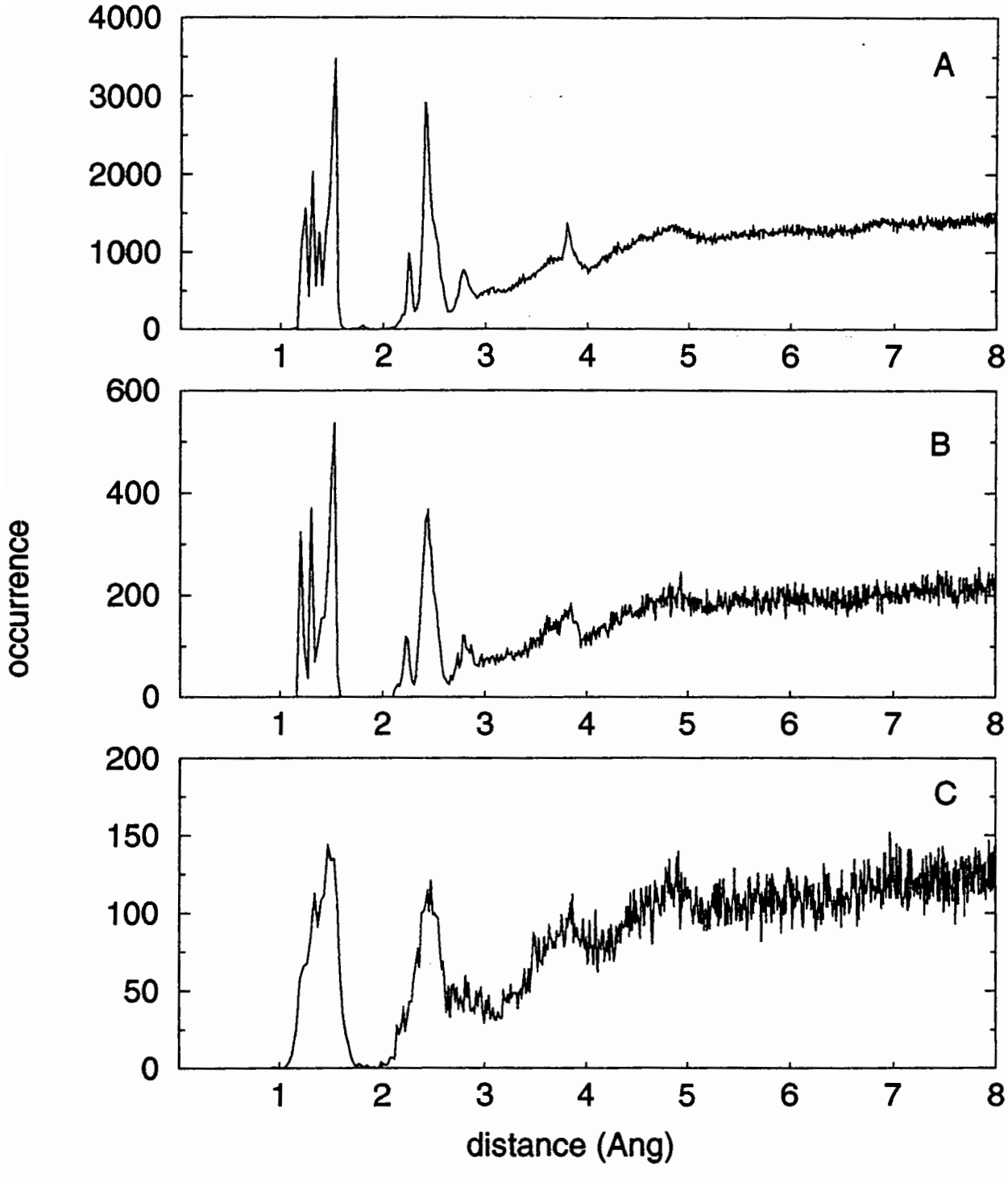


Figure 1, Radial distribution histograms for non-hydrogen atoms in protein structures: a) summed histogram for nine structures from the Protein Data Bank (Bernstein *et al.*, 1977): 1MBO, 1REI, 2ACT, 2AZA, 2PAB, 2PRK, 2PTN, 3TLN and 5CPA; b) an example of an individual structure, 3TEC, displaying a similar pattern as the summed histogram; and c) an observed deviation, 3DFR, from the average pattern: the features in the radial distribution histogram are less resolved. This observation is in agreement with the large number of distorted geometries as listed by PROCHECK (Laskowski *et al.*, 1993).

Atomic B-factors

For a reliable optimization of atomic *B*-factors a near-correct model at ca. 2.5 Å must be available. Clearly, this is not the case in our application. Moreover, an overall *B*-factor suffices for a near-correct model at low resolution (ca. 3 Å); at intermediate resolutions (2.5 to 3 Å) *B*-factor restraints should be applied. In our optimization, we estimate the atomic *B*-factors. Figure 2 shows that a correlation exists between the number of neighbouring atoms in a structure and the refined atomic *B*-factors. Based on this observation, we calculate atomic *B*-factors using an exponential function from the number of neighbouring atoms as observed in the atomic configuration. These approximate *B*-factors serve two goals: *i.* *B*-factors are estimated, and thus the number of parameters can be reduced from 4 to 3 per atom; and, *ii.* the estimated *B*-factors reflect the local atomic density, and thus atoms in dense regions (likely "protein" regions) result in relatively sharp peaks in the electron density, whereas atoms residing in sparse regions (likely "solvent" regions) are given very broad peaks.

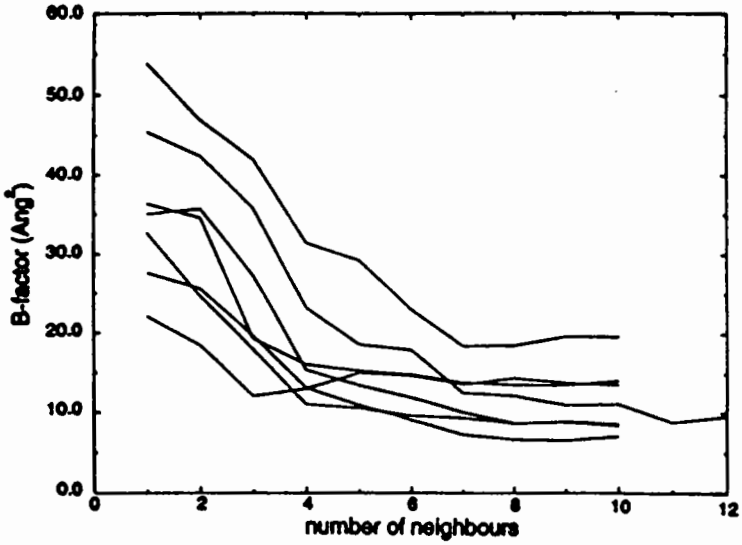


Figure 2, Atomic *B*-factors as a function of the number of neighbouring atoms; lines shown are obtained from 1MBO, 1REI, 2ACT, 2AZA, 2PRK, 2PTN, 3TLN and 5CPA. The radius used for counting the number of neighbouring atoms was 2.65 Å.

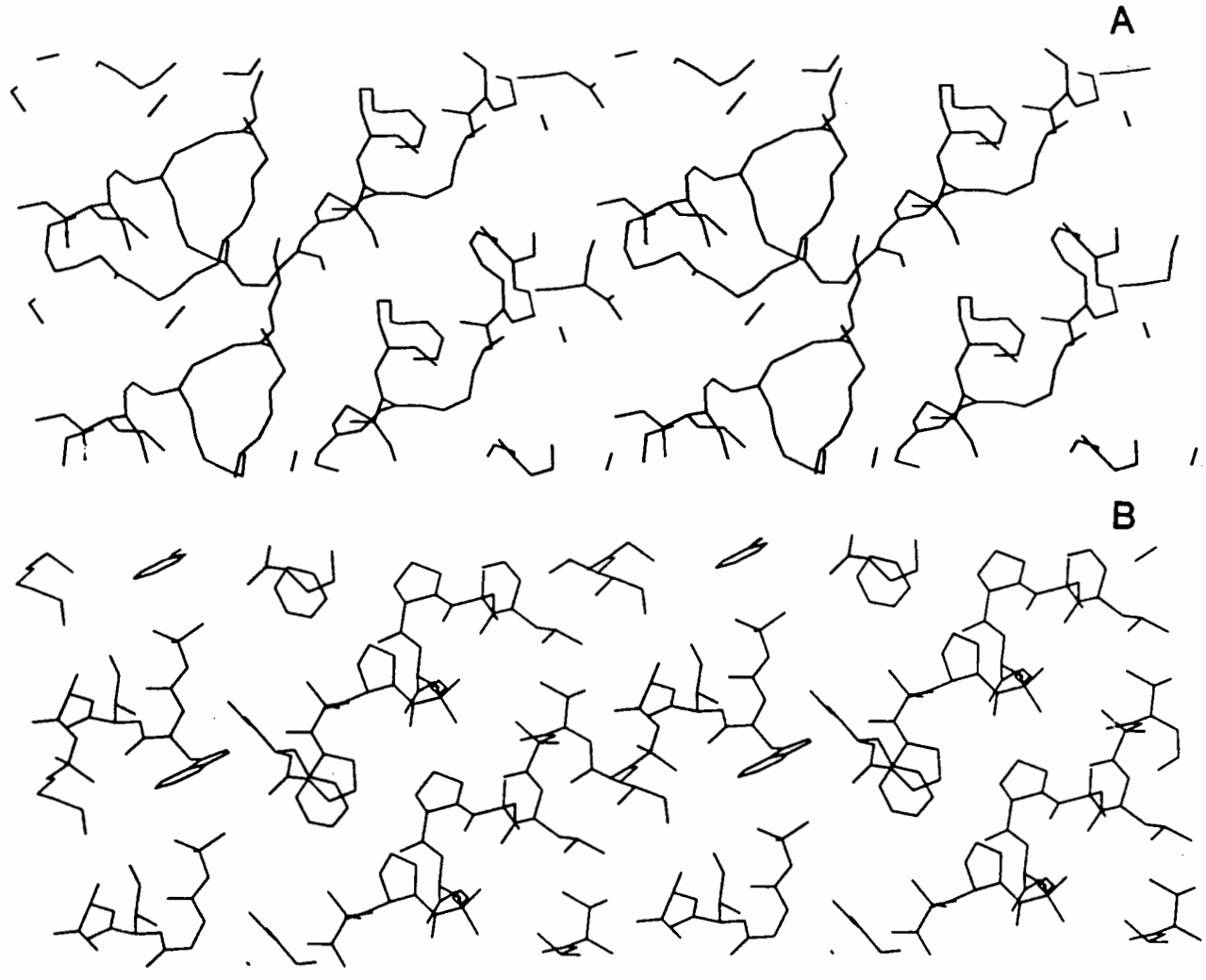
Scaling of Amplitudes

Our initial model consists of atoms distributed randomly in the unit cell (or asymmetric unit). A constraint of 2 Å minimal interatomic distance is applied during the random positioning. This model is correct only at infinitely, to very low, resolution. The resulting amplitudes displays a strong fall-off with resolution, because large *B*-factors are assigned to the dispersed atoms. These model-derived amplitudes have to be scaled linearly with respect to the observed amplitudes. To account for the difference in overall *B*-factor between the two amplitude sets, the data is first scaled non-linearly.This yields a linear scale factor *k* and a resolution-dependent scale factor *B*. Subsequently, only the linear scale factor *k* is applied. This procedures avoids over-estimation of the linear scale factor *k*.

Results

In the previous sections we described some of the general aspects of *ab initio* structure optimization. Based on the considerations given above we have implemented a method for phasing by *ab initio* modelling (AIM) in the program X-PLOR (Brünger, 1993). In this section, results are given of the AIM-optimization starting from a random model.

The test case shown concerns a hexadecapeptide zervamicin IIA analog (Karle *et al.*, 1987). The space group of the crystal is $P1$ with $a = 9.09$ Å, $b = 10.41$ Å, $c = 28.19$ Å, $\alpha = 86.13°$, $\beta = 87.90°$ and $\gamma = 89.27°$. The structure optimization was performed with measured data up to 2.5 Å resolution; data were kindly provided by Dr Isabella Karle. 126 atoms were included in the model; no bulk-solvent model was included. The initial model was obtained by placing the atoms randomly in the unit cell with minimal interatomic distance of 2 Å. AIM-optimization was performed for 250 cycles, each cycle consisting of scaling (scale factor k), 1000 steps Molecular Dynamics and 50 steps of Energy Minimization. The AIM-parameter set "wp72" was used (Gros, Brünger, van Gunsteren & Kroon, in preparation). The Molecular Dynamics calculations were performed at constant temperature ($T = 300$K).



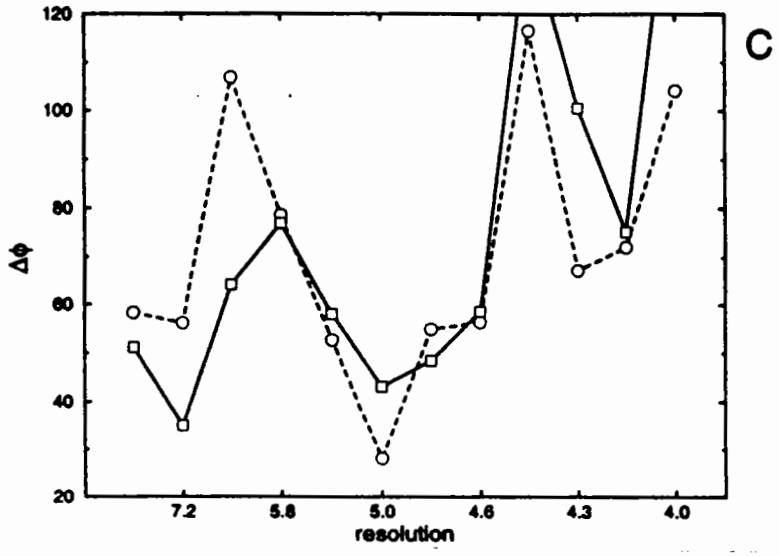


Figure 3, Comparison of the refined crystal structure of a 16-residue zervamicin IIA analog peptide with the structure obtained from the *ab initio* structure optimization procedure AIM: a) bonded structure obtained after 250 cycles AIM optimization; and, b) refined crystal structure (Karle *et al.*, 1987). 2 by 2 cells are shown for clarity. c) Phase differences between the refined structure and the structure obtained after AIM-optimization: unweighted phase differences (dashed lines) and weighted phase differences (solid lines). The weigthed phase differences were computed from the "vector *R*-factor" ($\Delta\phi = 2\ \sin^{-1}(R_V/2)$ with $R_V = \Sigma\ |F_{true} - k\ F_{model}|\ /\Sigma\ |kF_{model}|$; where F_{true} consists of the observed structure factor amplitudes and the phases from the refined crystal structure and F_{model} are the structure factors obtained in AIM.) Prior to comparison the common enantiomer and origin of the two structures was determined. The translation vector for common origin selection was computed using the phased translation function (Read & Schierbeek, 1988).

The data displayed in Figure 3, shows that the model obtained after AIM-optimization correlates at low resolution with the refined crystal structure. Starting from a random collection of atoms the correct overall shape of the molecule is obtained. However, the structure obtained from AIM is clearly not correct in detail. Phase analysis (Figure 3c) shows that phase information up to ca. 4.5 Å resolution is contained in the AIM model. Similar results have been obtained for other oligopeptide data sets.

Concluding Remarks

Starting from a random model the *ab initio* optimization procedure AIM yields correct but largely inaccurate models. Phase analysis shows that the model is correct to approx. 4.5 Å resolution. This indicates that at low resolution the correct (desired) minimum is found. Thus, our starting hypothesis "the true crystal structure corresponds to the global minimum of the target function" is validated. However, the models are inaccurate, corresponding to large phase errors for reflections beyond ca. 4.5 Å resolution. So, it appears we have not yet introduced sufficient restraints to define the answer to the desired resolution of $d \leq 3$ Å.

Acknowledgements

I gratefully acknowledge Profs Wilfred F. van Gunsteren (ETH-Zürich, Switzerland), Axel T. Brünger (Yale University, New Haven, USA) and Jan Kroon (Utrecht University, The Netherlands) for support and stimulating discussions. I thank Dr Isabella Karle (Naval Research Laboratory, Washington, USA) for generously supplying diffraction data. This work is supported by the Netherlands Foundation for Chemical Research (SON) with financial aid from the Netherlands Organization for Scientific Research (NWO).

References

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F. Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., "The Protein Data Bank: A computer based archival file for macromolecular structures", *J. Molecular Biology* **112**, 535-542 (1977).

Bricogne, G., "Direct phase determination by entropy maximization and likelihood ranking: status report and perspectives", *Acta Crystallographica* **D49**, 37-60 (1993).

Brünger, A.T., *X-PLOR Version 3.1 Manual*, Yale University, New Haven, CT, USA (1993)

DeTitta, G.T., Weeks, C.M., Thuman, P., Miller, R. and Hauptman, H.A., "Structure solution by minimal function phase refinement and Fourier filtering: theoretical basis", *Acta Crystallographica* **A50**, 203-210 (1994).

Hendrickson, W.A.,"Stereochemically restrained refinement of macromolecular structures", *Methods in Enzymology* **115**, 252-270 (1985).

Karle, I.L., Flippen-Anderson, J., Sukumar, M. and Balaram, P., "Conformation of a 16-residue zervamicin IIA analog peptide containing three different structural features: 3_{10}-helix, α-helix, and β-bend ribbon.", *Proc. Natl. Acad. Sci. USA* **84**, 5087-5091 (1987).

Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M., "PROCHECK:a program to check stereochemical quality of protein structures", *J. Applied Crystallography* **26**, 283-291 (1993).

Lunin, V.Y., Lunina, N.L., Petrova, T.E., Vernoslova, E.A., Urzhumtsev, A.G. and Podjarny, A.D., "On the *ab initio* solution of the phase problem for macromolecules at very low resolution: the few atoms model method", *Acta Crystallographica* **D51**, 896-903 (1995).

Read R.J. and Schierbeek, A.J.,"A phased translation function", *J. Applied Crystallography* **21**, 490-495 (1988).

Rice, L.M. and Brünger, A.T., "Torsion-angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement", *Proteins* **19**, 277-290 (1994).

Roberts, A.L.U. and Brünger, A.T., "Phase improvement by cross-validated density modification", *Acta Crystallographica* **D51**, 990-1002 (1995)

Sheldrick, G.M. and Gould, R.O., "Structure solution by iterative peaklist optimization and tangent expansion in space group P1", *Acta Crystallographica* **B51**, 423-431 (1995).

Subbiah, S., "Low-resolution real-space envelopes: an approach to the *ab initio* macromolecular phase problem", *Science* **252**, 128-131 (1992).

Wang, B.C., "Resolution of phase ambiguity in macromolecular crystallography", *Methods in Enzymology* **115**, 90-111 (1985).

Woolfson, M.M., "Direct Methods - from birth to maturity", *Acta Crystallographica* **A43**, 593-612 (1987).

What can we Learn from Anisotropic Temperature Factors ?

Thomas R. Schneider

European Molecular Biology Laboratory (EMBL) c/o DESY, Notkestr. 85, 22603 Hamburg

Introduction [1]

Brighter X-ray sources, sensitive area detectors and the use of cryogenic techniques enable the collection of atomic resolution (i.e. $d_{min} < 1.2$ Å [1]) data on an increasing number of protein crystals [2]. Such data provide a sound basis for the refinement of models with more parameters than previously acceptable. One possible approach is to replace the individual isotropic B-factor model by an anisotropic approximation. Besides resulting in much clearer electron density maps and frequently giving crystallographic R-values below 10 %, refined anisotropic temperature factors provide an interesting new piece of information: the direction dependence of the atomic mean square displacements. Some basic concepts to access this information and their application to models of protein molecules in the crystalline state will be described in this article. More comprehensive introductions can be found in the articles by Dunitz et al. [3, 4] and Trueblood [5] and in the book by Willis & Pryor [6].

Basics

The subject of an X-ray diffraction experiment and the subsequent analysis is not a single molecule at an instantaneous point in time. Instead the data collected on a crystal correspond to an ensemble of zillions of molecules observed for a time which is very long compared to typical time-scales of molecular motions. Therefore the result of the experiment, the so-called 'crystal-structure' of a molecule, does not provide a sharp position for each atom but, due to the time and space averages, a three dimensional probability density that is characterized by a mean position and some quantity that is related to the mean displacement of the particular atom from this mean position.

In many cases the atomic PD is approximated by a spherical Gaussian centered at the mean position of the atom. The width of this Gaussian corresponds to the MSD $\langle u^2 \rangle$ of the respective atom, which in turn is related to the isotropic B- or temperature factor by $B = 8\pi^2 \langle u^2 \rangle$. Atomic positions and B-factors are adjusted by optimizing the agreement between observed structure factors (F_{obs}) and structure factors calculated on an actual model (F_{calc}). In the isotropic B-factor approximation the calculated structure factor is of the form:

[1] Abbreviations: 'PD' probability density; 'ESD' estimated standard deviation; 'MSD' mean square displacement; 'ADP' anisotropic displacement parameter; 'TPP' Triphenylphosphine oxide

$$F_{calc}(\vec{h}) = \sum_j f_j exp\left(-\tfrac{1}{4}B_j \vec{h}^t \vec{h}\right) exp\left(2\pi i \vec{h}^t \vec{x_j}\right), \quad (1)$$

i.e. three coordinates $\vec{x_j} = (x_j, y_j, z_j)$ and one isotropic B-Factor B_j are refined for each atom j (f_j is the respective scattering factor, $\vec{h}$ a reciprocal lattice vector). The overall mean displacement of an atom originates from several sources:

- different conformations in different unit cells ('internal static disorder')
- vibration or dynamic transitions within molecules ('internal dynamic disorder')
- lattice defects
- lattice vibrations (acoustical phonons)

From the variety of these contributions it is clear that an isotropic description of mean atomic displacements is only a very crude approximation. In contrast to small molecules the refinement of more detailed models by introducing more parameters into the refinement process is unfortunately not supported by the number and quality of the X-ray data for most macromolecules. The situation is different if atomic resolution data are available. Due to the large number of observables (typically on the order of 30 to 50 reflections per non-hydrogen atom) the isotropic model for the shape of the PD (1 parameter corresponding to the radius) can be upgraded to an anisotropic model (6 parameters to describe the orientation and the elongation of an ellipsoid (Fig. 1)). The 6 parameters

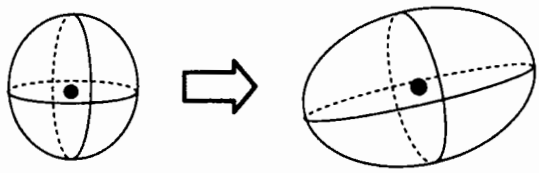


Figure 1: In the anisotropic case the PD for an atom is approximated by an ellipsoidal instead of a spherical distribution.

for the anisotropic description of the PD of an atom can be written as a symmetric matrix $\mathbf{U}_j$ which enters the structure factor equation in a way very similar to the isotropic B-factor:

$$F_{calc}(\vec{h}) = \sum_j f_j exp\left(-2\pi^2 \vec{h}^t \mathbf{U}_j \vec{h}\right) exp\left(2\pi i \vec{h}^t \vec{x_j}\right), \quad (2)$$

resulting in 6+3=9 instead of 1+3=4 (eq. 1) parameters to be refined per atom. The elements of the matrix $\mathbf{U}$ are referred to as anisotropic displacement parameters (ADP's).

In order to obtain meaningful results from a refinement of ADP's for a macromolecule in most cases restraints have to be employed to supplement the experimental data. The

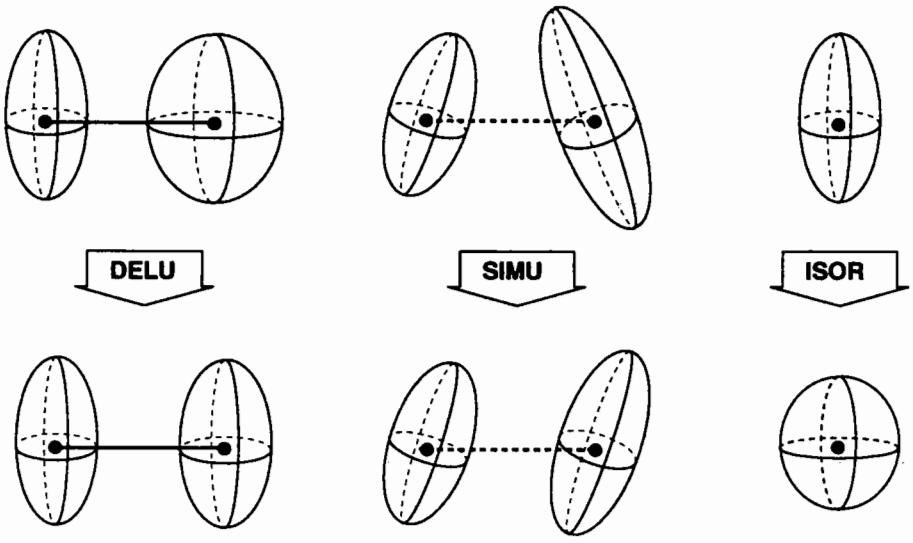


Figure 2: Restraints for ADP's available in SHELXL93. The DELU ('Δ-U') restraint is based on the fact that a covalent bond between two atoms is fairly rigid so that if the atoms move they will move in phase and therefore have the same MSDA along the bond. The SIMU ('SIMilar U') restraint is based on the assumption that displacements of atoms that are spatially close will have similar amplitudes and similar directions. This restraint is an extension of the restraints commonly used in isotropic B-factor refinement [9]. The ISOR ('ISOtropy Restraint') is mainly intended to prevent water molecules from diverging by keeping them more or less isotropic.

restraints on ADP's that are available in the program SHELXL93 [7, 8] are described in Fig. 2.

It is difficult to give a general rule as to when restrained refinement of anisotropic displacement parameters is justified by the experimental data. Caution should be exercised and it is advisable to apply the R_{free}-test [10] or the recently proposed extended Hamilton test [11] to establish the validity of ADP refinement for each case. In our experience R_{work} drops by about 5 % at any limiting resolution between 2.0 and 1.0 Å. Corresponding drops in R_{free} at 2.0, 1.5 and 1.0 Å resolution are 0.0, 2.5 and 5.0 % respectively. Hence, depending on the quality of the data, restrained ADP refinement becomes a reasonable option for data extending to a resolution somewhere between 1.5 and 1.0 Å.

B. Analysis of Anisotropic Displacement Parameters

The raw result of a refinement in the first place is nothing but a huge list of numbers (3 coordinates, 1 occupancy and 6 ADP's plus the same number of ESD's per atom). In the same way that atomic coordinates only come to life by visualizing them graphically and by calculating bond lengths and angles, ADP's also need to be translated into a more comprehensible format. A number of programs is available to represent ADP's on a computer screen (e.g. [12, 13, 14]) as well as to derive numbers that are more intuitive than the straight **U**-matrices. Although all of these programs are intended to be used for

small molecules, most of them can, after some massaging of the input files and selecting only parts of the respective structure, be used for macromolecules as well. A typical example of the graphical representation of ADP's as 'vibrational ellipsoids' is given in Fig. 3.

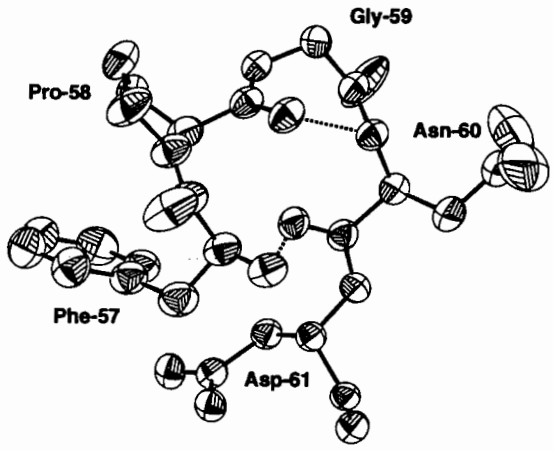


Figure 3: Representation of the ADP's of the atoms in a loop-region in a serine protease refined at atomic resolution. For each atom the U-matrix is translated into an ellipsoid, which gives an impression of the amplitude and the direction of the mean displacement. Hydrogen bonds are drawn as dashed lines. Fig. prepared with PLATON [14].

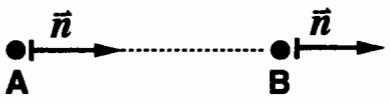
The sidechain of Asn-60 is exposed to solvent and exhibits pronounced anisotropic disorder. In contrast both carboxylate oxygens of Asp-61 are tied down by well defined intramolecular hydrogen bonds. The phenyl ring and most of the atoms in the peptide chain also show relatively small isotropic mean displacements. The non-hydrogen bonded carbonyl oxygens of Phe-57 and Gly-59 display pronounced disorder at right angles to the carbonyl planes, whereas the carbonyl oxygens of Pro-58 and Asn-60 are tied down by hydrogen bonds.

Based on the matrix of ADP's, $\mathbf{U}_A$, the mean square displacement $\langle u_A^2 \rangle_{\hat{n}}$ of an atom A in an arbitrary direction characterised by a unit vector $\hat{n}$ can be calculated via a quadratic form:

$$\left\langle u_A^2 \right\rangle_{\hat{n}} = \hat{n}^t \mathbf{U}_A \hat{n} \tag{3}$$

Based on the MSD's of two atoms along their interatomic vector, a condition for the envolvement of these two atoms in a common rigid body can be defined: if two atoms belong to a rigid body their distance will be constant and they will move in phase. This behaviour will be reflected in the ADP's by a similar displacement along the interatomic vector (Fig. 4 and Eq. 4). It must be noted, however, that eq. 4 is only a necessary and not a sufficient condition for two atoms belonging to a rigid body. Δ_{AB} may be zero for example for a planar or linear model with modest 'perpendicular' vibrations [3].

Disorder of a rigid molecule or a rigid group of atoms can be described in full generality by three matrices **T**, **L** and **S** ('TLS-model' [16, 17]). **T** and **L** are symmetric matri-



$$\begin{aligned} \Delta_{AB} &= \left\langle u_A^2 \right\rangle_{\hat{n}} - \left\langle u_B^2 \right\rangle_{\hat{n}} \\ &= \hat{n}^t \mathbf{U}_A \hat{n} - \hat{n}^t \mathbf{U}_B \hat{n} \\ &\approx 0 \end{aligned} \qquad (4)$$

Figure 4: Rigid-body criterion: if two, not necessarily covalently bonded, atoms A and B belong to a rigid body the displacements along the interatomic vector (dashed line), $\langle u_A^2 \rangle_{\hat{n}}$ and $\langle u_B^2 \rangle_{\hat{n}}$ should be the same, i.e. the difference Δ_{AB} should be zero within experimental error [15].

ces and describe translational and librational disorder, correlations between the two are represented by the non-symmetric **S** ('screw')-matrix. The elements of the **T**, **L** and **S**-matrices are adjusted by minimizing the difference between the ADP's calculated from an actual TLS-model ($\mathbf{U}_{j,kl}^{TLS}$) and the ADP's resulting from the crystallographic refinement ($\mathbf{U}_{j,kl}^{obs}$) by a least-squares procedure:

$$\sum_j \sum_{kl} \left(\mathbf{U}_{j,kl}^{obs} - \mathbf{U}_{j,kl}^{TLS} \right)^2 \longrightarrow \text{Min}, \qquad (5)$$

with index j running over all atoms and indices k, l running over all matrix elements.

A classical example of a thorough analysis of ADP's is the study on Triphenylphosphine oxide by Brock et al. [18]. TPP is a small molecule consisting of three phenyl rings and an oxygen bound to a central phoshor in an approximately tetrahedral conformation (Fig. 5). The first step of the analysis was the calculation of Δ_{AB}-values for all possible

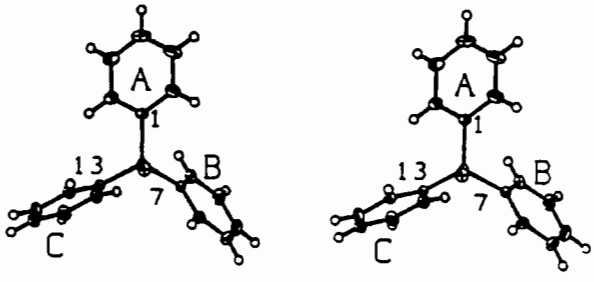


Figure 5: Stereo drawing of TPP. Fig. reproduced from [18] with permission.

combinations of atoms A and B. The resulting values were arranged in a so-called Δ-matrix (Fig. 6). The triangles labeled A,B and C contain Δ-values for the three phenyl rings ('intra-ring'). The three rectangular blocks AB, AC and BC contain Δ-values for pairs of atoms belonging to different rings ('inter-ring'). The intraring Δ values are close to zero and significantly smaller compared to the interring Δ-values. In other words: the rings themselves fullfill the rigid-body criterion but they move relative to one another. The rigid-body displacements of the individual phenyl-rings were then analysed by fitting TLS-models to the ADP's of the respective atoms giving the result that for all three rings the dominant contribution to the displacements is a libration around the bond connecting the ring and the central phosphorus atom.

atom	C18	C17	C16	C15	C14	C13	C12	C11	C10	C9	C8	C7	C6	C5	C4	C3	C2	C1	P1
O1	0	-8	-15	-6	-6	-19	18	8	10	35	32	9	2	-15	-22	9	48	-10	-20
P1	33	31	23	25	31	10	38	32	22	31	35	22	29	25	16	24	41	13	
C1	48	28	-2	-11	-5	-9	40	44	11	-5	-3	0	11	5	5	10	21		
C2	15	5	-23	-44	-46	-34	-73	-89	-100	-77	-57	-61	-2	-7	5	-6			
C3	49	26	-22	-45	-38	-20	-21	-31	-74	-78	-57	-42	-11	-22	2				
C4	56	12	-53	-59	-34	-25	36	67	12	-31	-29	-15	-10	-16		A			
C5	14	-32	-78	-68	-41	-47	-25	2	-25	-40	-29	-33	-11						
C6	17	-10	-34	-32	-18	-30	-50	-34	-51	-49	-32	-44							
C7	35	42	23	20	19	3	2	-1	0	1	-6								
C8	19	13	-19	-23	-23	-36	8	-9	-9	1									
C9	19	12	-27	-32	-30	-36	5	-5	-3										
C10	27	44	13	4	-1	-12	-5	-8		B									
C11	-2	21	1	-3	-5	-17	-7												
C12	6	24	4	-1	-1	-11													
C13	7	16	13	7	3														
C14	0	5	7	2															
C15	-8	-13	-11																
C16	2	4		C															
C17	-5																		

Figure 6: Δ-matrix for TPP as given in [3]. For atom numbering see Fig. 5. Positive Δ's mean that the MSDA along the interatomic vector between A_1 and A_2 is larger for A_2 than for A_1, negative Δ's mean the reverse. The estimated standard deviation of the Δ_{AB}-values is about 7 pm^2. Fig. reproduced from [18] with permission.

C. Analysis of ADP's in a protein

In cases where data of very high quality are available the above concepts can be applied to protein structures. One example is the crystal structure of SP445 (Fig. 7), a serine protease from *Nocardiopsis*, for which data to 0.97 Å resolution were collected at 120 K on beamline X11 at EMBL Hamburg. The overall R_{merge} for the 97 % complete data set is 3.3 %. The structure was refined using SHELXL93 employing restrained anisotropic displacement parameters for all non-hydrogen atoms. The current model has an R_{work} of 8.0 and an R_{free} of 10.3 %. It is hopeless to calculate and analyse the Δ-matrix for the more than 1400 non-hydrogen atoms of this molecule, but subsets can be selected for analysis. For example atoms of tyrosine rings are in some sense in a situation similar to the atoms forming the phenyl rings in TPP, the difference being that the phosphorus atom is replaced by the polypeptide backbone. Including hydroxyl oxygens the five ordered tyrosine rings in SP445 result in a 35×35 Δ-matrix (Fig. 9). Inspection of this Δ-matrix reveals that the intraring Δ-values marked by the grey boxes along the diagonal are relatively small: the rings themselves fulfil the rigid-body criterion. Apart from the Δ-values between Tyr-9 and Tyr-80, which are spatially close and both hydrogen bonded to the backbone via their hydroxyl oxygens, the inter-residue Δ-values are generally much larger indicating that different rings move independently. It should be kept in mind that the rigidity of the individual rings might be imposed by the restraints used in the refinement. Refinements using weak or no restraints on APD's in stable regions of the protein are under way.

Another interesting object to study is the polypeptide-backbone. The systematic increase of B-factors from the interior towards the surface of SP445 (Fig. 8) has been observed for a number of macromolecular structures and a librational rigid-body displacement of entire

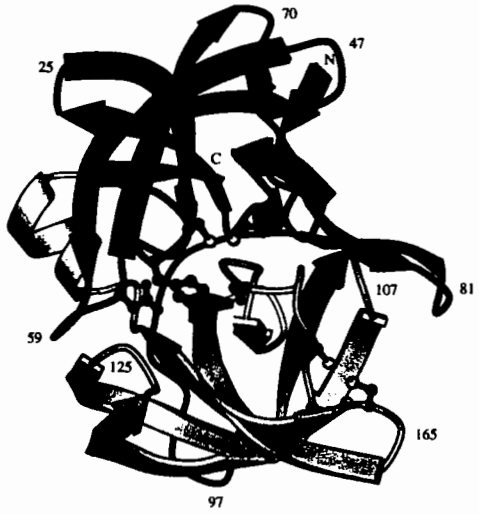


Figure 7: Schematic view of SP445. The molecule consists of 188 amino acids and exhibits a trypsin-like fold. The catalytic triade is located at the interface between the two β-barrel domains (shown in dark and light grey). The Figure was prepared using MOLSCRIPT [19].

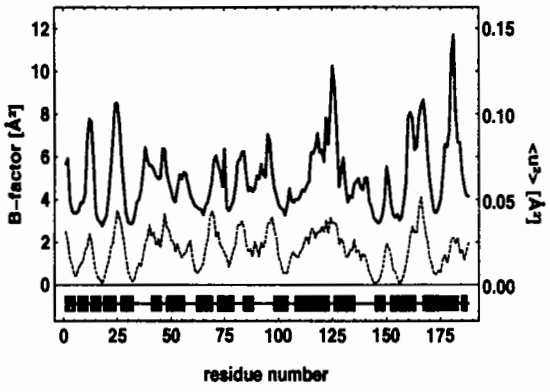


Figure 8: Isotropic $B/\langle u^2 \rangle$-values averaged for N-C_α-C atoms versus residue number (full line). The squared distance of the respective C_α atom from the centre of mass of the molecule is represented by the dashed line. Secondary structure elements are indicated by black (β-sheet) and grey (α-helix) rectangles.

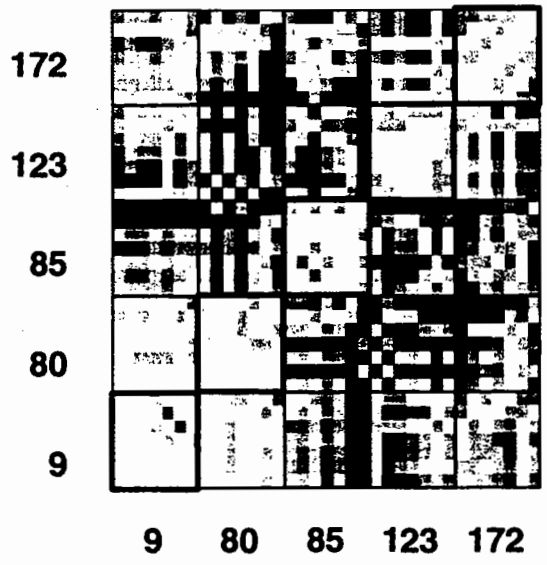


Figure 9: Δ-matrix for tyrosine ring atoms in SP445. Instead of numbers colours are used to represent the Δ-values : white corresponds to $\Delta \leq 1\sigma_\Delta$ and black to $\Delta > 3\sigma_\Delta$, values with $1\sigma_\Delta \leq \Delta \leq 3\sigma_\Delta$ are shown in different shades of grey. ESD's for the Δ-values were derived from the ESD's of the ADP's obtained from block-matrix refinement by an approach suggested by Irmer [20].

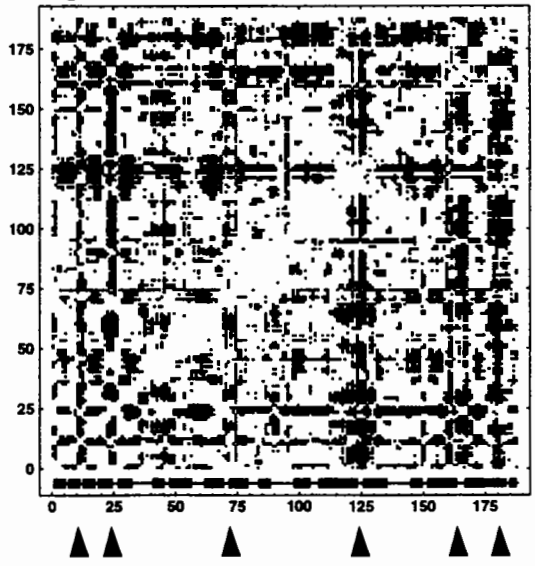


Figure 10: Reduced Δ-matrix for backbone atoms in SP445. The 564×564 matrix corresponding to all 564 N-C_α-C atoms has been reduced by averaging over the 9 interactions for each residue pair resulting in a 188×188 matrix. Average Δ-values smaller than 3σ are shown in white, greater than 3σ are shown in black.

molecules has been suggested as a possible source (e.g. [21, 22, 23]). The information about directions contained in the ADP's can be used to test this hypothesis.

First the rigid-body test can be applied to distinguish between potentially rigid and flexible parts of the protein. The corresponding Δ-matrix is shown in Fig. 10. The black streaks correspond to non-rigid parts of the polypeptide backbone, namely surface loops and the C-terminal α-helix.

If librational disorder of entire molecules is the origin of the observed distance dependence of the equivalent isotropic B-values, the radial and tangential displacements of atoms should exhibit qualitatively different distance dependencies: radial displacements should stay constant and tangential displacements should increase from the centre towards the surface of the molecule (see Fig. 11 for illustration). These distance dependencies can be regarded as a necessary condition for a librational disorder of entire molecules in the crystal. After excluding all backbone atoms that failed in the rigid-body test radial and tangential displacements relative to the centre of mass of the protein were calculated for the remaining 402 N-C_α-C atoms. The results are shown in Fig. 12: radial displacements are constant and tangential dispacements show a parabolic behaviour with increasing distance from the centre of mass. The offsets for tangential and radial displacements (0.027 and 0.031 Å^2) are due to overall translational disorder. The mean libration angle

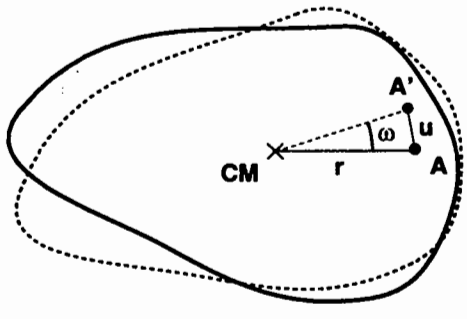


Figure 11: Effect of an overall librational disorder on the mean square displacements of an atom A in different directions. If a molecule is librating around its centre of mass with an average libration angle ω, the mean square displacement of an atom A in the radial direction vanishes for all distances r from the centre of mass. The tangential displacement can be calculated as $u = \omega r$ giving rise to a quadratic distance dependence of the mean square displacement.

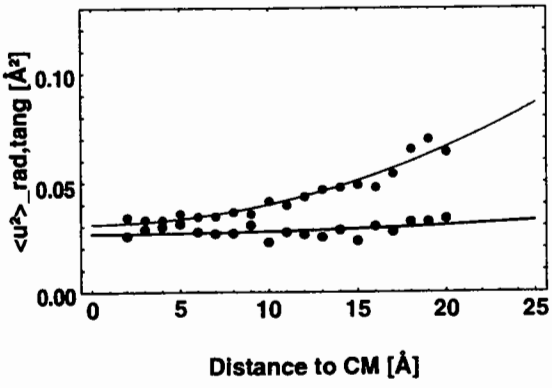


Figure 12: Radial (black) and tangential (grey) displacements relative to the centre of mass of the protein. Displacements were calculated for all 402 N-C_α-C atoms fulfilling the rigid-body criterion. To suppress contaminations by contributions due to effects other than libration only the minimum displacements in 1 Å shells around the centre of mass were plotted against the distance from the centre of mass of the molecule. Lines correspond to fits of a function $a+br^2$ to the radial (black) and tangential (grey) displacements ($f_{rad}(r) = 0.026605 + 0.00001\ r^2$, $f_{tan}(r) = 0.031051 + 0.000088\ r^2$).

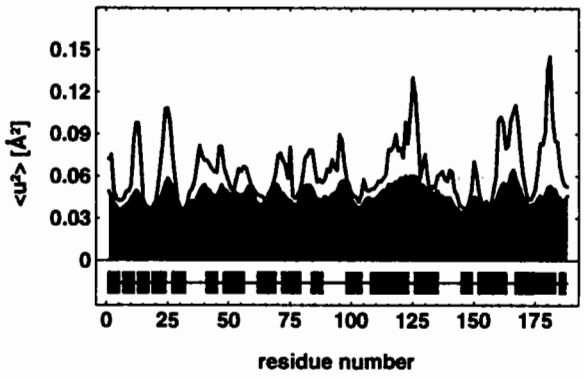


Figure 13: Mean square displacements $\langle u^2 \rangle$ averaged for N-C_α-Catoms of each residue (white curve) and mean square displacements derived from the TLS model described in the text. Secondary structure elements are indicated as in Fig. 8.

corresponding to the function fitted to the tangential displacements is 0.54°.

To fully characterise a rigid-body disorder of the molecule as a whole, a TLS-model was fitted to the refined ADP's of the 402 previously selected atoms. Such a simple least-squares fit of a TLS-model will always overestimate the rigid-body contributions. This problem can be partly alleviated by scaling the initial **T** and **L** such that no TLS derived displacement is larger than the respective observed displacement (otherwise negative displacements would be obtained after correcting the observed displacements for rigid-body disorder). However even after this correction the derived rigid-body displacements should merely be considered as an upper limit for this contribution, since the TLS model will to some extent 'mop up' contributions that are due to the internal normal modes which are not present in this simple model [24]. After performing the above mentioned scaling of the initially obtained **T** and **L** tensors the TLS-model gives a mean translational displacement of 0.034 Å^2 and an average libration angle of 0.49°, values which are in good agreement with the results derived from the analysis of radial and tangential displacements. In Fig. 13 the deconvolution of equivalent isotropic displacements into external and internal contributions based on the TLS model determined above is shown. The curve corresponding to the TLS-model has a high correlation (correlation coefficient 0.79) with the curve for observed displacements and takes up a fairly large part of the latter. Again it must be emphasized that the TLS-contribution only represents an upper limit for external effects. Correction of the atomic displacement for these consequently only leads only to lower limits for the internal contributions.

Conclusions and Perspectives

Provided atomic resolution data are available, restrained anisotropic displacement parameters can be refined for protein molecules. Visual inspection and numerical analysis of ADP's can lead to new ideas about the different contributions to the overall mean displacements of atoms in a crystal structure. In particular rigid-body disorder of entire molecules

or parts of molecules can be detected and characterised by adjusting TLS-models to best fit the observed ADP's. The interpretation of results requires great caution, since the TLS approach in its simple implementation overestimates external contributions: only upper limits for external and consequently only lower limits for internal disorder can be derived. In addition it must be kept in mind that based on Bragg reflections it is not posible to derive rigorous conclusions about crystal or molecular vibrations [24]. In principle the information on such correlated displacement is available in the thermal diffuse scattering signal ('TDS') but currently difficult to access [25, 26]. Nevertheless dynamic and static effects can be distinguished by multiple temperature experiments and models obtained on small proteins refined at atomic resolution might help in defining models of disorder for larger proteins where atomic resolution data are generally not accessible.

Acknowledgements

I am grateful to NOVO Nordisk for providing samples of SP445 and I thank both my PhD. supervisors Prof. Keith Wilson and Prof. Fritz Parak for continuing support and many discussions. This work was supported by an EMBL Predoctoral Fellowship and by a grant from the European Union (BI02-CT92-0524).

References

[1] G.M. Sheldrick. Phase Annealing in SHELX-90: direct methods for larger structures. *Acta Cryst.*, A46:467–473, 1990.

[2] Z. Dauter, V.S. Lamzin, and K.S. Wilson. Proteins at atomic resolution. *Curr.Op.Struct.Biol.*, 5:784–790, 1995.

[3] J.D. Dunitz, V. Schomaker, and K.N. Trueblood. Interpretation of Atomic Displacement Parameters from Diffraction Studies of Crystals. *J.Phys.Chem.*, 92:856–867, 1988.

[4] J.D. Dunitz, E.F. Maverick, and K.N. Trueblood. Atomic Motions in Molecular Crystals from Diffraction Measurements. *Angew.Chem.Int.Ed.Engl.*, 27:880–895, 1988.

[5] K.N. Trueblood. Diffraction studies of molecular motions in crystals. In A. Domenicano and I. Hargittai, editors, *Accurate molecular structures: Their determination and importance*, pages 199–219. Oxford University Press, Oxford, 1992.

[6] B.T.M. Willis and A.W. Pryor. *Thermal Vibrations in Crystallography.* Cambridge University Press, London, 1975.

[7] G.M. Sheldrick. *SHELXL-93.* University of Göttingen, Germany, 1993.

[8] G.M. Sheldrick and T.R. Schneider. SHELXL:High Resolution Refinement. In Carter C. and Sweet B., editor, *Methods in Enzymology*, volume in press, pages 000–000. Academic Press Inc., Orlando, Florida, 1996.

[9] J.H. Konnert and W.A. Hendrickson. A Restrained-Parameter Thermal Factor Refinement Procedure. *Acta.Cryst.*, A36:344–350, 1980.

[10] A.T. Brünger. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355:472–475, 1992.

[11] A. Bacchi, V. Lamzin, and K.S. Wilson. Self Validation: An extended Hamilton-Test. *CCP4 Study Weekend*, this volume:000–000, 1996.

[12] C.K. Johnson. ORTEP. Report ORNL-3794. Technical report, Oak Ridge National Laboratory, Oak Ridge, Tennesee, USA., 1965.

[13] W. Hummel, J. Hauser, and H.B. Bürgi. Peanut: Computer graphics program to represent atomic displacement parameters. *J.Mol.Graphics*, 8:214–220, 1990.

[14] A.L. Spek. *PLATON-92.* University of Utrecht, The Netherlands, 1992.

[15] R.E. Rosenfield, K.N. Trueblood, and J.D. Dunitz. A test for rigid-body vibrations, based on a generalization of Hirshfeld's 'rigid-bond' postulate. *Acta Cryst.*, A34:828–829, 1978.

[16] D.W.J. Cruickshank. The Analysis of the Anisotropic Thermal Motion of Molecules in Crystals. *Acta Cryst.*, 9:754–756, 1956.

[17] V. Schomaker and K. Trueblood. On the Rigid-Body Motion of Molecules in Crystals. *Acta Cryst. B*, 24:63–76, 1968.

[18] C.P. Brock, W.B. Schweizer, and J.D. Dunitz. Internal Molecular Motion of Triphenylphosphine Oxide: Analysis of Atomic Displacement Parameters for Orthorombic and Monoclinic Crystal Modifications at 100 and 150 K. *J.Am.Chem.Soc.*, 107:6964–6970, 1985.

[19] P. Kraulis. Molscript: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.*, 24:946–950, 1991.

[20] E. Irmer. *Restraints in der Kristallstrukturverfeinerung und systematische Überprüfung von Datenbankstrukturen auf höhere Symmetrie.* PhD thesis, Universität Göttingen, 1990.

[21] P.J. Artymiuk, C.C.F. Blake, D.E.P. Grace, S.J. Oatley, D.C. Phillips, and M.J.E. Sternberg. Crystallographic studies of the dynamic properties of lysozyme. *Nature*, 280:563–568, 1979.

[22] H. Hartmann, F. Parak, W. Steigemann, G.A. Petsko, D. Ringe-Ponzi, and H. Frauenfelder. Conformational substates in a protein: structure and dynamics of metmyoglobin at 80 k. *Proc.Natl.Acad.Sci. USA*, 79:4967–4971, 1982.

[23] M. Eisenstein, H. Hope, T.E. Haran, F. Frolow, Z. Shakked, and D. Rabinovich. Low-Temperature Study of the A-DNA Fragment d(GGGCGCCC). *Acta Cryst.*, B44:625–628, 1988.

[24] A. Kidera and N. Go. Normal Mode Refinement - Crystallographic Refinement of Protein Dynamic Structure. 1. Theory and Test by Simulated Diffraction Data. *J.Mol.Biol.*, 225:457–475, 1992.

[25] J.P. Benoit and J. Doucet. Diffuse scattering in protein crystallography. *Quart.Rev.Biophys.*, 28:131–169, 1995.

[26] T. Thüne and J. Badger. Thermal diffuse X-ray scattering and its contribution to understanding protein dynamics. *Progr.Biophys.Mol.Biol.*, in press, 1995.

SHARPENED MAPS FOR MORE EFFECTIVE REFINEMENT IN PROTEIN CRYSTALLOGRAPHY

Susanna Butterworth [i,ii], Victor S. Lamzin [i] & Keith S. Wilson [i,iii]

[i] European Molecular Biology Laboratory (EMBL)
c/o DESY, Notkestraße 85, 22603 Hamburg, Germany
[ii] University of Durham, South Road, Durham, DH1 3LR, U.K.
[iii] University of York, Heslington, York, YO1 5DD, U.K.

Abstract

Refinement of the model structure obtained from an X-ray diffraction experiment improves its fit to the data in reciprocal space and to the electron density in real space. Real space modifications can be carried out manually, with the aid of graphical simulation, or automatically, using computer programs. The quality of the electron density maps used is crucial for the success of refinement.

Due to the falloff of scattering intensity with increasing sinθ, high resolution data are generally weak. Structure factors can be normalised to remove their resolution dependence, so the high resolution terms make a more significant contribution to maps, giving sharper atomic peaks. Since the high resolution intensities are weak, they have large associated errors. Their upweighting leads to a concurrent magnification of the associated errors and the appearance of spurious peaks in the density map.

An optimal degree of sharpening of the data leads to maps in which atomic peaks are sharp and well defined, while the noise contribution is minimal. The desirable degree of sharpening varies with the characteristics of the structure, the resolution of the data and overall B factors. Various methods can be employed in the determination of the most informative level of sharpening for maps, including inspection of electron density distributions and trial-and-error refinement runs.

E's, F's and electron density maps

A structure factor is the result of the summation of the scattering of all the atoms in the unit cell in a given direction. It possesses phase and amplitude. The amplitude is proportional to the atomic scattering factor. The phase is determined by the position of the atom in the unit cell with respect to the origin. As only amplitudes are available by experiment, estimated phases must be obtained through the application of one or more structure solution methods. From the combination of the model phases and observed structure factors electron density maps are computed.

An atomic scattering factor is composed of the atomic form factor for a spherical atom, which decreases with increasing sinθ due to interference between waves scattered from different parts of the electron density within the atom and, in addition, a term accounting for the static and dynamic disorder of the atom, also resolution dependent. If the atom is considered to be a point scatterer, the expression for the scattering factor can be divided by all the resolution dependent terms, producing a resolution independent normalised structure factor, E-value (1).

$$|E_{hkl}|^2 = |F_{hkl}|^2 / \Sigma f_j^2 \quad (1)$$

f = atomic form factor, j = 1 to N, N = atoms per unit cell

A normalisation factor may be determined, assuming randomly distributed atoms, from the gradient of the Wilson plot (Wilson,1942): ln ($\Sigma f_j^2(s) / |F_{hkl}|^2$) against s^2 ($s = \sin\theta/\lambda$). Normalised structure factors can also be calculated by the K curve method (Karle & Hauptman, 1953). The data are divided into resolution bins, then $E^2 = F^2 / \langle F^2 \rangle$. The CCP4 (1994) program ECALC utilises the K curve approach. The data are sorted into overlapping resolution bins and smoothing is applied to the intensities, before calculation of the average values.

Models

This work utilised 4 structures, Table 1. The data were collected using synchrotron radiation at the EMBL outstation at DESY, Hamburg. The models were refined using similar protocols, details of which are described elsewhere. The resolution of the data covers the range in which the sharpening of maps would be expected to be beneficial, from atomic to 2 Å.

Table 1. Crystal structures

Structure	Space group	Resolution (Å)	Wilson Plot B factor (Å^2)	Reference
Rubredoxin	$P2_1$	20.0 - 0.92	15	Dauter *et al.*, 1992
Protein G	$P2_12_12_1$	10.0 - 1.1	20	Derrick & Wigley, 1994
Eglin	$P4_3$	10.0 - 2.0	37	Betzel *et al.*, 1993
Transthyretin	$P2_12_12$	10.0 - 1.9	44	Damas *et al.*, 1996

Maps

(F_o-F_c,α_c) and ($3F_o$-$2F_c$,α_c) maps with varying sharpness and resolution limits were computed using FFT (Ten Eyck, 1973), ECALC and other programs from the CCP4 suite. Following the application of artificial resolution cuts to data, 20 cycles of restrained least-squares refinement were run with the CCP4 version of PROLSQ (Konnert & Hendrickson, 1980), to minimise the memory of the high resolution data in the model. A scale factor of $(\Sigma F^2 / \Sigma E^2)^{1/2}$ was applied to the E-values, so that maps calculated with different degrees of sharpness would be comparable.

The shape of electron density

Effective placing of atoms and improvement of their position in the electron density, particularly by automatic means, is possible if atoms are resolved in the density map. The visibility of atoms in the density for maps at different resolutions and with different degrees of sharpening was investigated by examining the shape of the density between neighbouring atomic centres. The maps analysed were of the form ($F_o^x E_o^{1-x}$,α_c) with $0 \le x \le 1$. All pairs of atoms separated by distances of 1.9 Å or less were selected. The electron density at the two atomic centres and at nine equally spaced points along the line connecting them was calculated. Pairs of points equidistant from the midpoint were then averaged. The resulting values for the density between each pair of atoms were normalised to give a value of unity at the atomic centre. These values give a representation of the average shape of the density between neighbouring atoms, Figure 1.

If r_{atom} is the averaged, normalised density at the atom centre, $r_{midpoint}$ is the density at the midpoint between two atoms and $\rho_{difference}$ is the difference between ρ_{atom} and $\rho_{midpoint}$, then the atoms can be said to be 100 % resolved if:

$$\rho_{atom} = 1 \text{ and } \rho_{midpoint} = 0 \quad \text{therefore} \quad \rho_{difference} = 1$$

The atoms are 40 % better resolved on the E-map than on the F-map if:

$$\rho_{difference}\ (\text{E-map}) - \rho_{difference}(\text{F-map}) = 0.4$$

The plot of $\rho_{difference}$ against resolution cutoff, Figure 2, shows that the atoms in the E-map are 40 % better defined at atomic resolution. The resolvability of atoms in maps of all degrees of sharpness declines as the high angle data are cut. The extra interpretability of E-maps over F-maps is also reduced. At 1.5 Å the difference in resolvability of atoms is 8 % and at 2.0 Å, 4 %, but E-maps retain an advantage until a threshold resolution around 2.25 Å to 2.5 Å.

The shape of the density between atoms, Figure 1, is principally determined by the resolution of the data. At atomic resolution there is a pronounced minimum at the midpoint. This minimum is lower in the E-map. The depth of the minimum decreases for all maps as the resolution is cut. At around 2.5 Å the shape of the density undergoes qualitative changes and there is a maximum at the midpoint, Figure 3. Between 2 Å and 2.5 Å it may be possible to resolve atoms in the E-map as there is still minimum although there is not in the F-map.

Figure 1. Averaged shape of Interatomic Density: Rubredoxin

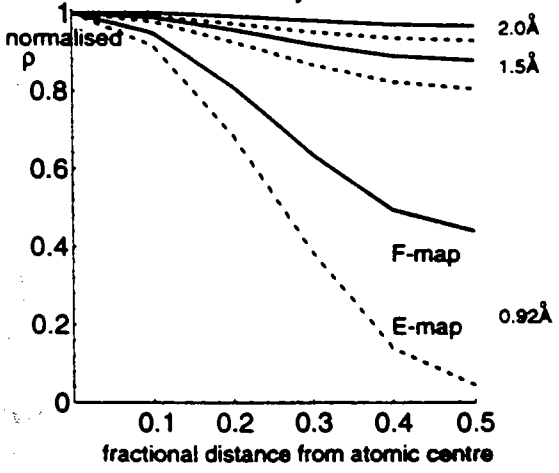


Figure 2. Resolution dependence of $\rho_{difference}$: Rubredoxin

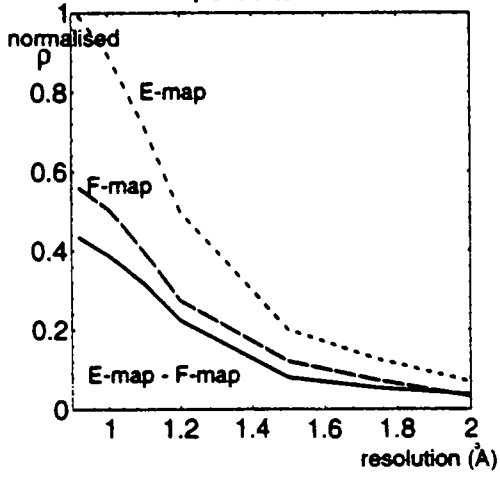
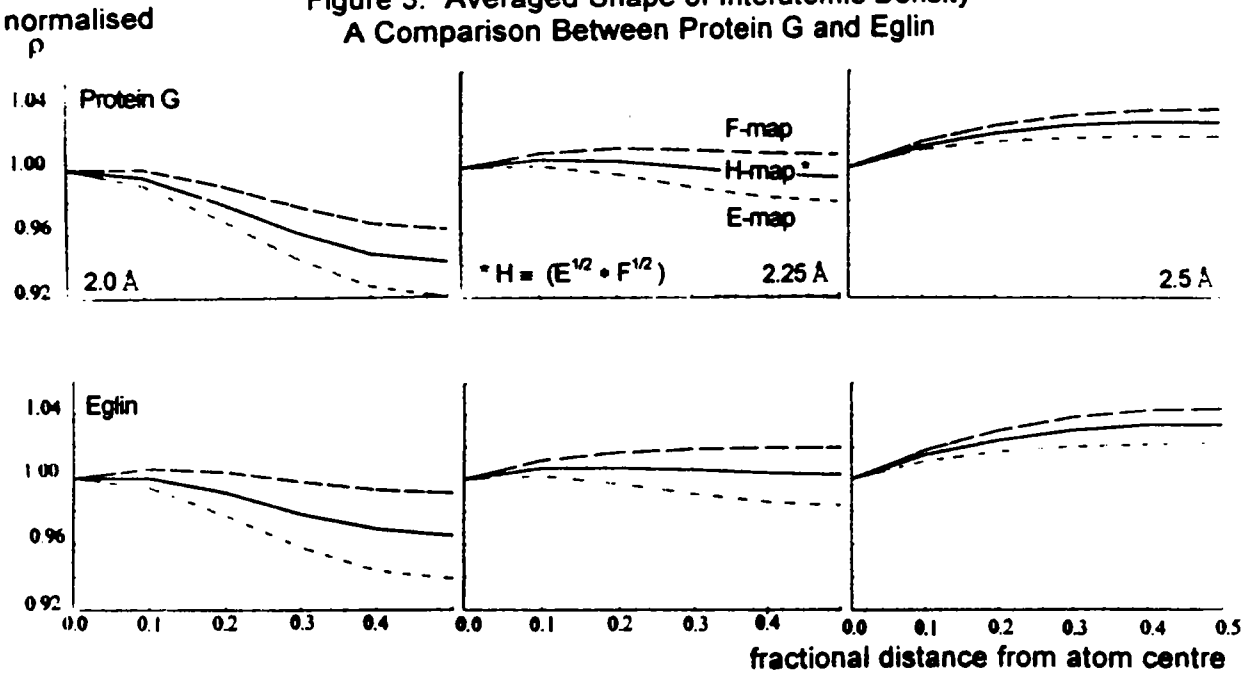


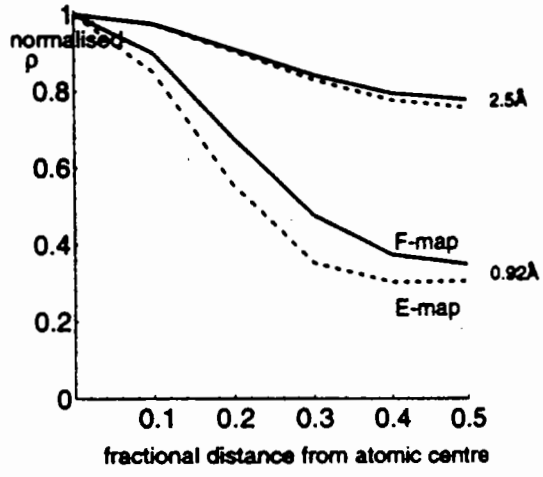
Figure 3. Averaged Shape of Interatomic Density
A Comparison Between Protein G and Eglin



The shape of the density in the F-map is influenced by the thermal parameters of the structure. For a structure with high B factors the atoms are less resolved. A map in which the resolution had been artificially cut can be distinguished from one for which data are present up to the diffraction limit, thus there is a substantial difference between the shape of the density in the F-map for protein G at 2.0 Å, and that for eglin at 2.0 Å, Figure 3. Since thermal effects should be removed during the calculation of E-values, there is a greater similarity between the shape of the density in the E-maps.

The density around solvent atoms was analysed, as described above for the protein atoms, by selecting each solvent atom and its nearest neighbour and calculating the density along the line joining the atomic centres, Figure 4. The change in shape of the density with resolution cutoff was much more gradual in this case. This demonstrates that the high resolution data contain little information about the solvent. The separation between solvent sites is greater than that between adjacent atoms in the protein, so solvent molecules are still resolved at a lower resolution.

Figure 4. Averaged density between solvent sites: Rubredoxin



Electron Density Histograms

The electron density in the unit cell may be assumed to be composed of Gaussian atoms on a fairly flat background (Main, 1990). Thus, the histogram of electron density is comprised of two components, a low density background, with an approximately random distribution, which can be described by a Gaussian function and a high density contribution from the atomic peaks. The convolution of these two distributions is a function with a characteristic shape. At atomic resolution, the distribution is highly skewed with a sharp maximum lying close to $\rho = 0$, a long tail in the $\rho > 0$ region and a short tail in the $\rho < 0$ region. As the resolution is cut, the peak becomes lower and broader and the distribution less skewed.

Examination of the electron density histogram allows assessment of the characteristics of the entire map. The density histogram of a map, generated using correct phases, will be highly skewed. Increasing the phase error results in a more Gaussian distribution. Thus, skewness of the distribution and phase quality are related. It has been proposed that for a given structure the distribution generated from the best set of phases will possess maximum skewness (Cochran, 1952; Podjarny & Yonath, 1977; Lunin, 1993).

Maps of the form $(F_o^x E_o^{1-x}, \alpha_c)$, $0 \le x \le 1$, were calculated, with the application of varying resolution cutoffs to the four datasets. It was suggested that the most informative map should be that for which skewness is maximum. The skewness of the density distribution was calculated for each map. For each set of maps, for a single dataset at a specific resolution, skewness was plotted againts x, where x is the power of F_o, Figure 5. The value of x at which skewness is maximum is defined as xmax , so $F_o^{xmax} E_o^{1-xmax}$ is the map possessing the degree of sharpness which yields the maximum skewness distribution. xmax against resolution cutoff is shown in Figure 6 for each dataset.

Figure 5. Variation of Skewness with X $F^x E^{(1-x)}$ maps: Rubredoxin

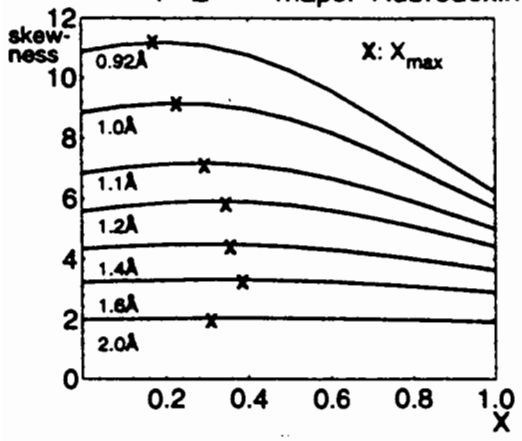
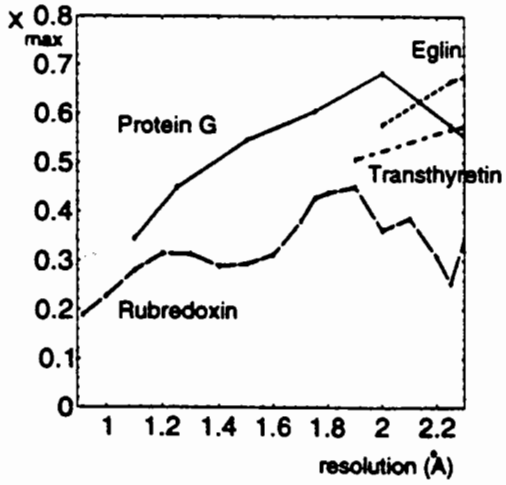


Figure 6. Resolution dependence of Xmax



At atomic resolution the highly sharpened map, ($F^{0.2}$ $E^{0.8}$), is found to contain the most interpretable features. As the resolution is decreased, the optimal degree of sharpening is reduced. This can be explained by the fact that as more of the high resolution data are removed, termination errors increase the noise level. Thus, a greater 'F' contribution to the structure factors is necessary to dampen the noise and the F/E balance swings towards F.

Refinement using Sharpened Maps

Models derived from an atomic resolution structure of Rubredoxin from *Desulfovibrio vulgaris* were subject to a refinement procedure, involving the use of maps of varying degrees of sharpness. The improvement brought about by refinement was assessed.

Models for Refinement

I: An almost fully refined model, with 82 water molecules and R and Rfree values of 8.3 % and 11.2 %. The SHELXL-93 diffuse solvent correction, which is based on Babinet's principle (Langridge *et al.*, 1960) but differs from that described by Tronrud (1996), had been applied during the previous cycles of refinement. The resolution range of the data was 20 to 0.92 Å. 5 % of the reflections had been isolated from the working dataset for calculation of Rfree.

II: Solvent atoms with B factor > 30 $Å^2$ were removed from ***I***, leaving 32 waters. R and Rfree values were 10.4 % and 12.2 % respectively.

III: A random deviation with rms 0.3 Å was introduced into model ***I*** and, in addition, the co-ordinates were shifted by 0.5 Å along the *a* axis. This mimics inaccuracies which could be present in a model obtained by molecular replacement. R and Rfree were 44.8 % and 45.9 %.

IV: Molecular replacement was carried out using AMORE (Navaza, 1994). The search model was the 1.4 Å structure of Rubredoxin from *Desulfovibrio gigas* (Frey *et al.*, 1988), which has an rms displacement of 0.65 Å from ***I*** for CA atoms. There are 14 differences between the two sequences. Side chain shortening mutations were carried out where this was appropriate. 7 residues in the molecular replacement model were mutated to Ala and 2 to Ser, leaving 9 sequence differences between the starting model and ***I***. R and Rfree were 39.2 % and 41.0 %.

V: A loop region in ***I*** comprising residues Pro 20 to Val 24 was removed. R and Rfree were 20.9 % and 24.3 %.

Refinement

Restrained least-squares refinement of atomic positions and thermal parameters was performed, using PROLSQ and also SHELXL-93 (Sheldrick,1993). Standard geometric restraints were applied to bonding distances and thermal parameters during both types of refinement. Hydrogen atom positions were not refined but calculated using a riding model.

The Automated Refinement Procedure (Lamzin & Wilson, 1993) was employed for modification of the structures in real space. ARP identifies atoms for removal by inspection of the ($3F_o$-$2F_c$,α_c) map, together with the application of distance constraints. Atoms which approach each other within a given distance are merged, leaving a new atom at the midpoint. Sites for addition of new atoms are found from the difference map, (F_o-F_c,α_c), with distance constraints applied. Real space refinement matches the expected and actual shape of density around an atom in the ($3F_o$-$2F_c$,α_c) map and moves the atom to improve its sphericity. The refinements were repeated with variation of the sharpness of the ($3F_o$-$2F_c$,α_c) map input to ARP; unsharpened (F), fully sharpened (E) and half-sharpened ($F^{1/2}E^{1/2} \equiv H$) structure factors were used in turn.

For ***I*** & ***II***, five cycles of SHELXL-93 anisotropic refinement were run, followed by a cycle of ARP. This was iterated ten times. ARP was used for modification of solvent only. The distance range for addition of new atoms was set to 2.2-3.3 Å and the merging distance to 0.6 Å. Refinement was run with and without the application of the SHELXL-93 diffuse solvent correction, and with and without real space refinement. The number of atoms to be added and removed in each cycle was set to 0 or 5 for ***I*** and 10 or 15 for ***II***.

For ***III***, ***IV*** & ***V***, PROLSQ refinement was performed until convergence. Following each cycle, ARP was run. Real space refinement was carried out on all atoms, but only those designated as solvent were cut and added. Atoms were added at distances of 1.0-3.3 Å from existing ones, and merged if they lay within 0.6 Å. These limits were set to allow for the fact that some of the 'solvent' may represent protein atom sites. Observation of the change in R factors was useful in determining the convergence point of a refinement, while the refinement parameters were being tuned. However, a comparison of R factors does not give a good assessment of how well the refinement process corrects the deliberate mistake introduced since R factors refer to the whole model.

Results of Refinement

In the final stages of refinement the well defined part of the model remains virtually unchanged, while improvements are made in the fitting in the disordered regions and solvent. Such was the case for the refinements of ***I*** and ***II***. The effect of sharpening is to upweight the high resolution terms. Since scattering from the regions which were modified by these refinements does not contribute greatly to the high resolution data, the effect of sharpening was not dramatic. The use of fully sharpened maps was ineffective, causing the maximum number of atoms to be removed and added on each cycle, arguing that the noise level in these maps was too high. The semi-sharpened maps were more useful than plain F-maps. Real space refinement assisted in the equilibration of the solvent building during the reconstruction of the solvent network of model ***II***. This can be explained by the observation that diffuse atoms tend to drift towards the edges of the density, a problem corrected for by real space fitting.

The effect of the application of the SHELXL-93 diffuse solvent correction was much more noticeable, since this correction is specific to the low resolution data. When the solvent correction was implemented during the building of a very incomplete solvent network, the addition of solvent was slowed down. When the virtually complete model ***I*** was refined, with the solvent correction turned on, the resulting model had fewer solvent molecules and a lower

value of R_{free}. This can be ascribed to the removal of peaks which were present due to incorrect scaling of low resolution terms.

Models ***III***, ***IV*** and ***V*** roughly approximate to structures at earlier stages in refinement with significant errors in the well defined part of the density. The degree to which refinement has corrected the inaccuracies which were introduced can be assessed by observing change in the rms displacement of the main chain atoms from those of model ***I***, Table 2. ***IV***, the molecular replacement model, possesses 3 regions in which the position of the chain is seriously in error and requires interactive graphical rebuilding. When these regions are not used in the calculation of rms displacement of CA atoms from those in ***I***, the values obtained closely mirror those found for model ***III***. R_{free} values reflect the success in correcting the mistake in the main chain, while R values are insensitive. In all three cases, refinement using E-maps produced the best final model. The results from using H-maps were similar to those obtained with E-maps, while the F-map based models were considerably worse.

Table 2. The refinement of models ***III***, ***IV*** & ***V***

III	R (%)	Rfree (%)	rmsd (Å)
starting model	44.8	45.9	0.269 *i*
using E	15.6	18.0	0.041 *i*
using H	15.5	18.5	0.044 *i*
using F	15.5	19.0	0.057 *i*
IV			
starting model	39.2	41.0	0.398 *ii*
using E	17.3	19.6	0.043 *ii*
using H	16.9	19.8	0.046 *ii*
using F	16.9	20.9	0.058 *ii*
V			
starting model	20.9	24.3	-
using E	16.1	18.5	0.054 *iii*
using H	15.8	18.8	0.054 *iii*
using F	15.6	19.4	0.081 *iii*

i	rms deviation of CA atoms from those in ***I***
ii	rms deviation of CA atoms from those in ***I*** calculated for residues excluding; Met(1), Pro(34), Ala(35), Lys(46) & Ser(47)
iii	rms deviation of closest peaks in the model obtained from ***V*** from main chain atoms in the loop region of ***I***

Conclusions

The first part of this investigation examined density between adjacent atoms; the variation of peak shape with sharpness and resolution. Fully sharpened maps possess the best resolved peaks. From the analysis of density histograms, which reflect the character of the whole map, it was concluded that the optimal level of sharpness increases with resolution, and is around $E^{0.8}\ F^{0.2}$ at atomic resolution. The test refinements demonstrate the effectiveness of sharpening during the early stages, while showing that this approach confers a lower advantage towards the end of refinement. For an atomic resolution structure, the optimal degree of sharpening appears to lie between fully (E) and half-sharpened ($E^{1/2}\ F^{1/2}$) maps, in agreement with the value obtained from the density histogram study.

Use of sharpening during automated improvement of the model in real space enhances the accuracy with which atoms can be placed within the density. Where significant errors are there in well-defined regions, improvement in the model is accelerated and enhanced. In the closing

stages of refinement, sharpening becomes less important, as most of the atoms are already well positioned and modifications are made to peripheral, weakly scattering parts. The level of sharpness which produces the most informative map is strongly resolution dependent. Sharpening should be advantageous for any refinement at a resolution higher than 2.5 Å, a resolution range which encompasses more than half the structures at present in the Protein Data Bank (Lamzin *et al.*, 1995).

References

Betzel, C., Dauter, Z., Genov, N., Lamzin, V., Navaza, J., Schnebli, H.P., Visanji, M. & Wilson, K.S. (1993) Structure of the proteinase inhibitor eglin C with hydrolysed reactive centre at 2.0 Å resolution. *FEBS Lett.* **317**, 185-188.

CCP4 (1994) Collaborative Computational Project Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr.* **D50**, 760-763.

Cochran, W. (1952) A relation between the signs of sturcture factors. *Acta Crystallogr.* **5**, 65-67.

Damas, A.M., Ribeiro, S., Lamzin, V.L., Porto, J.A. & Saraiva, M.J. (1996) The crystal structure of Val-122-Ile variant transthyretin - a cardiomyopathic mutant. *Acta Crystallogr. D.* in press.

Dauter, Z., Sieker, L.C. & Wilson, K.S. (1992) Refinement of Rubredoxin from *Desulfovibrio vulgaris* at 1 Å with and without restraints. *Acta Crystallogr.* **B48**, 42.

Derrick, J.P. & Wigley, D.B. (1994) The third IgG-binding domain from streptococcal protein G. An analysis by X-ray crystallography of the structure alone and in a complex with Fab. *J. Mol. Biol.* **243**, 906-918.

Frey, M. Sieker, L.C., Payan, F., Haser, R., Bruschi, M., Pepe, G. & Le Gall, J. (1987) Rubredoxin from *Desulfovibrio gigas*. A molecular model of the oxidised form at 1.4 Å. *J. Mol. Biol.* **197**, 525.

Karle, J. and Hauptman, H. Application of statistical methods to the naphtalene structure. *Acta Crystallogr.* **6** (1953) 473-476.

Konnert, J.H. & Hendrickson, W.A. (1980) A restrained-parameter thermal-factor refinement procedure. *Acta Crystallogr.* **A36**, 344-350.

Lamzin, V.S. & Wilson, K.S. (1993) Automated refinement of protein models. *Acta Crystallogr.* **D49**, 129-147.

Lamzin, V.S., Sevcik, J., Dauter, Z. & Wilson, K.S. (1995) Implications of atomic resolution. *Making the most of your model. Proceedings of the CCP4 Study Weekend, 6-7 January, SERC Daresbury Laboratory, Daresbury, Warrington, England*, 33-40.

Langridge, R., Marvin, D.A., Seeds, W.E., Wilson, H.R., Hooper, C.W., Wilkins, M.H.F. & Hamilton, L.D. (1960) The molecular configuration of Deoxyribonucleic Acid. II. Molecular models and their Fourier transforms. *J. Mol. Biol.* **2**, 38-64.

Lunin, V.Y. (1993) Electron-density histograms and the phase problem. *Acta Crystallogr.* **D49**, 90-99.

Main, P. (1990) A formula for electron density histograms for equal-atom structures. *Acta Crystallogr.* **A46**, 507-509.

Navaza, J. (1994) AMoRe: An automated package for molecular replacement. *Acta Crystallogr.* **A50**, 157-163.

Podjarny, A.D. & Yonath, A. (1977) Use of matrix direct methods for low-resolution phase extension for tRNA. *Acta Crystallogr.* **A33**, 655-661.

Sheldrick, G.M. (1993) SHELXL-93, program for crystal structure refinement, *University of Göttingen*, Germany.

Ten Eyck, L. (1973) Crystallographic fast Fourier transforms. *Acta Crystallogr.* **A29**, 183-191

Tronrud, D.E. (1996) The TNT Refinement Package. In *Methods in Enzymology* (Carter, C. & Sweet, B. eds.) **In press**.

Wilson, A.J.C. (1942) Determination of absolute from relative X-ray data intensities. *Nature* **150**, 151-152

Removing Bias From A Model For HIV-1 Reverse Transcriptase By Real-Space Averaging Between Different Crystal Forms

Robert Esnouf[1,2,†], Jingshan Ren[1], Yvonne Jones[1,2], David Stammers[3] and David Stuart[1,2]

[1] Laboratory of Molecular Biophysics, The Rex Richards Building, South Parks Road, Oxford, OX1 3QU, UK;
[2] Oxford Centre for Molecular Sciences, New Chemistry Building, South Parks Road, Oxford, OX1 3QT, UK;
[3] Structural Biology Group, Glaxo–Wellcome Research Laboratories, Langley Park, Beckenham, BR3 3BS, UK,
[†] Current address: Rega Institute for Medical Research, Katholieke Universiteit Leuven, Minderbroedersstraat 10, B-3000 Leuven, Belgium.

Introduction

As the major enzyme target for anti-AIDS therapies, the reverse transcriptase (RT) of HIV has been the focus of intensive study. The enzyme has at least three functions: it is an RNA-directed DNA polymerase, a DNA-directed DNA polymerase and it also has RNase H activity. It functions as a heterodimer: the first chain (p66; 560 residues) contains the key catalytic residues for all three functions, the second chain (p51; 440 residues) is an *N*-terminal portion of the p66 resulting from proteolysis and is not known to be associated with any function.

Structural studies of RT have been directed mainly at HIV-1 RT and different groups have obtained a wide variety of crystal forms (for examples from our group see Jones *et al.*, 1993). However, few of these crystal forms showed useful diffraction. The first detailed structure report for RT was based on crystals diffracting to 3.4Å resolution (Kohlstaedt *et al.*, 1992) which were grown in the presence of a non-nucleoside RT inhibitor (NNRTI), nevirapine (Merluzzi *et al.*, 1990). The resolution of this structure determination has since been extended to 2.9Å (Smerdon *et al.*, 1994). A second medium-resolution structure (data to 3.0Å resolution) was also reported for RT in complex with DNA and an Fab (Jacobo-Molina *et al.*, 1993). Both structures showed an open, asymmetrical arrangement of five domains for the p66 subunit 'sitting' on the more compact base of the p51. The p51 subunit comprised the first four domains of the p66, but in a very different relative arrangement. The p66 structure was likened to a right hand, able to grip the DNA/RNA and the domains were named accordingly (domain names are shown in Figure 4). A comparison of these two structures showed that the molecule was very flexible and that large domain shifts were possible.

A crystal form of HIV-1 RT which reorders on dehydration

We had also obtained a crystal form for HIV-1 RT in complex with NNRTIs, in particular with the nevirapine analogue 1051U91 (Hargrave *et al.*, 1991). These crystals were of space group $P2_12_12_1$ with one heterodimer in the asymmetric unit and showed diffraction to a high-resolution limit of 3.4Å (Stammers *et al.*, 1994). We had observed that the crystals exhibited substantial non-isomorphism, apparently clustering into two groups with either $a\approx147$Å or $a\approx143$Å and $b\approx112$Å, $c\approx79$Å (designated cell forms A and B, respectively). Crude molecular replacement structures for these cell forms had been elucidated based on the Kohlstaedt *et al.* (1992) model (Esnouf, submitted) and showed electron density for all domains except the thumb domain of the p66 subunit. This crystal form promised to yield little new structural information until a chance observation was made whilst screening for heavy atom derivatives. During a series of exposures overnight on an in-house detector, repeated disordering and re-ordering of the diffraction pattern was observed from a single crystal. Initially well-ordered

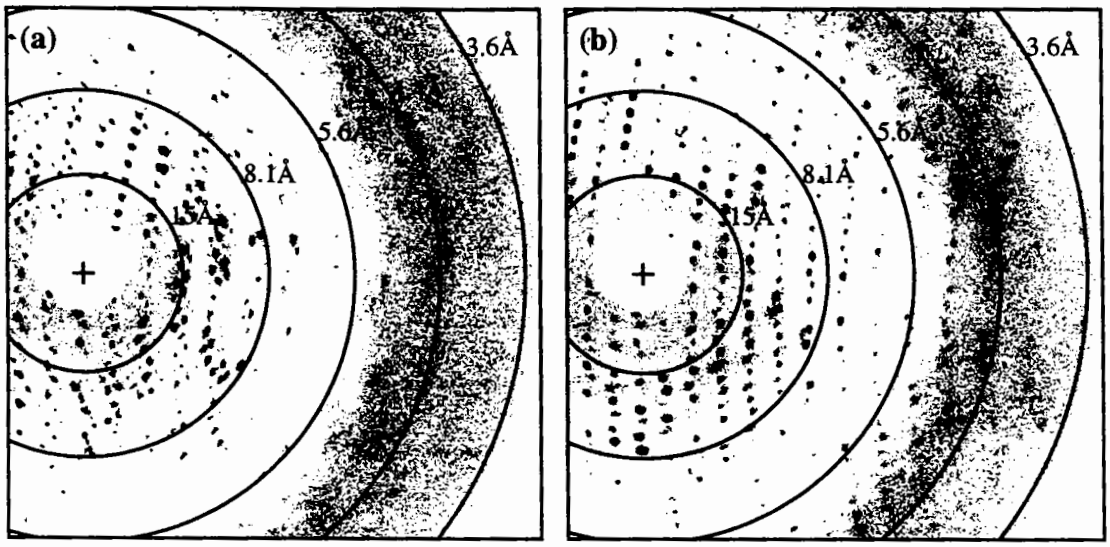


Figure 1: Diffraction images from an RT crystal changing cell form due to dehydration (a) image 27: during transition between forms C and D, (b) image 40: cell form D

diffraction had changed to show distinct evidence of a second lattice by the 4th image and by image 27 (Figure 1(a)) the diffraction was very disordered. However, diffraction recorded on images between these two and after image 27 was much better (see image 40, Figure 1(b)). On checking the crystal it was immediately apparent that the capillary tube containing it was imperfectly sealed and the mother liquor had evaporated.

Data frames for this crystal were processed individually and showed that the loss of mother liquor had led to changes in the unit cell dimensions of the crystal without changing the space group. Initially the cell form was as above with $a \approx 147$Å (form A), but over the first 8 images a conversion to the form with $a \approx 143$Å (form B) had occurred. The disorder in the diffraction during this conversion showed that they were, indeed, two distinct cell forms. However, this was only the tip of an iceberg and by image 40 (Figure 1(b)) the cell dimensions were $a \approx 142$Å, $b \approx 116$Å, $c \approx 66$Å, a contraction of 13Å in the c-axis. The last image from which a reasonably reliable unit cell could be measured (image 67) showed even greater dehydration: $a \approx 137$Å, $b \approx 113$Å, $c \approx 63$Å. Estimates for the mosaicity of the crystal at each image also varied as dehydration occurred showing maxima with disordered images (Figure 2). A fuller analysis

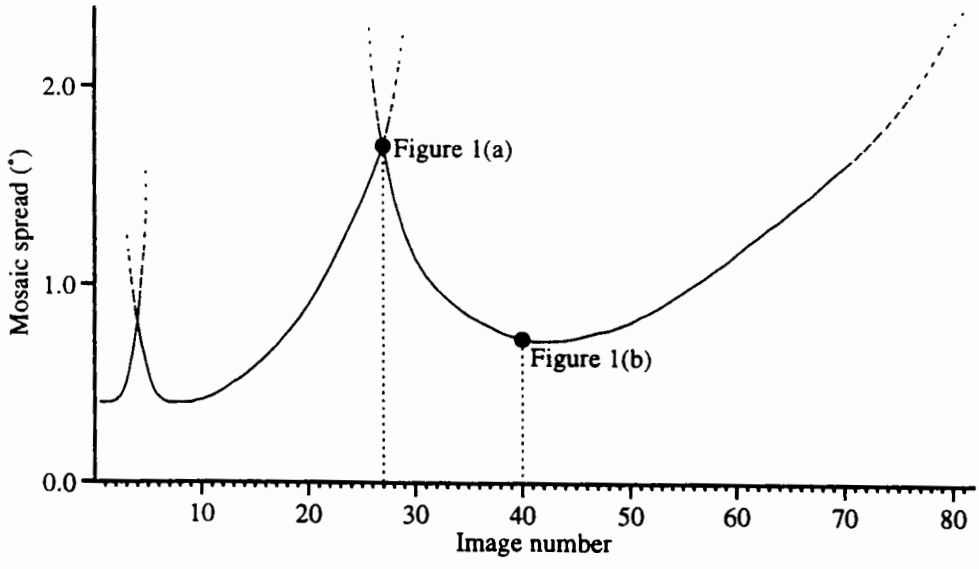


Figure 2: Variation in mosaic spread of the RT crystal during dehydration

of the changes occurring on dehydration will be presented elsewhere (Esnouf *et al.*, in preparation). That report shows that for images from the well-ordered dehydrated crystal states the fall-off of diffraction intensity with increasing resolution appears to be slower than for the less dehydrated cell forms, suggesting that dehydrated crystals might be capable of diffraction to higher resolution (especially using a synchrotron source). This observation prompted a search for ways of inducing crystal dehydration in a more controlled manner.

Data from deliberately dehydrated crystals

A dehydration protocol was developed based on transferring crystals between wells containing PEG 3400 solutions of increasing concentration (Stammers *et al.*, 1994). Although a significant fraction of crystals are damaged by the dehydration process, the loss is acceptable and those that survive are stable in a solution containing 46% w/v PEG 3400 for some months. Data collected from these crystals show that they are dehydrated to one of two end-points: either $a\approx141$Å, $b\approx111$Å, $c\approx73$Å (cell form C) or $a\approx142$Å, $b\approx116$Å, $c\approx66$Å (cell form D). Both crystal forms are capable of diffraction to a high resolution limit of at least 2.2Å at a suitable synchrotron source (Stammers *et al.*, 1994). When these crystals are used for cryo-crystallography a further cell reduction is observed yielding two further crystal forms, E and F (Esnouf *et al.*, 1995; Ren *et al.*, 1995b).

Crystal form	Form A	Form B	Form C	Form D
Diffraction limit	3.7Å	3.4Å	2.2Å	3.2Å
Number of reflections	13149	15872	43009	16158
R_{merge}	5%	13%	9%	7%
a-axis	147Å	143Å	141Å	142Å
b-axis	112Å	112Å	111Å	116Å
c-axis	79Å	79Å	73Å	66Å
Solvent content	56%	54%	50%	48%

Table 1: Datasets from the four different RT crystal forms

For the original structure determination, four datasets were available (summarised in Table 1). The original molecular replacement model was fitted to the data for cell form C by rigid-body refinement of the individual domains using X-PLOR (Brünger, 1992). Using sequence information from Jacobo-Molina *et al.* (1993), the structure was carefully refined against the high-resolution dataset (cell form C). Cycles of manual rebuilding and simulated annealing eventually produced a model with an R factor of 0.285 for all data from 10–2.3Å resolution. Whilst the model clearly contained errors, there was little evidence for how to correct these errors in electron density maps phased from it. Although our R factor was worse than for the published RT structures, this was indicative of the more extensive data providing more stringent constraints, rather than the model being worse. Indeed, when the R factor was calculated for data in the resolution shell 10–3.0Å after applying a 3σ cut-off on intensity to the data (*c.f.* Jacobo-Molina *et al.* (1993)) the R factor was only 0.181.

Real-space electron density averaging between cell forms

In the continued absence of useful experimental phasing information, partly because of the non-isomorphism of the crystals, an alternative refinement strategy was required: real-space electron density averaging between the cell forms relying on the troublesome non-isomorphism to supply phase restraints. That the datasets for different cell forms have phases largely

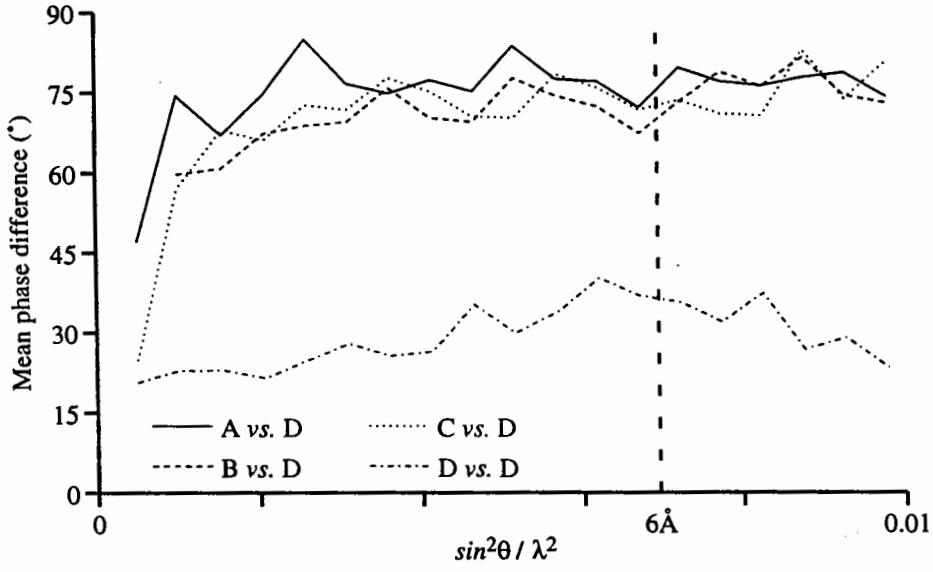


Figure 3: Phase independence of data from the four different RT crystal forms

independent of each other is demonstrated in Figure 3. The mean phase differences from the cell form D data were calculated as a function of resolution based on our current best models for each cell form. For comparison, we also show the mean phase difference between two cell form D datasets (RT in complex with two different NNRTIs). Since completely uncorrelated phases would have a mean phase difference of 90°, it can be seen that even at moderate resolutions these datasets are largely independent.

As well as averaging between cell forms, the domain structure of the RT heterodimer (Figure 4) also has a degree of internal non-crystallographic symmetry (NCS) which can be used for *intra* cell form averaging. The five domains at the top of the figure are from the p66 subunit and the first four of these are repeated in the p51 subunit below. Although the different domain arrangement of the subunits causes some change in the internal structure of equivalent domains, 'core' regions of each domain can be defined where the structures are similar (using the program SHP (Stuart *et al.*, 1979)). Since we had no model for the p66 thumb domain at this stage we were left with three pairs of domain cores for internal NCS averaging.

Averaging was performed using the program GAP (Grimes and Stuart, unpublished) in conjunction with the CCP4 suite for FFTs (CCP4, 1994). Our first protocol was based solely

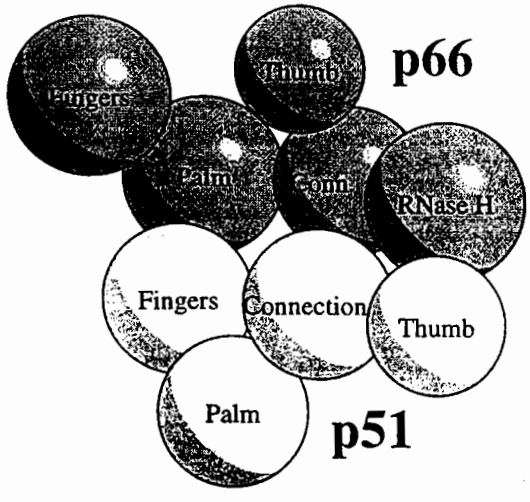


Figure 4: Domain structure of the RT heterodimer

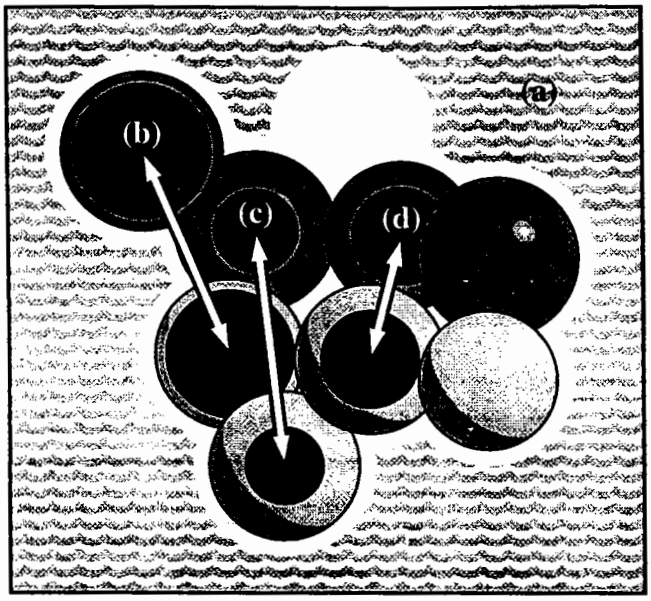


Figure 5: Protocol using solvent flattening and approximate NCS between domains (a) solvent flattening (b) fingers- (c) palm- and (d) connection-domain averaging

on cell form C using data to a high resolution limit of 3Å (Figure 5): a starting $2F_{obs}$-F_{calc} map was solvent flattened using the Wang method (Wang, 1985), the NCS operators were refined and then two-fold averaging of the core regions of the fingers, palm and connection domains was performed. The averaged map was back-transformed, scaled to the F_{obs} and used to calculate a $2F_{obs}$-F_{calc} map using the new structure factor amplitudes and phases. This cycle was repeated until the statistics used to monitor progress showed it had converged (correlation coefficients and R factors both in real and reciprocal space). This internal averaging required the definition of 8 envelopes for the molecule and for individual domains and the use of 6 NCS operators relating the pairs of domains. Not unexpectedly, the averaged maps from this procedure showed little improvement on the starting maps. However, the work was useful for ironing out difficulties in the protocol before extending it to more complex cases including *inter* crystal form averaging.

When extending the protocol to include averaging between crystal forms the difference in the quality of the datasets has to be considered. Firstly, the diffraction for crystal form C was much stronger than that from the other crystal forms. We chose to compensate for this as simply as possible by sharpening the data for each of the other crystal forms to match the form C data using an isotropic B factor. Secondly, the other datasets did not extend to as high a resolution as the crystal form C data (not even as far as the 3Å cut-off used in the averaging protocol). As a means of increasing the effectiveness of the cross-averaging for the higher resolution data, synthetic structure factors were used in the map calculations. For the first few cycles no synthetic data were used and then they were gradually introduced in thin shells of increasing resolution up to the 3Å limit.

For cross-averaging we took our current 'best' form C model and fitted it to the other datasets by rigid-body refinement of individual domains. This fitting procedure gave us starting 'NCS' operators relating the position of each domain in one cell form to its counterparts in the other cell forms. We assumed that the internal structure of domains was very similar in all cell forms and so defined envelopes enclosing virtually the whole of each domain. As the refinement proceeded our envelope definition became more sophisticated, but essentially confirmed this initial assumption. The averaging cycle was modified to include synthetic data in the map calculation, to refine the *inter* cell form 'NCS' operators and to

perform *inter* cell form electron density averaging. Other refinements included monitoring phase shifts and a 'free *R*' calculation. When including cross averaging we were careful to keep the process equivalent for each cell form. Internal Wang solvent flattening and two-fold averaging for each cell form was performed first in the manner described above. These internally averaged maps were copied and the duplicate map for each cell form was then *n*-fold averaged (n = 2, 3, or 4) with the original (internally averaged) maps for the other cell forms. Averaged maps for each cell form were then back-transformed, scaled to the appropriate datasets and used to calculate new maps for the next cycle.

Averaging this way between cell forms B and C required the definition of 28 'NCS' operators and 32 envelopes, but again the result was disappointing. Including data for cell form D took the number of operators up to 66 and the number of envelopes to 48. This increase in complexity proved worth the effort, however, and the averaged maps provided clear guidance for manual rebuilding. Amongst the improvements in the averaged maps were evidence for several errors in sequence alignment and the appearance of electron density allowing the positioning of the p66 thumb domain (see below). This rebuilt model was refined by simulated annealing (using X-PLOR (Brünger, 1992)) and then used to provide a better starting point for further averaging. One final data set (for cell form A) was also 'stirred into the pot' along with the newly-found p66 thumb domain. This required a total of 140 'NCS' operators to be defined in order to average density in 76 envelopes.

Results of the averaging protocol

Monitoring the progress of such a complex protocol is no easy task. Each averaging step on each domain of maps in each cell form produces a real-space *R* factor and correlation coefficient. The scaling of each set of averaged structure factors to the relevant dataset produces statistics for the reciprocal space part of the cycle. With so many operators and envelopes, individual errors were easy to make and such errors did not necessarily have catastrophic effects on the averaging as a whole. While setting up the protocol each indicator was followed separately and the number of pixels being averaged was cross checked as far as possible. Space does not permit a full analysis here, but a few representative numbers illustrate the success of the procedure.

The improvement in the real-space correlation coefficients for the internal NCS averaging (Figure 6) was quite dramatic. Not only were the correlation coefficients higher, but also an

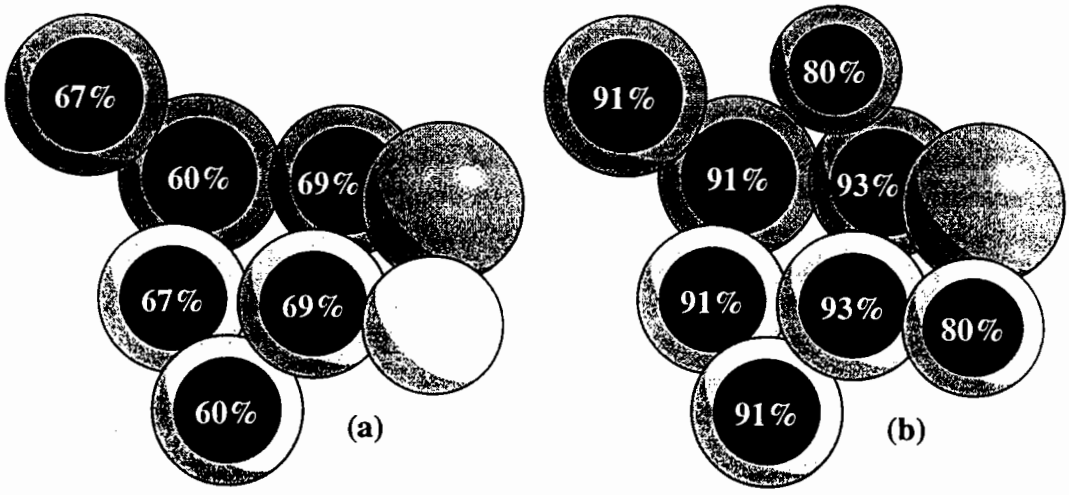


Figure 6: Mean correlation coefficients of domains for internal NCS averaging
(a) at start of NCS-only averaging, (b) at end of 4-fold cross-averaging

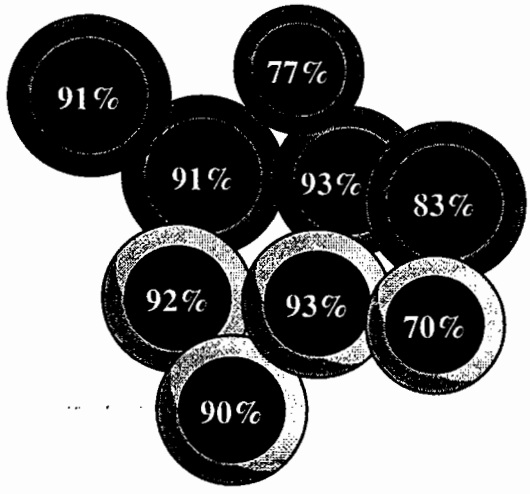


Figure 7: Final mean correlation coefficients of domains for 4-fold cross-averaging

increasing percentage of each domain could be included in the definition of the 'core' envelopes after each rebuild. Whilst some very significant differences in the internal structures of the p66 and equivalent p51 domains remained, many of the previously reported smaller differences turned out to be artifactual. The two fingers domains showed the greatest degree of convergence: the root-mean-square difference in the Cα positions of our 'core' region for these domains dropping from 1.74Å to 1.09Å by the end of the refinement against the four cell forms.

The cross-averaging statistics are even more numerous and we just give the final mean values (over all four cell forms) of the real-space correlation coefficients for each domain (Figure 7). The values for the fingers, palm and connection domains are all very good (>90%). The correlation for the RNase H domain is somewhat lower (83%) and this appears to reflect differing degrees of disorder as well as conformational differences for this domain amongst the cell forms. Lower still are the correlation coefficients for the thumb domains, especially the p51 thumb. This may be due to the definition of the envelopes for this domain being rather sub-optimal since the conformation of the p51 thumb domain is affected by the neighbouring RNase H domain. The internal averaging can then account for the 'knock-on' effect of a lower correlation for the p66 thumb domains.

For the final cycle of four cell form averaging the mean correlation coefficient in reciprocal space was 92% and the mean *R* factor was 17%. However, the relevant measure of success of the protocol was how well the averaged maps had escaped the original model bias and showed how our model should be rebuilt. From this perspective, it was the maps resulting from the three crystal form protocol that were the most valuable, and over two rounds of manual rebuilding the *R* factor for our model dropped substantially. One of the major factors in this improvement was the emergence of connected density for the p66 thumb domain in the averaged maps (Figure 8). The model for the p51 thumb domain was found to superpose well on this density and so the completeness of our model was improved dramatically. As a result of the averaging / rebuilding / simulated annealing cycle the *R* factor for our model of the RT heterodimer in cell form C was reduced from 0.285 for data from 10–2.3Å resolution to 0.214 for all data from 25–2.2Å resolution (Ren *et al.*, 1995a).

Conclusions

Real-space averaging is a powerful tool for exploiting the phase restraints that result from non-crystallographic symmetry — in ideal cases with many-fold NCS the results can be

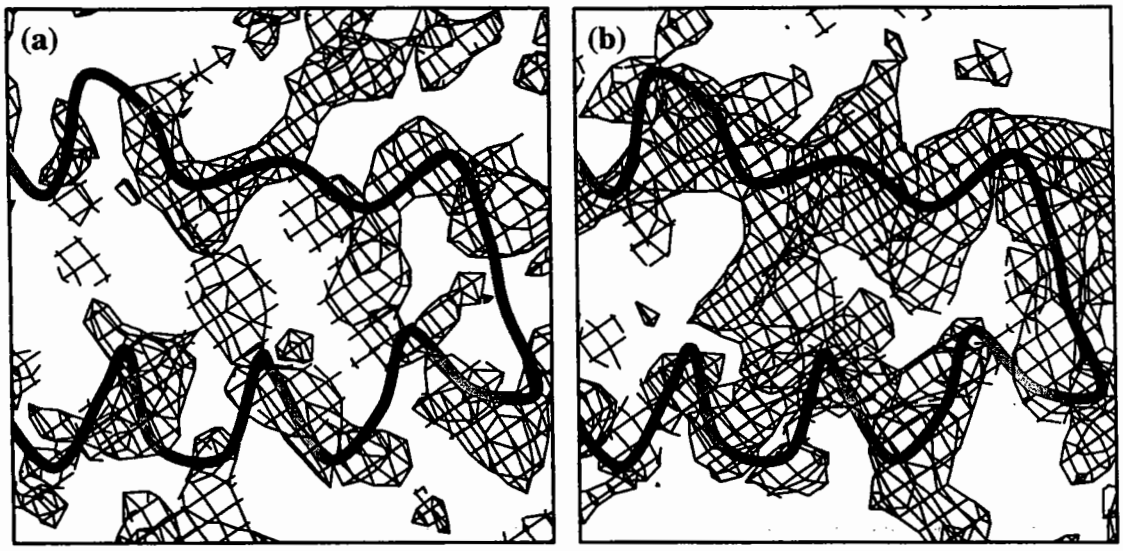


Figure 8: Density for part of the p66 thumb domain and the final chain trace
(a) at the start of 3-crystal form averaging, (b) after 14 cycles of averaging

dramatic. However, even in cases where the symmetry is only approximate the benefit can be crucial in allowing a structure to be refined successfully. For the refinement of HIV-1 RT the method was extended to include averaging between maps obtained from different datasets. To allow this required further development and testing of the real-space averaging program GAP, as well as devising a protocol to cope with the differing (and indifferent!) quality of some of the data. The setting up of such a protocol was a very time-consuming exercise.

The averaging / rebuilding / annealing cycles were also time-consuming, keeping three workstations almost fully occupied for two months. However, computer time is relatively cheap and there was no other way of reliably improving the phases. Thus, the non-isomorphism of the RT crystals which had bedevilled efforts to determine the structure for so long, was, in the end, turned to our advantage and allowed us to produce a well-refined structure. Work on RT has not stopped and our current best model (refined against cell form D data in the resolution range 25–2.2Å) has an R factor of 0.186 (Ren *et al.*, 1995a).

Whilst cross crystal form averaging could hardly be recommended as a refinement strategy of choice, it has been used (albeit in somewhat simpler situations) for a number of other studies. These studies range from the early incomplete example of influenza neuraminidase (Varghese *et al.*, 1983) and include the more recent example of HLA class II (Brown *et al.*, 1993). It is likely that there will be other cases where no other avenue is open. It is worth noting that the unit cells for two crystal forms do not have to be very different for (at least higher resolution) data to have substantial phase independence: a difference in one axis of 2–5Å may well be sufficient. Such changes in unit cell dimensions are not uncommon when crystals are flash-cooled and hence averaging between data collected at room temperature and data collected at cryogenic temperatures may well be a relatively general and useful method of phase improvement (as, for instance, in the application to SIV matrix antigen (Rao *et al.*, 1995)).

References

Brown, J. H. *et al.*, *Nature*, **364** (1993), 33.
Brünger, A. T., *X-PLOR v. 3.1*, Yale University, New Haven, CT (1995).
CCP4, *Acta Cryst.*, **D50** (1994), 760.
Esnouf, R. M. *et al.*, *Nature Struct. Biol.*, **2** (1995), 303.
Hargrave, K. D. *et al.*, *J. Med. Chem.*, **34** (1991), 2231.
Jacobo-Molina, A. *et al.*, *Proc. Natl. Acad. Sci. USA*, **90** (1993), 6320.
Jones, E. Y. *et al.*, *J. Crystal Growth*, **126** (1993), 261.
Kohlstaedt, L. A. *et al.*, *Science*, **256** (1992), 1783.
Merluzzi, V. J. *et al.*, *Science*, **250** (1990), 1411.
Rao, Z. *et al.*, *Nature*, **378** (1995), 743.
Ren, J. S. *et al.*, *Nature Struct. Biol.*, **2** (1995a), 293.
Ren, J. S. *et al.*, *Structure*, **3** (1995b), 915.
Smerdon, S. J. *et al.*, *Proc. Natl. Acad. Sci. USA*, **91** (1994), 3911.
Stammers, D. K. *et al.*, *J. Mol. Biol.*, **242** (1994), 586.
Stuart, D. I. *et al.*, *J. Mol. Biol.*, **134** (1979), 109.
Varghese, J. N. *et al.*, *Nature*, **303** (1983), 35.
Wang, B. C., *Methods Enzymol.*, **115** (1985), 90.

Improving Electron Density Maps Calculated from Weak or Anisotropic Data

S.J. Gamblin†, D.W. Rodgers* and T. Stehle*

† Protein Structure Division, NIMR, Mill Hill, London, NW7 1AA
* Department of Molecular & Cellular Biology, Harvard University
7 Divinity Avenue, Cambridge, MA 02138

Introduction

Given the choice one would always prefer to have crystals which diffract strongly and isotropically to high resolution. However, there may be rare occasions when protein crystallographers are forced to work with less ideal diffraction data. This paper will discuss two particular shortcomings of diffraction data and procedures which can be used to minimise their effects; significant overall anisotropy and high Wilson B-factors. Several other papers in these proceedings deal with the determination of improved weighting schemes for crystallographic refinement protocols. This paper will be concerned with the modification of observed structure amplitudes for calculation of 'improved' electron density maps. These two approaches are closely related but have different immediate objectives. The examples used to illustrate these approaches will be based mainly on our experience with the crystallographic structure determination of HIV-1 RT and SV40 virus; these examples have 4 and 5-fold non-crystallographic symmetry respectively. Non-crystallographic symmetry is not necessary for the corrections described but does not detract from their effectiveness.

Anisotropic Diffraction

Sometimes it is evident from simply looking at raw data images that a crystal diffracts more strongly in one direction than another. Even if this property is not immediately evident from visual inspection, a given crystal may well be better ordered in one or two directions than the others. Indeed, in the absence of an overwhelming argument (such as a cubic space group), it is always safest to assume that diffraction is anisotropic.

Given that the diffraction behaviour of crystals often displays some overall anisotropy, what are the consequences of this and what action should be taken? Viewed at its simplest, anisotropic diffraction is a consequence of anisotropy in the packing of the crystal. This anisotropy arises from the way in which the molecules pack in the crystal lattice, not because of any innate asymmetry in the order of the molecule itself. It is evident then, that an

overall anisotropic correction (described by 6 parameters) of the observed amplitudes will often be required to optimise the information that can be extracted from it. Of course, the asymmetry in the order of the crystal has a cost; there is a lack of detailed information about the molecule in a certain direction.

In practice, the treatment of anisotropy corrections depends on whether the principle axes describing the anisotropy are aligned with the reciprocal space axes of the crystal. If these axes are aligned then the correction will not materially affect the reduction of the observed diffraction measurements subsequent to the data scaling. In other words, symmetry related reflexions will be equally affected by the anisotropy correction. In these cases it is quite legitimate to calculate and apply the correction after scaling and reduction of the observed diffraction data to the unique segment of reciprocal space. In the case where the axes of the anisotropy of the crystal are not aligned with the reciprocal space axes, then the correction must be calculated and applied prior to the averaging of symmetry equivalent reflexions.

The calculation of the anisotropy correction can either be done based on the distribution of observed intensities within the dataset itself or with respect to some external reference (such as a set of calculated amplitudes from atomic coordinates). Many programs deal with the latter situation but the former case has often been dealt with in a rather *ad hoc* fashion. Brenda Temple at UNC-Chapel Hill has modified versions of CCP4 programs in order to calculate scale factors and anisotropic B-factors for data merging. These modified programs have been shown to work well and are available from her on request.

The potential advantage to be gained from applying even a relatively modest anisotropic correction prior to carrying out a molecular replacement calculation is illustrated by the example below. This example is taken from a molecular replacement calculation carried out by T. Barrett at NIMR on the lectin protein TCA. The correlation coefficients shown represent the values obtained from first fitting a single molecule and then both of the molecules present in the asymmetric unit. The anisotropy correction was such that the B-factor difference between the strongest and weakest direction was approximately 30Å^2.

	Correlation Coefficient	
	uncorrected	corrected data
Molecule 1	0.251	0.294
Molecule 1 & 2	0.549	0.595

Sharpening Weak Data

As the technology of protein crystallography advances, increasing numbers of important biological problems become amenable to crystallographic analysis. Enormous developments have occurred in the intensity and brilliance of synchrotron sources, X-ray detectors and cryogenic procedures. This means that crystals which just a few years ago would have been; too small, too poorly ordered, or the unit cell just too large, are now tractable. What this often means with difficult crystals is that diffraction data is measured which falls off rather quickly as a function of resolution. It is quite feasible to collect reasonable quality, high resolution, diffraction data which has a Wilson B-factor in the rather alarming range of 50-80 $Å^2$. Three questions arise; firstly, what causes this kind of behaviour, secondly, what is the effect on electron density maps and finally, what can be done to ameliorate the effect.

High Wilson B-factors reflect poor order in the packing of the crystal. The disorder in the crystal can arise in many ways but the fact is that the crystal may still produce diffraction to high Bragg angles. In other words the molecule in question (or at least certain parts of it) are still giving rise to relatively high resolution information. As before, the relatively high disorder in the crystal (be it static or dynamic) is a consequence of the crystal lattice and not just the characteristics of a single molecule.

Electron density maps calculated from this kind of uncorrected data will be dominated by low resolution reflexions. Higher resolution detail will generally be missing from the electron density maps simply because all of the terms which represent these details in the Fourier calculation have small amplitudes. These weak data may or may not be accurately measured. Whilst it is easier to make reliable measurements of strong amplitudes, there is no absolute reason why weaker reflexions should not be measured to good accuracy. Given suitable multiplicity of observations, even reflexions that are measured at just 1 to 2 sigma on a given data image may nonetheless be determined with some certainty. Ultimately, the quality of electron density maps are limited by the resolution, reliability and quantity of measurements. The point that we will illustrate here is that the interpretability of electron density maps may well be enhanced considerably by increasing the relative contribution of high resolution Fourier terms. The sharpening procedure we have used is to apply a negative B-factor to the observed data amplitudes.

The use of a simple exponential sharpening factor is an intuitive, rather than rigorously derived, correction factor but it does meet the requirement of producing better electron density maps. The size of the B-factor applied to the

dataset depends on; the Wilson plot of the data, how reliable the measurements are, how complete the data is and the amount of non-crystallographic symmetry and solvent present. The examples below deal with crystals with 4 and 5-fold non-crystallographic symmetry and so the innate value of the weak high resolution data is approximately twice what it would be in the absence of this symmetry. The correction factors used in the examples were determined by inspection of various electron density maps.

HIV-1 Reverse Transcriptase

The size and flexibility of the HIV-1 RT molecule caused problems for many years in producing diffraction quality crystals. Indeed the molecule was eventually crystallised either by the addition of non-nucleoside inhibitors (1) or antibodies and DNA duplex (2). To date, no strongly diffracting crystals of apo-RT have ever been produced. In response to this we decided to pursue the structure determination from relatively small (150 x 50 x 30 microns) crystals with high (70%) solvent content and a largish asymmetric unit (~500K/au) (3). These crystals were only modestly well ordered (diffraction limit ~3Å) but the presence of 4-fold ncs suggested that the project was worth pursuing. The real problem was that the diffraction beyond 6Å was extremely weak. We developed cryogenic procedures which enabled us to transport a library of pre-screened crystals to CHESS. This was necessary because the exposure times required (even using the very brilliant F1 beamline) was of the order of 20-30 minutes for each 0.5° oscillation. The consequence of this was that even at 100K, single crystals deteriorated rapidly after collecting 5-8° of data. Finally, a dataset was produced with observations from 34 crystals. The data statistics are summarised below as a plot of $I/\sigma I$ (figure 1a). Clearly, these data are somewhat underwhelming. Nevertheless, with a sharpening B-factor of 50 and non-crystallographic symmetry averaging, the final electron density map was extremely informative for many parts of the molecule. There were some parts of the density map which were not well defined but they represent parts of the molecule which are highly mobile. This structure remains the only true unliganded crystal form of RT and provides a wealth of information with regard to domain mobility and drug binding and function. The figures 1b and 1c show samples of electron density maps at various stages of the structure determination. The dramatic improvement afforded by data sharpening stems not only from the enhancement of the high resolution terms directly but also from the concomitant improvement in the averaging transformations.

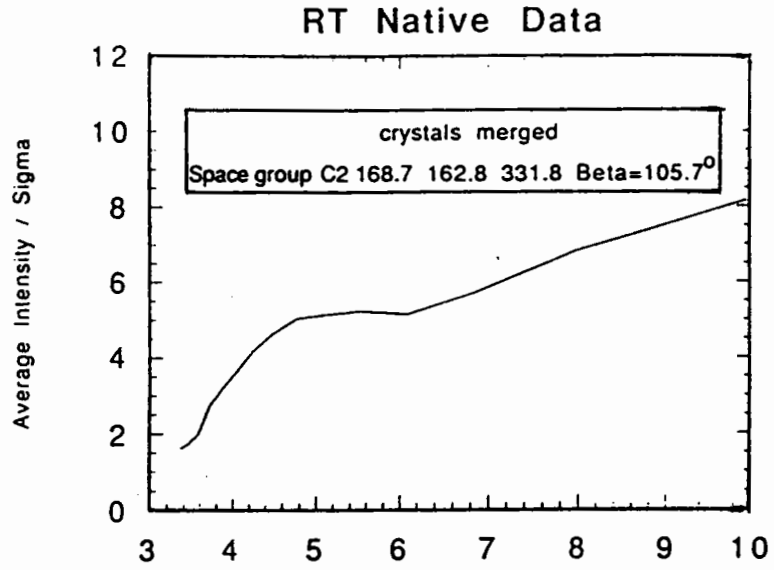


Figure 1a

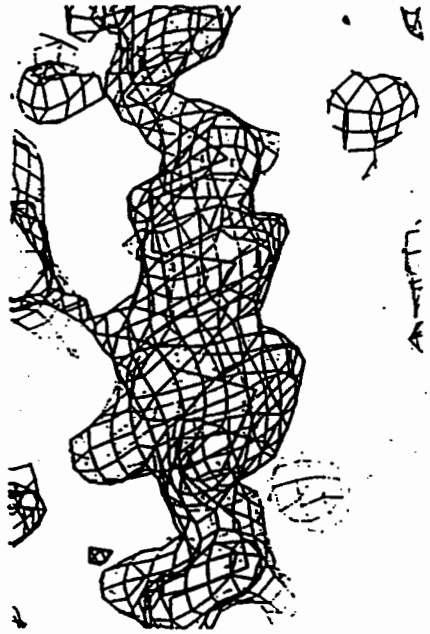


Figure 1b

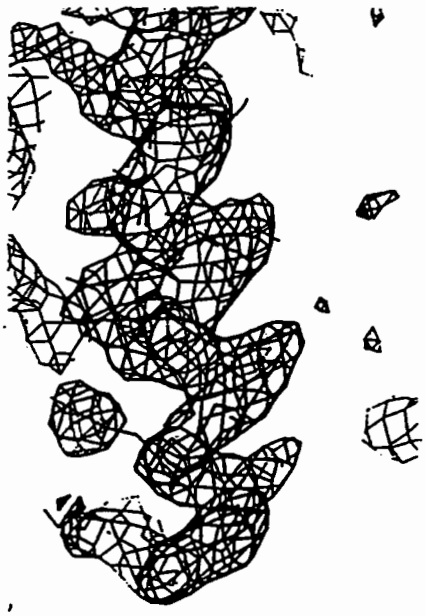


Figure 1c

SV40

SV40 was originally solved to 3.8Å resolution (4) and the problem of extending it to higher resolution was essentially one of crystal longevity in the beam. Before we knew how to successfully handle virus crystals at 100K there was no alternative but to collect and merge data from many crystals. The crystals belong to space group I23 with a=558Å and contain two complete virions in the unit cell. The crystals of SV40 used for data collection were 500-800 microns in thickness and single 0.2° oscillations were recorded from each unique volume of the crystal. In total, 45 crystals were used to collect a 3.1Å dataset (5). The Wilson plot of this dataset is shown both before and after application of a sharpening factor (B=40Å^2) in figure 2a. The final electron density map was produced by 5-fold ncs averaging and phase extension

starting from initial phases from 12.0-5.0Å. The accuracy of this phase extension process for the sharpened and unsharpened data is illustrated in figure 2b. By monitoring both the free correlation coefficient at each phase extension step and by inspections of the electron density maps it was evident that the phase extension was much more powerful using the sharpened dataset. Indeed in the absence of sharpening, the phase extension procedure would not proceed beyond 3.8Å. The final averaged electron density map was of exceptional clarity showing most side chains and the orientation of many main-chain peptide bonds. This is illustrated in Figures 2c and 2d, which show a segment of the average electron density map before and after data sharpening.

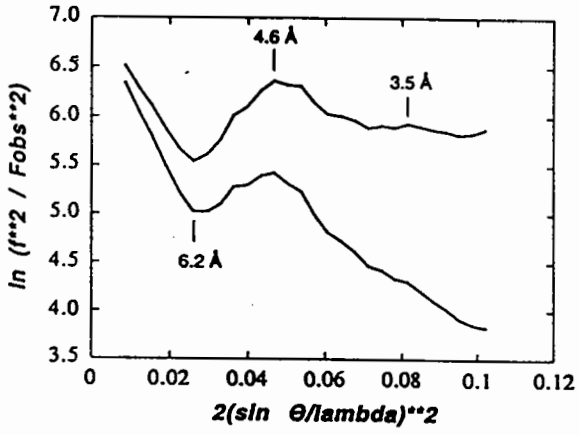


Figure 2a

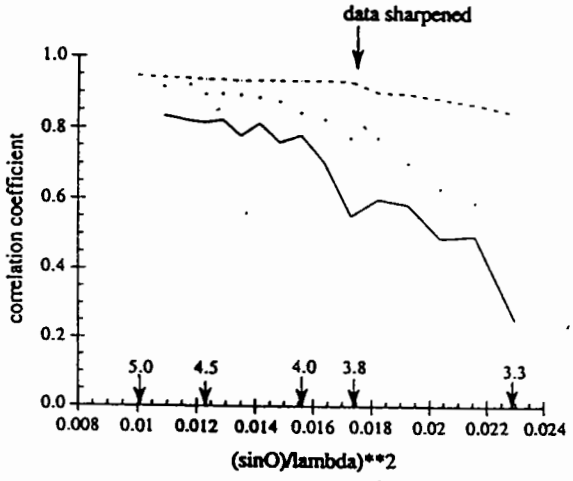


Figure 2b

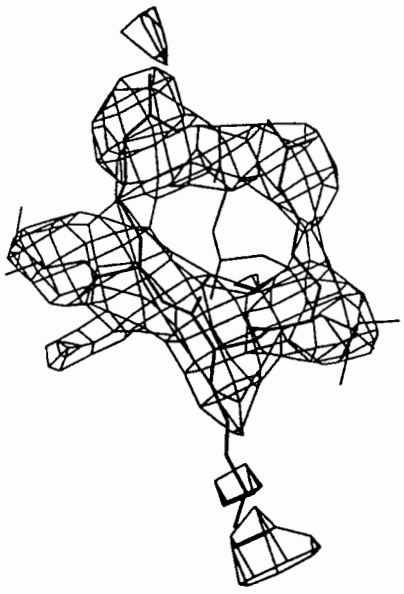


Figure 2c

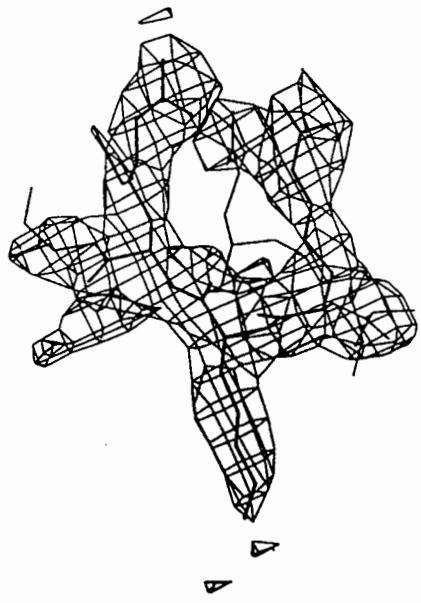


Figure 2d

Acknowledgements

Most of the work described here was supported by NIH grants CA-13202 and GM-39589 to S.C. Harrison, to whom we are also indebted for encouragement and vision. We thank also our colleagues both at Harrison/Wiley laboratories and at NIMR.

References

1. Kohlstaedt, L.A., Wang, J., Friedman, J.M., Rice, P.A. & Steitz, T.A. (1992) *Science* **256**, 1783-1790.

2. Jacobo-Molina, A., Ding, J., Nanni, R., Clark, A.D., Lu, X., Tantillo, C., Williams, R.L., Kamer, G., Ferris, A.L., Clark, P., Hizi, A., Hughes, S.H. & Arnold, E. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 6320-6324.

3. Rodgers, D.W., Gamblin, S.J., Harris, B.A., Ray, S. Culp, J.S., Hellmig, B., Woolf, D.J., Debouck, C. & Harrison, S.C. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 1222-1226.

4. Liddington, R.C., Yan, Y., Moulai, J., Sahli, R., Benjamin, R.L. & Harrison, S.C. (1991) *Nature* **354**, 278-284.

5. Stehle, T., Gamblin, S. J., Yan, Y. & Harrison, S. C. (1996) Structure Vol 4, No 2, 165-182.

Pseudo Symmetry

David Watkin
Chemical Crystallography Laboratory
9 Parks Road
OXFORD OX1 3PD, England

Abstract: This paper reminds readers that the conventional least-squares technique will inevitably fail if used to refine a model in a low symmetry space group generated from a model in a space group of higher symmetry related to the real one by a centre of symmetry or transition to a supercell. It describes some techniques which can be used to improve the conditioning of the matrix, and speculates on methods which might be suitable for determining whether the result of a refinement is meaningful.

Background: When a crystal structure contains motifs which are repeated such that there is a simple spatial relationship between their orientation and displacement, the motifs are said to be related by symmetry. The space group symmetry elements constitute well-characterised groups of symmetry relationships. If the matrix relating one motif to another is not a space group symmetry operator, then the symmetry is said to be 'non-crystallographic'.

There are two broad classes of pseudo-symmetry which occur in crystal structure analysis. One class is generally beneficial, the other aggressively disruptive. In the first class, the pseudo-symmetry does not degrade the least squares normal matrix. This non-crystallographic symmetry has been widely exploited in protein crystallography during structure development. It has also been widely exploited in small molecule analyses during least squares refinement.

The second class is concerned with the analysis of structures which are nearly centro-symmetric, or have a near perfect superlattice. The problem has been perceived for a long time (J.S. Rollett, (1970), Crystallographic Computing, ed Ahmed, Munksgaard, Copenhagen, O. Ermer & J. Dunitz, (1970), Acta Cryst A26, 163, V. Schomaker & R.E. Marsh, (1979), Acta Cryst B35, 1933-1934). No exact mathematical tool exists for the treatment of this problem.

Least squares: The principal procedures for refinement not dependent upon direct modification of a computed electron density map are based upon least-squares minimisation. Because the structure factor expression is non-linear in the parameters to be evaluated, the equation must first be linearised, usually through a truncated Taylor series.

Equation 1 shows how, in the linear case, an observation y is related to some known quantities a by n unknown but sought after quantities x.

When there are exactly n independent observations, the simultaneous

$$\sum_{i}^{unknowns} a_i \,.\, x_i = y \qquad 1$$

$$A \,.\, x = y \qquad 2$$

$$\left(A^t.A\right).\, x = A^t \,.\, y \qquad 3$$

$$x = \left(A^t.A\right)^{-1} .\left(A^t \,.\, y\right) \qquad 4$$

$$x = N^{-1} .\left(A^t \,.\, y\right) = N^{-1} \,.\, y' \qquad 5$$

$$x_n = N_{n,n}{}^{-1} .\left(A^t{}_{n,m} \,.\, y_m\right) = N_{n,n}{}^{-1} \,.\, y'_n \qquad 6$$

$$\sum^{atoms} f_i e^{-T} Cos\left(\underset{\sim}{h}.\underset{\sim}{x}\right) = Fc \approx Fo \qquad 7$$

$$Fc + \sum_{i}^{parameters} \frac{\partial Fc}{\partial x_i} .\, \delta x_i = F'c \approx Fo \qquad 8$$

$$\sum_{i}^{parameters} \frac{\partial Fc}{\partial x_{\ i}} \,.\, \delta x_i \approx Fo - Fc \qquad 9$$

$$x = \begin{vmatrix} * & * & * \\ * & * & * \\ * & * & \left(\sum_{m}^{observations} \frac{\partial F}{\partial x_i} \cdot \frac{\partial F}{\partial x_i}\right)^{+} \end{vmatrix} \cdot \begin{vmatrix} * \\ * \\ \sum_{m}^{observations} \frac{\partial F}{\partial x_i} \,.\, (Fo - Fc) \end{vmatrix} \qquad 10$$

equations can be solved for the unknowns. The operations in equations 2-6 show one way the problem can be solved when there are more observations than unknowns - this solution is the least squares solution to the problem.

Equation 7 gives the form of the structure factor equation, where the unknowns are x and terms in the temperature factor (adp), the observations are the structure factors, and the 'knowns' are the reflection indices. This equation is non-linear in the unknowns, but with a first order Taylor expansion and rearrangement, gives equation 8 & 9, which are analogous to 1, and so can be solved by the least squares method. Equation 10 shows some of the elements of the normal equations in detail. This derivation is in terms if the structure magnitudes $|Fo|$, but it may easily be modified to use the squared magnitudes. Several authors (J.S. Rollett, T.G. McKinlay & N.P.H. Haigh, (1976), Crystallographic Computing, ed. F.R. Ahmed, Munksgaard, E. Prince (1994), Springer-Verlag, Berlin) have shown that both procedures should yield the same parameter values if appropriate weighting schemes are used.

Observations of restraint: In equation 10, the observations are the structure magnitudes, and the unknowns are the atomic and other parameters. Other *observational equations* can be written in terms of 'observable' quantities and the atomic parameters, and if these equations are also expanded (so that the unknowns appear as parameter shifts) and are properly weighted, they can be added into the summations in equation 10. Some useful observations of restraint are:

1. Geometric features. Accumulated experience gives crystallographers good ideas about the expected values of geometrical features, such as interatomic distances, inter-bond angles, and torsion angles, and other features such as planarity or chirality. These features can generally be expressed as a function of the atomic coordinates, a target value assigned, together with an estimate of the reliability of the target value.

   e.g. for a distance restraint $(\Delta x)^t\, G\, (\Delta x) = D^2$

   (Watkin, (1988) Crystallographic Computing 4, ed N.W. Isaacs & M.R. Taylor, Oxford University Press, Oxford)

2. Molecular similarity. When molecular fragments repeat in a structure, it may be appropriate to propose that molecular parameters are similar, without actually pre-assigning a value to the parameters. These restraints can be used impose symmetry (2 or 6 fold) on phenyl groups, to impose non-crystallographic symmetry on peptide residues if 1-2, 1-3 and 1-4 distances are all restrained, or to impose local symmetry on peptide chains if only 1-2 and 1-3 distances are restrained to be similar.

3. Shift limiting restraints. These are mathematically related to the Marquardt-Lavenberg method for improving the convergence of least squares refinement, but without numerical optimisation. The equation of

restraint is trivial, and states that the value of the parameter should be the same after the refinement as it was before. This equation must be weighted, and is added in with all the other equations which have some effect on the parameter, so that the parameter may in fact move, but the movement will be restrained. This restraint is particularly useful when the normal matrix contains little or no information about one or more parameters being refined, for example when reducing the space group symmetry.

$1.0 \; . \; x_{new} = x_{old}$

(Watkin, (1988) Crystallographic Computing 4, ed N.W. Isaacs & M.R. Taylor, Oxford University Press, Oxford)

4 Non-crystallographic symmetry. If some quite large part of the structure appears to be related to another part of the structure by some non-crystallographic operator, this information can be encoded into the refinement as a matrix of constraint (in which case the non-crystallographic symmetry will be obeyed exactly), or as observations of restraint. Since equations of restraint are not binding upon the refinement, they permit some deviations from exact pseudo-symmetry. Equations of constraint reduce the number of variables being adjusted, and so might be a useful way to start the refinement, but it seems physically unlikely that they will be appropriate in the final stages of the analysis. Imagine two molecular fragments approximately related by the operator

x, .5-y, z+.25

This is a kind of glide plane. Similar operators are often found relating 'independent' molecules in an asymmetric unit. If the molecules are chiral, then of course the glide plane cannot hold exactly. The matrix of constraint is

$$y = \begin{matrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{matrix} y' + \begin{matrix} .5 \\ \\ +.25, \end{matrix}$$

which can be differentiated to give the shift constraining matrix.

Contributions from the atom at x′ are simply added into the summations for the base atoms at x.

The three equations of restraint are

$$x = x'$$
$$0.5 - y = y'$$
$$z - .25 = z',$$

which can be differentiated to give the contributions to be added into the normal equations.

Uses of restraints: Restraints are used to 'tell' the mathematics something

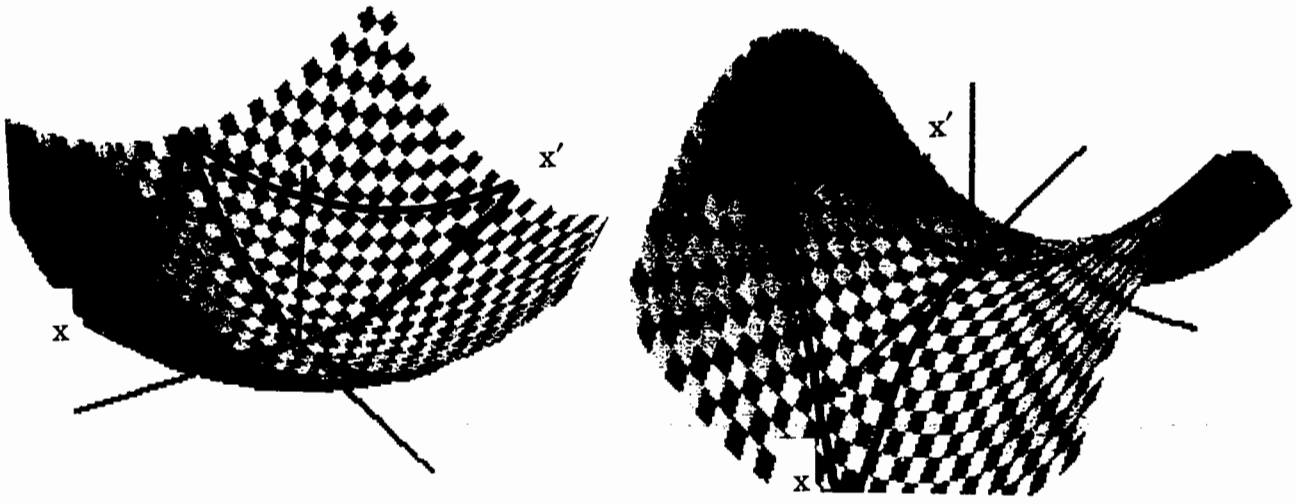
which the analyst knows about the real world. It is not uncommon in mathematics for a set of equations to have more than one solution. In some situations, all solutions are interesting, in others only a subset or perhaps even only one. Due to the finite precision of computers, it may happen that even more solutions are approximately equivalent. Restraints permit us to indicate, in advance, the kind of solution we would like to see and so eliminate some of the extraneous solutions.

The situation is even more complicated in non-linear least squares, since in this case there is not a single computation which leads us from the observations to the unknowns. Instead, we are required to postulate trial values for the unknowns (the initial model), and then refine these. Unfortunately, many of the terms in the normal equations can only be obtained from the model, and for a poor model, they will inevitably be in error. Therefor, it should not be surprising that there will exist many occasions where a refinement will not lead to an improvement in the model, and may even lead to a deterioration. Under these conditions, the equations of restraint try to ensure that the refined model remains physically viable.

Pseudo-symmetry: One of the essential requirements for a satisfactory refinement by least squares is that there is no degeneracy in the normal equations, that is, no row is a linear combination of other rows. When this occurs, it indicates that the model has been over-parameterised, and that some parameters are simply related to other parameters. In linear least squares, singular value decomposition or eigen value (principal component) filtering will identify these redundant parameters. In non-linear least squares these techniques are much less successful, because the terms in the Design and Normal matrices are not independent of each other, but are computed from the model (see equation 10). When the symmetry of a model is reduced, for example by removing a potential centre of symmetry, new atoms have to be added to the model to replace those formerly generated by the centre. Clearly these atoms will be exactly (to machine precision) related to the ones generating them, so that the matrix will be singular, and not yield a proper solution. If, for reasons of machine precision, the matrix is not exactly singular, it will be enormously ill-conditioned, and so yield meaningless parameter shifts. Use of shift limiting or pseudo-symmetry restrains should hold the model together, so that it does not 'blow up', but even though the shifts have been contained, they are probably meaningless. Eigen value filtering should hold the structure exactly symmetrical. The reason for this is that the normal matrix, being computed from a symmetrical model, contains no information to lead to a valid less symmetrical model.

Figures 1 & 2 illustrates the minimisation function (vertical axis) as a function of a pair of symmetrically related parameters, all other parameters being held constant. When the space group of higher symmetry is used, the values of the two parameters are constrained to lie along the diagonal between their axes, and a minimum is reached. Once the symmetry is permitted to fall, the solution can move from this diagonal. In general we know nothing of the topology of this minimisation space, and we cannot know if there is a local minimum near at hand

(Figure 1), or if the parameters can slide away to non-sensical values (figure 2). The problem is almost impossible to visualise.



If there are p atoms, then the minimisation space is in $3p$ dimensions. The figures are a two dimensional slice through this space, at fixed value for all other parameters. Should any of the other parameters be altered, then we need to look at a different slice, in which the minimisation function for the parameters we are examining may be quite different. This is why the refinement of a low symmetry model derived from a higher symmetry one almost always falls into chaos. **There is no information in the first cycle of refinement to justify any parameter shifts at all!** Shifts may be generated, but these only come from rounding errors. By chance, some may be in the right direction, but their magnitude will be worthless, and their effect may be concealed by false shifts to other parameters. Many, many cycles of refinement with restraints to hold the structure together may eventually reveal a valid structure, but this procedure is only a very low-efficiency Monte Carlo method. I do not believe that there is an exact mathematical method for breaking symmetry.

Breaking symmetry: Two broad classes of procedures exist for breaking pseudo-symmetry.

1 Using external information. When such information exists, this is much the most successful method. The external information may be chemical, physical or historical. Familiarity with a class of compounds may enable the analyst to postulate deviations from symmetry, or theoretical considerations may give clues.

2 Monte Carlo methods. Small perturbations to the structure are generated, and their effect on the minimisation function is evaluated. I suspect that there is a lot of research mileage here in designing perturbation regimes, and figures of merit for assessing their value. Several workers have devised simulated annealing protocols to assist in the solution and refinement of structures (A.T. Brunger, (1988) Crystallographic Computing 4, ed N.W. Isaacs & M.R. Taylor, Oxford University Press)

04/01/96 DJW

A synthetic example: Appended to this brief article is an extract from the Lead Article 'The Control of Difficult Refinements' (D.J. Watkin (1994) 'The Control of Difficult Refinements', Acta Cryst A50, 411-437.). In this example, contrived to provide a manageable and analysable problem, a variety of techniques are examined. At the point where the centre of symmetry is removed and additional atoms are introduced, the model becomes critically unstable. In effect, three models are potentially viable - a centrosymmetric structure with either large adps or disorder, or an ordered non-centrosymmetric solution. These solutions are not mathematically distinguishable at the moment of reducing symmetry. Even a Fourier synthesis will only reveal an elogngated region of density in the region of the problematic atoms. Since the 'non-centrosymmetric' structure is in reality still centro-symmetric, so also will be phases computed from it, and thus so also the map. In this case, the canny analyst may be able to predict a valid non-centrosymmetric model, but in most real cases this will not be the case. In the event that the analyst postulates an invalid structure, this will almost certainly be recovered from any Fourier syntheses, since the centro-symmetric components of the model will dominate the phases.

Evaluating the result: If the true structure eventually turns out to be far from pseudo symmetric, then the normal equations will eventually become well conditioned, and a conventional refinement will lead to a minimisation function (or R factor etc) which clearly indicates that the new structure is better than the original. In more marginal cases, other tools may be required.

1 Molecular geometry. If geometry restraints are not being used, the molecular parameters must refine to reasonable values. If restraints are used, they must be obeyed (ie, the restraint R factor should be very low).

2 R factor. In general, the Hamilton R-factor ratio test is insensitive to this kind of problem. R-free changes may be more informative, but may be influenced by the weighting scheme, and dominated by reflections not sensitive to the changes being tested.

3 Demonstration of the stability of a structure. If a minimisation is to gain our confidence, it must be shown to be stable to small perturbations in the parameters. If a parameter is deliberately perturbed, the refinement should return to the original values. To some extent, the stability is shown by the parameter e.s.d, but these only reflect the steepness of the walls of the minimisation well, not their height nor the presence of other minima near by.

Other ideas: Perhaps people concerned with refinement strategies could learn something from the work on Direct Methods for structure solution. For small and medium size structures, these procedures have become enormously successful, yet the underlying calculations are much as they were 20 years ago. The high success rate is due to several things.

12/01/96 DJW

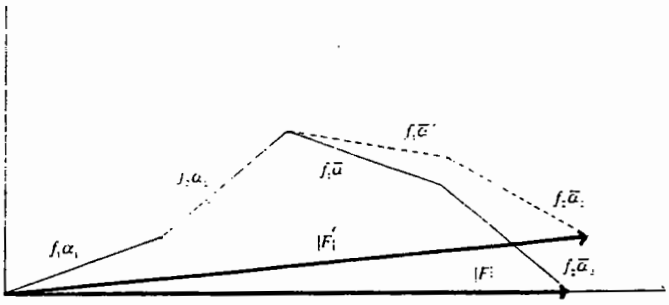
1 Not being afraid to burn computer time when there is a good probability that a computation will eventually yield a structure.

2 Optimisation of the code which is executed very many times.

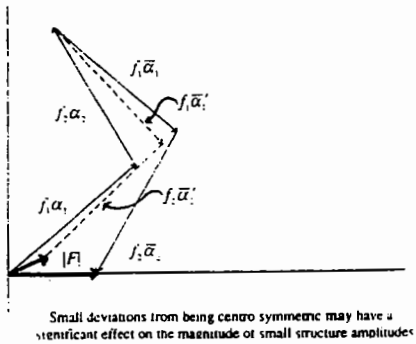
3 The design of figures of merit which enable low probability results to be weeded out efficiently.

4 Selection of subsets of the data to be used during the initial stages.

Crystallographers have traditionally completed their refinements using all the available good data, where 'good' used to imply above a certain signal to noise ratio. Dunitz and others showed that for certain kinds of problems, the key was held amongst the weak reflections (Figures 3 & 4), so that there is now a growing movement to use all data, including the negative observations (J.D. Dunitz (1979). X-ray Analysis and the Structure of Organic Compounds, Cornel University Press).



Small deviations from being centro symmetric have little effect on the magnitude of strong structure amplitudes



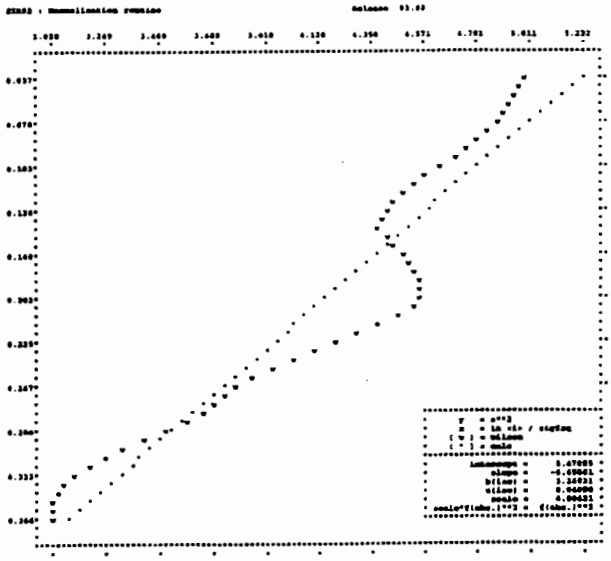
Small deviations from being centro symmetric may have a significant effect on the magnitude of small structure amplitudes

The tumbling price of area detector machines will soon make it possible for almost all crystallographers to collect as many 'unobserved' reflections as they feel they need, in order to get some pre-determined observation to parameter ratio. It doesn't take much thought to realise that uncontrolled dumping of hundreds of unobserved reflections into a refinement will not improve it. The data used need to be relevant. Rollett (J.S. Rollett, T.G. McKinlay & N.P.H. Haigh, (1976), Crystallographic Computing, ed. F.R. Ahmed, Munksgaard,) described selective use of derivatives, and Milledge (H.J. Milledge, M.J. Mendelssohn, C.M. O'Bien & Webb, G.I. (1985), Structure & Statistics in Crystallography, ed. A.J.C. Wilson, Adenine Press New York) suggested strategies for selecting reflections to measure with especial care. Atkinson (A.C. Akinson, 1985, Plots, Transformations and Regression, Oxford Science Publications) gives a wide range of techniques for assessing the impact of individual data items on an analysis. Prince (1994) shows

how to compute the leverage of reflections, but to compute this for a full data set might be difficult.

Our current interest is centred on the Wilson Plot. On of the problems with pre-screening X-ray data for bad measurements (outliers) is that unlike many other experiments, it is difficult to predict in advance the expected magnitude of individual observations, and so gauge if an observation is in gross error. The Wilson statistics enable us to make some predictions about average properties.

Figure 5 is a typical plot (from SIR92). The straight line traces the expected local average for a random arrangement of atoms. The wavy line traces the local average for the observed data. A molecular structure clearly cannot be a random collection of atoms, and the deviations from the straight line reflect recurrent motifs in the structure (inter atomic distances, relationships between residues etc). Most of the observed data will lie astride these plots. For a reflection to lie far from the local average, there must be something special about the contributions to the structure factor equation - that is, reflections lying far from the Wilson Plot are information rich.



It is immediately evident that weak reflections at high angles are nothing special, but they are very important at lower resolution since they indicate atomic arrangements which lead to cancellation in the diffracted wave front. Strong reflections at high angles are likely to be dominated by contributions from heavy atoms, or repeated aligned motifs. Princes projection matrix may be a tool for analyzing the information content of these anomalous reflections. If reflections sensitive to some structural feature (pseudo symmetry, chirality, partial isomorphous replacement) can be identified, then these reflections can be used to devise figures of merit to asses the progress of Monte Carlo methods of structure development.

12/01/96 DJW

Conclusion: Pseudo-symmetry continues to pose serious difficulties for the analyst. Symmetry breaking remains an art, and will probably use substantial computational reserves. Future advances will probably lie in devising suitable figures of merit for choosing between multiple solutions. It is likely that these will rely heavily on weak reflections, and so will require data of a very high standard. Particular care will be needed to avoid over estimation of these weak reflections when learning algorithms are used to extract structure intensities. When area detectors (image plate or CCD) are used, the lack of good energy discrimination by the detection system may lead to harmonic degradation of the data.

LEAD ARTICLE

Acta Cryst. (1994). **A50**, 411–437

The Control of Difficult Refinements

BY DAVID WATKIN

Chemical Crystallography Laboratory, 9 Parks Road, Oxford OX1 3PD, England

(*Received* 12 *September* 1989; *accepted* 17 *November* 1993)

A synthetic example

Raising the symmetry of a refinement rarely poses any serious computational problems (there are sometimes practical ones, *e.g.* origin shifts as well as parameter averaging) but lowering the symmetry is usually much more problematic. Once the symmetry has been reduced, a number of refinement strategies are available. Some of the strategies have catastrophic outcomes and so should be avoided at all costs. Others lead with differing degrees of success to acceptable solutions. The analyst is of course restrained by the programs he has available. However, most modern programs contain some features that can, more or less simply and with more or less ingenuity, lead to satisfactory refinements.

The following example was devised (Watkin, 1986) to demonstrate some features of the different procedures described above. The known structure of *trans*-1,4-dimethylcyclohexane ($P2_1/c$, half a molecule in the asymmetric unit) was remodelled into the *cis*-1,4 isomer in $P2_1$ with a whole molecule in the asymmetric unit and structure factors computed to be used as 'observations' in the subsequent analysis. We now have to pretend not to know what the structure is, to doubt the systematic absences and erroneously to take the space group as $P2_1/c$. $h0l$ reflections with l odd were eliminated.

This pseudostructure was solved with *SHELXS*86 (Sheldrick, 1985) for half a molecule in the asymmetric unit. Isotropic refinement converged at 36%, at which point the methyl group, which had a large U_{iso}, was refined anisotropically. This refinement converged at 19%. Apart from the large R factor, other symptoms of a poor refinement were the short C-methyl bond length (1.41 Å) and the very aspherical methyl temperature factor (Fig. 2).

In accordance with the above suggestions, the methyl carbon was replaced by two half-methyl-carbon atoms, one at each end of the thermal-ellipsoid long axis, and this disordered model was refined isotropically. The final

R was 11%. To complete the example, the converged isotropic model (R + 36%) was recovered and a second half-molecule generated, with all atoms isotropic, using the centre of symmetry, and the space group was reduced to $P2_1$. Several strategies were used to refine this highly symmetric starting model. Table 4 records the R factors and minimum and maximum C-C bond lengths for each refinement.

(i) *Full matrix with Choleski inversion*

Rollett remarked, 20 years ago, that some analysts were surprised that such a strategy often led to uncontrolled shifts in parameters or singular matrices (Rollett, 1970). Though the reason was described again in detail by Dunitz almost ten years later (Dunitz, 1979), the problem continues to surprise beginners.

The actual behaviour depends upon details of the least-squares program. Commonly, the matrix inversion proceeds *via* the Choleski method. If rounding errors are large, then the inversion may seem to have been successful in that it executes to completion. However, the shifts of parameters that were initially equivalent are likely to be large and are in any case valueless. The old strategy of using partial shift factors to try to contain the disruption is sometimes successful but not necessarily so. 10% of a calculated shift of 100 Å is still a big shift! [Shift factors could have a place in structure refinement in the hands of sensitive operators (Rollett, McKinlay & Haigh, 1976). These authors show that careful use of factors greater than unity can be used to accelerate a well behaved refinement. There seems to be no evidence from the literature that this strategy is in common use.] If the computation is to greater precision, smaller shifts may be computed for some parameters but eventually related parameters become pivots of the method and the latent singularities are discovered, and usually the corresponding shifts are set to zero. Thus, of a pair of originally related parameters, one is modified and the other is not. If the analyst is fortunate and the random shifts thus applied are sufficiently small and more or less in the right direction, further cycles of refinement of the now nonsymmetric structure may proceed satisfactorily. This is rarely so and the refined structures usually show all sorts of curious anomalous geometries. In fact, the refinement generally 'blows up'. In this example, the R factor rises continuously and bond lengths become worthless.

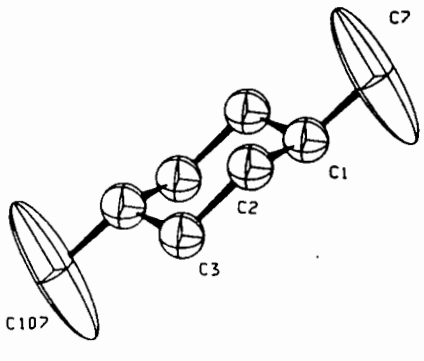


Fig. 2. Structure of pseudo-dimethylcyclohexane at a false minimum.

(ii) *Blocked matrix*

The analyst, dismayed at discovering singularities in his full-matrix refinement, either refines the related fragments each in its own matrix block or refines alternate fragments in alternate cycles. These techniques differ slightly in detail but suffer from the same problems. In the first method, the structure factors and derivatives are all calculated from the same model and the matrix blocks are accumulated at the same time. Used with care in well behaved refinements, this is probably the most cost-effective method of refining 'medium-sized' structures (the definition of 'medium' depends of course upon the size and speed of the available computer). In the second method, the model is updated after each block of atoms has been refined and so different structure factors are available for subsequent blocks. For structures not showing pseudosymmetry, this latter technique is used as the basis for cascade refinement and has been shown to be very cost effective on computers with limited memory. If there are discrete molecular fragments and the analyst is not too concerned with intermolecular distances, the method can even be used as a crude procedure for fixing origins in polar directions. However, in the current situation, both techniques suffer from the same catastrophic disadvantage.

As discussed above, the failure of the full-matrix method (normally the safest method of refinement) is caused by the latent singularities arising out of 100% correlations (except for rounding errors) between the original model and the fragment generated by symmetry. Partitioning the matrix and discarding the off-diagonal elements that relate the two fragments does not cure the problem but only blinds the mathematics to it. As a result, the refinement seems to proceed satisfactorily and no singularities are observed. In fact, in some cases the refinement may appear to be chemically satisfactory, particularly when the two fragments are whole unconnected entities.

However, in the majority of cases, the refinement is unacceptable, with the fragments showing anomalous geometries but with average values close to the accepted ones. This often becomes particularly evident when the two fragments are part of the same molecule and are joined across the former symmetry operator. In blocking the matrix, the analyst has actually thrown out the information that will eventually, once the model has settled down, ensure correct geometries. The serious danger in this procedure is that the program cannot give the user any warning that all is not well so there is real risk of

Table 4. *Course of refinements of a synthetic data set by different techniques, showing minimum and maximum* C-C *bond lengths*

Process used (see text)	*R* factor									C-C distances Minimum	Maximum
Full matrix (Choleski)	35.5	49.9	49.8	49.3	48.8	48.4	47.4	47.4	47.4	0.70	1.78
Two blocks	35.5	35.9	35.9	36.2	36.3	36.3	36.3	36.3	36.3	1.39	1.53
Full matrix (limited shifts)	35.5	35.4	35.2	33.4	26.6	14.9	4.6	1.7		1.52	1.52
Orthogonal coordinates	35.5	31.1	16.3	4.9	1.7					1.52	1.52
Antiriding constraints	35.5	34.7	15.6	3.9	1.7					1.52	1.52
Rigid-body constraints	35.5	35.2	26.2	11.9	2.9	2.1				1.52	1.53
Distance restraints	35.5	35.8	25.9	13.4	7.1	1.8				1.52	1.52
Eigenvalue filtering	35.5	35.5	35.5	35.5	35.5	35.5	35.5	35.5	35.5	1.41	1.52
Common sense	9.1	1.7								1.52	1.52

improperly refined structures being published, with little readers can do to recognize the situation. In the synthetic example, minor anomalies appeared in bond lengths but the structure remained approximately pseudosymmetric. It is only the high *R* factor, still over 30%, which makes us suspicious. If the pseudosymmetric structure had refined to, say, 12%, we might have accepted the model and assumed that there was something wrong with the data. *Blocking the matrix can never be recommended as a cure for singularities unless their source is well understood.*

(iii) *Full matrix with shift-limiting restraints*

In the case we are concerned with here, in which the original model has higher symmetry than the 'true' structure, the small (but otherwise uncontrolled) shifts permitted by shift-limiting restraints may mean that eventually the model drifts towards a correct one. The matrix then begins to contain terms computed from more or less correct derivatives, the shift-limiting restraints are over-ridden and the refinement proceeds to an acceptable solution. In this case, convergence occurs after eight cycles, at $R = 2\%$, with a model very close to that used to produce the pseudo-observations (Fig. 3).

The data *do* contain enough evidence for the original structure to be recovered and it is merely the unsatisfactory nature of the normal matrix (because of the over-symmetric model) that prevents proper refinement. However, this process, though semi-automatic, may be rather slow to start converging and, since it depends on fortuitous random shifts, cannot really be recommended *except as a method of last resort*. As outlined below, there are generally better methods. If the program being used permits shift-limiting restraints to be expressed explicitly in the same units as the parameter concerned, restrictions of about 0.1 Å seem to be workable values.

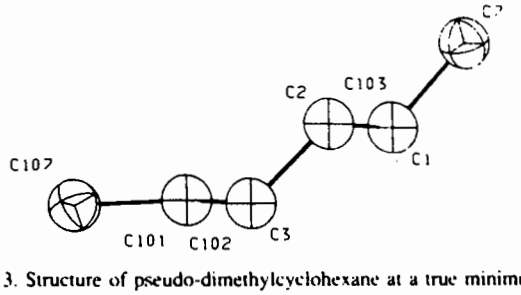


Fig. 3. Structure of pseudo-dimethylcyclohexane at a true minimum.

(iv) *Reparameterization to orthogonal coordinates*

Pairs of new coordinates are defined for refinement by least squares, one being the sum and the other the difference of the corresponding positional parameters of the symmetrically related atoms. The matrix for a structure that is exactly pseudosymmetric is still singular but, once the structure is perturbed, the refinement quickly settles down. Shift-limiting restraints are still necessary for the initial computations but they can be fairly slack and limiting the shift to be not more than one unit cell seems to work well. This means that, once the crystallographic derivatives become meaningful, they are not unduly damped by the restraint. Convergence is achieved in five cycles.

(v) *Anti-riding constraints*

When a structure is refined in a high-symmetry space group, the analyst may be applying all sorts of constraints to the solution. These are implicit in the space-group symmetry operators, which the model for the total contents of the cell must obey. If the analyst believes that the bulk of his structure conforms (at the resolution of his data) to the higher symmetry and only some atoms are in more general positions, he is at liberty to refine the structure in the lower-symmetry space group and then re-impose selectively the relationships previously defined by the space-group operators. These relationships can be set up as constraints. In this case, we might believe that atoms C(1) to C(3) are very close to being centrosymmetrically related to atoms C(101) to C(103) and we could impose this belief as anti-riding constraints by setting the shift in $C(1, x)$ to be of the same size but opposite sign as that in $C(101, x)$ and so on for all the pairs of positional parameters of the atoms in the central ring. Thus, only one least-squares parameter is computed for the shift in $C(1, x)$ and $C(101, x)$,

Table 5. *Eigenvalues and selected eigenvectors of the normal matrix for the centrosymmetric starting model*

(*a*) The 12 largest eigenvalues. The remaining 12 have values close to zero

2.68	2.53	1.37	2.45	1.52	1.57	1.72	2.15	2.01	1.93	1.97	2.02

(*b*) Components of the eigenvector corresponding to the first eigenvalue. Parameters are ordered *x*, *y*, *z* for each atom. Entries in the second row are centrosymmetrically related to the corresponding entries in the first row

0.32	−0.02	0.24	0.17	0.02	0.16	0.38	0.01	0.35	−0.10	0.03	−0.03
−0.32	0.02	−0.24	−0.17	−0.02	−0.16	−0.38	−0.01	−0.35	0.10	−0.03	0.03

(*c*) Components of the eigenvector corresponding to the 13th eigenvalue

0.20	0.63	0.04	0.15	0.03	0.02	0.04	−0.06	−0.06	0.14	0.11	−0.03
0.20	0.63	0.04	0.15	0.03	0.02	0.04	−0.06	−0.06	0.14	0.11	−0.03

another for C(1, *y*) and C(101, *y*) and so on for all the other pairs of parameters.

Because these are constraints, only three least-squares parameters are refined for each atom pair and after the matrix work appropriate shifts are applied to the related atoms. Formally, this is the same as working in the higher-symmetry space group, with the exception that the structure factors and derivatives have to be computed for all six atoms. The other atoms, the methyl C atoms, which through their temperature factors or anomalous bond lengths made us suspect lower symmetry, will of course be refined without this sort of constraint but may be the subject of reparameterization or shift-limiting restraints.

Once the refinement shows signs of stabilizing, the anti-riding constraints can be removed and, with mild shift-limiting restraints, full-matrix refinement can be used to finish off the task in a total of five cycles.

(vi) *Rigid-body constraints*

The centrosymmetric refinement yielded a core structure, the cyclohexyl ring, that made chemical sense. Another way to proceed to the lower-symmetry structure would be to refine the cyclohexyl ring as a rigid body with its current geometry and only the methyl groups as independent atoms. Replacing the 18 degrees of freedom of the core by only six rigid-body parameters reduces the number of ways in which the refinement can fall into ruin and ensures that the solution makes some chemical sense. As with the constrained refinement above, the rigid-body constraint should eventually be relaxed. Convergence was achieved in six cycles.

(vii) *Distance restraints*

The major problems with the centrosymmetric refinement were the anomalous temperature factor of the methyl group and its bond length from the ring C atom. The full-matrix refinement revealed its failure by both the *R* factor rising and the quite unacceptable C–C bond lengths. This suggests that another approach to a stable refinement might be to use bond-length restraints, both for the C-methyl bond and also for the bonds in the ring. In this example, we can make a well informed guess at suitable values. In more general cases, theoretical arguments may not provide actual bond lengths but may indicate that, by symmetry, bonds should have similar lengths. This similarity may be applied as a restraint. Convergence was achieved in six cycles.

(viii) *Eigenvalue filtering*

This method provides excellent diagnostics as to why the full-matrix refinement failed. Table 5 lists the eigenvalues of the scaled normal equations and the eigenvectors corresponding to eigenvalues 1 and 13. If one remembers that the first four atoms are centrosymmetrically related to the second four, it is instructive to note that the signs of the second 12 components of eigenvector 1 are the opposite of the first 12, while they are the same for eigenvector 13. This reveals straight away that, while the sums of corresponding parameters are well defined, the differences are not, and explains why the standard refinement is unstable. It also explains why the re-parameterization described above is a useful technique. In that case, a rotation was applied in which the components of the relevant eigenvectors were exactly zero or $2^{1/2}/2$.

Fig. 4 (*ad hoc* plotting program) represents the variation of the minimization function, $M = \sum(w\Delta^2)$, as a function of the value of C(1, *x*) and C(101, *x*). (Note that this is a two-dimensional section through a 24-dimensional space. Changing any other parameter in the model requires us to look at the section parallel to the given section but displaced in the direction of the perturbed parameter.) The dotted line lies in the plane $C(1, x) = -C(101, x)$, so that the minimum for the centrosymmetric structure must lie on this line. In this case, this minimum is also the local absolute minimum in the C(1, *x*)C(101, *x*) plane and lies at the bottom of a shallow bowl. Movement away from the dotted line causes the minimization function to rise, so that a centrosymmetric solution for these two parameters is best, even when noncentrosymmetric positions are available.

Fig. 5 is the corresponding contour map and shows that the minimum is well defined in the C(1, x) + C(101, x) = 0 direction (centrosymmetric) but not at right angles. Small perturbations along this direction will not affect the minimization function greatly and so are more or less equally acceptable.

Fig. 6 is a similar representation for C(7, x) and C(107, x). Again, the dotted line contains the minimum for the centrosymmetric structure. However, this is at a saddle point if the two coordinates are not required to vary synchronously; lower minima lie to either side of the line. The gradient of the surface perpendicular to the symmetry line should be zero for points immediately adjacent to the line, so there is no information in the normal matrix to tell the calculation to move the parameters off one way or the other. In the presence of rounding errors, a small gradient may be seen, indicating some distant minima, and large spurious shifts are computed. Eigenvalue filtering eliminates these spurious shifts. It differs from Choleski inversion, which can also trap large shifts, in that it recognizes special relationships between parameters and preserves these relationships.

In this example, the normal matrix contains no information at all about what shifts should be applied and the structure remains essentially unchanged after ten cycles of refinement. This is mathematically correct, though naturally disappointing for the hard-pressed analyst hoping for miracles. The standard Taylor expansion of the structure-factor equation and subsequent building of the normal matrix involves only the first-order derivatives. This helps to increase the range of convergence of the method, avoids saddle points and saves the expense of computing second derivatives, at the cost of a possibly nonquadratic convergence. However, we believe that inclusion of second derivatives would give eigenvalue filtering the information it needs to determine which parameters need to be perturbed and the correlation between these perturbations.

Once the model has been perturbed, the minimization surface (which is computed from the model) ceases to be symmetric and, if the perturbations are in the correct directions, the true minimum appears in the surface and the refinement proceeds correctly. We have not seen an example of the use of second derivatives in structure refinement.

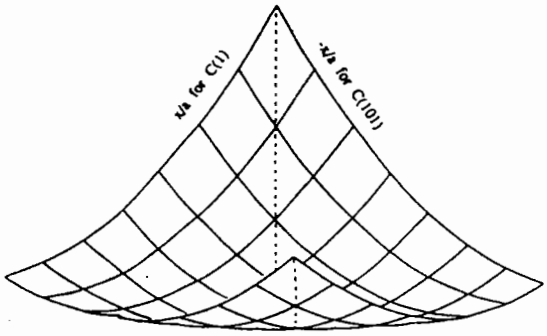


Fig. 4. Representation of a section of the minimization function for well resolved parameters in pseudo-dimethylcyclohexane.

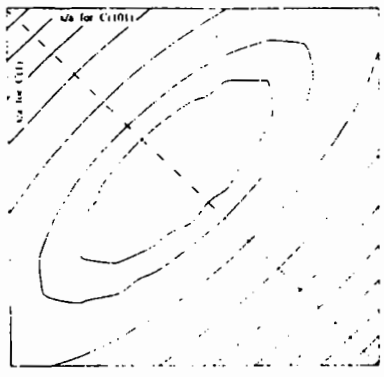


Fig. 5. Contour map corresponding to Fig. 4.

(ix) *Common sense*

The split-atom refinement (R = 11%) could have given us a clue about a possible model for the noncentrosymmetric space group. As with the ordered model, a second half-molecule could be generated using the pseudocentre, giving four half-methyl-carbon atoms (two at each end of the molecule). Taking a nonequivalent one from each pair and restoring it to full occupancy gives a model with asymmetric methyl groups. The R factor for this structure produced by trivial modelling techniques has a value of 9% and refines by any valid method in two cycles. The game, therefore, in all cases of near pseudosymmetry, is to provide the mathematics with as much evidence as possible drawn from sources external to the X-ray experiment and bearing upon noncontroversial issues in the analysis, and so permit concentration of the information contained in the X-ray data onto the real issues under investigation. The risk, of course, is of feeding in erroneous or prejudiced opinions.

Likelihood-weighted real space restraints for refinement at low resolution

J.P. Abrahams,
MRC Laboratory of Molecular Biology,
Hills Road, Cambridge, UK

INTRODUCTION

A correct electron density map agrees with all measured diffraction data and all prior knowledge.

The fundamental problem of crystallography, the inability to measure phases of structure factors, requires the introduction of additional knowledge to arrive at a solution. Many types of such information are currently incorporated in programs used for the refinement of protein crystal structures and we can say that this refinement currently hinges on knowledge of the following constraints and/or restraints:

- atomicity and positivity, atomic shapes defined as form factors
- topology, the chemical and electro-chemical connectivity of the structure
- stereochemistry, defined in terms of bond lengths, bond angles and chirality, steric, electrostatic and hydrophobic interactions as defined by torsion angles and force fields, planarity and temperature factor restraints
- non-crystallographic symmetry.

The apparent gap between the first three structurally local chemical restraints and the structurally global assumption of the preservation of macromolecular shape (ie. non-crystallographic symmetry), raises the question why current programs do not employ sub-global or supra-local restraints. Some sub-global restraints are in fact present in all current refinement programs, albeit in a hidden form. For example, most programs will conserve the folding pattern of a molecule because the sub-global restraint of the conservation of fold is inherent to the minimising algorithms. This paper discusses another type of supra-local restraint which can be introduced with relative ease in current refinement programs and has a wide range of applications.

The more you can measure, the less you need to know.

If data to a sufficiently high resolution are available, atomicity and positivity restraints are usually adequate to solve a structure. Because these essentially one-dimensional restraints are comparatively easy to formulate, the process of

structure determination has now been automated to a large extent for such well-ordered crystals, providing the structure is not too large. On the other hand, if the resolution of the data does not allow the separation of individual atoms, the solution of the structure requires more knowledge in the form of heavy-atom derivative data, anomalous data, or a molecular replacement model. Successful refinement of these solutions depends on the proper use of restraints on connectivity and stereochemistry; in their absence the refinement will not converge to the correct solution. Even when data to true atomic resolution are available, the need for such restraints quite often still exists, because not all of the molecule needs to be equally well ordered for a crystal to diffract to such a high resolution.

When the resolution of the diffraction data is lower still, refinement programs are prone to diverge from the proper solution even when all conventional stereochemical restraints are being used. It has been argued that this happens around a resolution of about 2.5 Å in the absence of non-crystallographic symmetry, because at a lower resolution the number of parameters of the model (the three coordinates and the temperature factor of each of the ordered non-hydrogen atoms) is greater than the number of observed reflections. It should be realised that this is a simplification. Refinement is not an exhaustive search for a global minimum in an N-dimensional space (where N is the number of parameters), but a minimisation of an N-dimensional function. The difference between these two procedures is dramatic: inclusion of additional parameters in the former case increases the difficulty of the problem exponentially, in the latter case additively. Other considerations are that atomic coordinates and temperature factors are not independent parameters because of the introduction of restraints and that not all diffraction data are equally restrictive, since weak data can, and low resolution data will contribute less information to the atomic detail of a structure. If a refinement diverges from the true solution (as measured by the free R-factor, for example), this is a reflection of the inadequacy of the model and its restraints, given correct data. In the end, it is the sequence of a protein which determines the structure, not the diffraction pattern.

The main reason for the breakdown of refinement algorithms at lower resolution is that none of the currently employed procedures is capable of describing poorly-ordered parts of the crystal with sufficiently few degrees of freedom. The tendency exists to model these poorly ordered parts by single conformations with sub-optimal stereochemistry, whilst in reality these parts are present in multiple conformations, almost certainly with good stereochemistry. This tendency is illustrated most clearly by the correlation between the stereochemical quality of a model and the resolution of the data. The program "PROCHECK", for example, is more tolerant of poor stereochemistry when the resolution of the data is low.

Since it not so much the number of parameters which renders refinement at low resolution a perilous task, but rather the degree of freedom each of these parameters has independent of the other ones, one way forward is to introduce

additional restraints, thus coupling the parameters more strongly. These restraints provide the boundaries within which the refinement routine is allowed to search for the solution and the tighter the boundaries, the smaller the chances of the routine losing its way. If this coupling of parameters works, refinement of crystals diffracting to a higher resolution might benefit too, if parts of these crystals are poorly ordered or are present in multiple conformations.

In absence of proof to the contrary, assume things stay the same.

As pointed out above, non-crystallographic symmetry (NCS) is one of the sources of additional restraints. If one is confident that NCS is exact, the weight of these restraints can be increased to infinity, turning the restraints into constraints. On the other hand, if one has evidence for the breakdown of NCS, the weight of the associated restraints should be decreased, thereby unfortunately compromising the refinement process. Determining the appropriate weight is complicated by the fact that the strictness of the NCS will vary locally: although the structures of two molecules can be identical in their cores, they might differ at lattice contacts.

These considerations prompted the development of a procedure which automatically determines a local weight of the NCS restraint for each of the atoms, rather than a global one. If the data suggest that the NCS breaks down locally, the associated restraints are weighted down too. The described procedure was implemented in the "NCS"-module of the TNT-suite.

After determining the NCS operator, the vector shift of each atom is determined by comparing the individual NCS related structures to the averaged structure. In conventional averaging, this vector is combined with the shift induced by the X-ray terms and the shift suggested by stereochemical and energetic terms, after globally weighting each of the vectors. In locally weighted averaging, the shift suggested by the NCS restraint is weighted first by the statistical significance of the differences between the positions of the individual NCS related atoms and their averaged position. The likelihood that an atom should in fact be at the averaged position can be calculated from its distance to that position and from prior knowledge of the coordinate error:

$$P = 2 / \sqrt{(2\pi)} \int_{d}^{\infty} e^{-z^2 / 2} \, \delta z$$

where:

P = likelihood, assuming Gaussian statistics

d = distance of the atom under consideration from the averaged position

$z = d / \sigma$

$\sigma = \sqrt{2}$ times the positional standard deviation

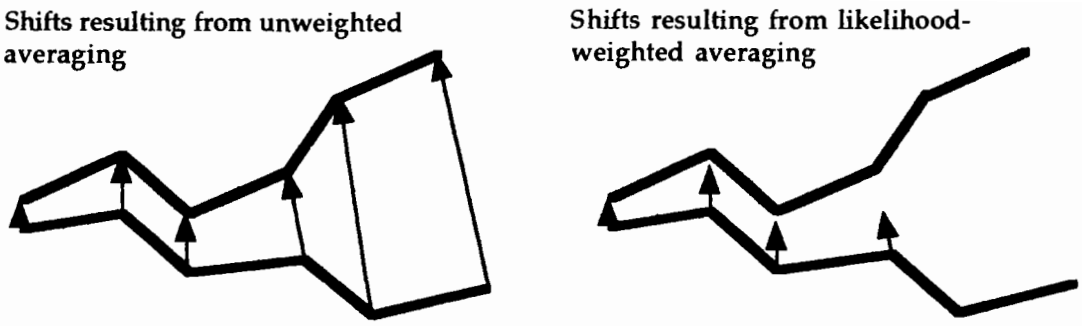


Fig. 1 *Unweighted vector shifts vs. likelihood-weighted vector shifts.*
Non-crystallographic symmetry restraints introduce vector shifts towards the averaged atomic positions as indicated in the figure. If the NCS restraint is unweighted, the length of the vector shift is proportional to the distance between the actual atomic position and the averaged one. However, if the NCS restraint is weighted by likelihood, the length of the shift vector is proportional to the chance that the atom really does occupy the averaged position, and therefore gets shorter the larger the distance between the actual and averaged atomic positions is.

It follows from the formula that the larger the distance between an atom and the averaged position, the smaller the weighted NCS shift towards this position becomes, because of the increased likelihood that the atom really does occupy a position different from the average. The difference between likelihood-weighted and unweighted averaging is illustrated in figure 1.

Various simplifications were made in the implementation because of my unfamiliarity with "FLEX", the computer language in which the TNT-suite is written, and because of the desire to test the merits of likelihood-weighted averaging quickly without getting bogged down in elaborate coding. Neither the determination of the NCS-operators, nor the determination of the averaged position towards which the atoms gravitate is weighted by likelihood, for example. The standard deviations of the atomic positions can be calculated from their temperature factors, but this finesse was omitted, and an overall standard deviation was assumed instead. However, these shortcuts did not prevent the procedure from proving its worth.

RESULTS

The procedure as outlined above initially was devised for the refinement of an inhibited form of F_1 ATPase (373 kD). The stoichiometry of this enzyme is $\alpha 3\beta 3\gamma\delta\varepsilon$, and the α and β subunits do not obey NCS because each of the three β subunits is in a different catalytic state. Data were collected at the SRS in Daresbury to a resolution of 3 Å, but unfortunately were only usable to 3.1 Å because the limiting aperture of the MAR Research image plate scanner used for data collection was not designed for a crystal-to-detector distance larger than 450 mm. The data are summarised in table 1.

The structure of native F_1 ATPase, which was refined to 2.8 Å (R_f: 17.2%, R_{free}: 25.4%, 603 waters, 0.016 Å rms bonds, 2.91° rms angles), was used as a replacement model for the inhibited form of the enzyme. After rigid body refinement using TNT (the native structure was broken up into domains and secondary structure elements), the free R-factor fell from 37.8% to 23.2%. A difference map allowed localisation of the inhibitor, a peptide with 16 non-canonical aminoacids. The presence of the inhibitor locally distorts the structure of F_1 ATPase and the domains of the enzyme had adopted slightly different orientations relative to one another, perhaps in response to freezing the crystal. These structural variations prompted some rebuilding of the model.

The model of the native structure contains 603 water molecules which in the rigid body refinement were treated as a single rigid body. If the water molecules are omitted, the free R-factor increases by 0.6%, so the presence of the water molecules is legitimate. On the other hand, because the ordered water structure was treated as a single rigid body, and because the individual domains of F_1 had shifted relative to one another, it was unlikely to be entirely correct.

All further attempts to refine the rebuilt structure using the conventional refinement techniques of TNT failed, as judged by monitoring the free R-factor. Various weights on the stereochemical restraints were tried, and the model could quite easily be refined to one with superb stereochemistry, unfortunately always accompanied by an increase of the free R-factor. Also, refining only part of the model did not improve its quality. Neither keeping the waters constant, keeping the temperature factors constant, nor keeping the poorly ordered parts of

Table 1: *Summary of the crystallographic data as a function of the resolution.*

	15Å to 5.5Å	5.5Å to 4.6Å	4.6Å to 4.0Å	4.0Å to 3.6Å	3.6Å to 3.35Å	3.35Å to 3.1Å	Overall
R_{sym}[1]	0.058	0.072	0.075	0.087	0.111	0.152	0.079
$\langle F/\sigma \rangle$[2]	70	48	43	33	24	16	37
mult.[3]	2.6	2.7	2.7	2.5	2.5	2.5	2.6
compl.[4]	0.97	0.96	0.90	0.87	0.85	0.83	0.86
R_f[5]		0.165	0.165	0.190	0.220	0.245	0.179
R_{free}[6]		0.205	0.190	0.235	0.295	0.275	0.225

[1] $R_{sym} = \Sigma\ |(I-\langle I \rangle)|\ /\ \Sigma(I)$, where I = observed density, and $\langle I \rangle$ is the mean density from multiple measurements after rejections (0.0015% of the data were rejected).

[2] The mean of the structure factor amplitude over the standard deviation as estimated from counting statistics (after modification based on the observed agreement between symmetry mates).

[3] Multiplicity of the data.

[4] Completeness of the data.

[5] The crystallographic R-factor.

[6] The free R-factor of 1% of the data (877 reflections) not included in the refinement.

the model constant (or combinations of these constraints) proved beneficial. The model could only be refined further upon the introduction of the likelihood-weighted local NCS restraints described above.

The scope for such an NCS-restrained refinement was quite large: not only were there three copies of each of the α and β subunits, but also the native model, refined with more complete data extending to a higher resolution, was available. Restraining the refinement to the native structure could quite easily be implemented as non-crystallographic averaging between crystal forms, though in this case the native model was kept constant. It was already quite clear that the individual subunits were not identical and the maps also indicated differences between the native and the inhibited form of the enzyme, yet all differences seemed to be mainly local. However, some domains did adopt slightly different orientations and some secondary structure elements had shifted slightly upon inhibition and/or freezing of the crystal. Therefore the structure was broken up into stretches of 4 aminoacids, and the NCS-operators were determined separately for each of these stretches. An overall standard deviation of the atomic positions of 0.3 Å was assumed, which is close to the expected mean standard deviation at 3 Å resolution.

Because of the implementation it was not practical simultaneously to restrain the refinement to the native structure and to the model itself through the averaging of the α and β subunits. The rebuilt model was therefore first subjected to 7 cycles of refinement restrained to the native structure, keeping the temperature factors and the water molecules constant. This improved the free R-factor from 23.1% to 22.9% and also improved the stereochemistry of the model. Subsequently the model was subjected to 3 cycles of internal likelihood-weighted averaging, which improved the free R-factor to 22.8%, and again the stereochemistry benefited. The free R-factor was further reduced to 22.7% by 4 additional cycles of refinement restrained to the native structure.

In order to further test the benefits of locally weighting the NCS restraints by likelihood, unweighted averaging was also tried on the resulting model, which was very close to the final model. As was the case for the likelihood-weighted averaging, the model was broken up into fragments of 4 aminoacids before determining the NCS operators. Several global weights on the NCS restraints were tried, but all resulted in an increase of the free R-factor by at least 1.5%. If the weight was set high, the stereochemistry suffered, and both the refined R-factor and the free R-factor increased. If the weight was set low, the quality of the stereochemistry was unaffected, the refined R-factor improved slightly, but the free R-factor still increased, indicating that the refinement was drifting away on model bias.

There was still the problem of refining the temperature factors of all the atoms and the positions of the water molecules. Neither could be altered without increasing the free R-factor using conventional refinement techniques, indicating the inadequacy of the restraints imposed on these parameters. However it

was possible to refine these parameters in real space. Refining the positions and temperature factors of all atoms present in the model, including the water molecules, improved the already excellent geometry of the model substantially to 0.006 Å rms deviation of bond lengths, 1.75° rms deviation of bond angles, the few remaining bad contacts involving water molecules disappeared, the R-factor dropped to 17.9% and the free R-factor dropped to 22.5%. The ϕ-ψ angles of 86.8% of the residues are in the most favoured regions of the Ramachandran plot, 12.8% of the residues are in additionally allowed regions, and 0.4% of the residues are in generously allowed regions. The map used for the two cycles of real space refinement was calculated using the observed structure factor amplitudes and the phases of the penultimate model. In a real space refinement the phases remain unchanged, thereby restraining the model to the map and thus allowing the otherwise unrestrainable temperature factors and the parameters of the water molecules to be refined.

DISCUSSION

The tests on F_1 ATPase suggest that even the rather crude implementation of likelihood-weighted NCS restraints described here provides a useful extension to the currently employed set of restraints. Refinement benefits because the procedure provides an automatic discrimination between regions where the NCS is properly maintained and those regions where the similarity between comparable structures is reduced. The procedure is independent of the actual refinement algorithm employed, and could well be used in conjunction with the techniques of maximum likelihood refinement as described in other papers in this volume, and with simulated annealing techniques. In particular the treatment of local symmetry using 1:4 distance restraints described in another paper in this volume, will quite probably enhance the usefulness of likelihood-weighted symmetry restraints.

It might be worthwhile to try to extend the usefulness of the procedure as described here to cases where there is neither (partial) NCS, nor another related structure for providing the restraints. In such cases, databases of unrelated high resolution structures could provide coordinates of peptide fragments to which parts of the structure under investigation could be restrained. If this approach proved to be useful, the procedure might be extended even further by allowing multiple conformations to describe poorly ordered regions, each of the conformations tightly restrained against fragments of known high resolution structures. Also, in this case the increase of the number of parameters should be offset against the tighter restraints.

ACKNOWLEDGEMENTS

I am greatly indebted to Alison Sutton and Andrew Leslie for carefully and critically reading the manuscript.

Weighting Diffraction Data

G. David Smith
Hauptman-Woodward Medical Research Institute, Inc.
73 High Street, Buffalo, NY 14203
and
Roswell Park Cancer Institute
Elm and Carlton Streets, Buffalo, NY, 14263

In small molecule crystallography, structures are typically refined by full-matrix least-squares using atomic resolution data for which there may be as many as ten observations per parameter. The results from such refinements typically allow hydrogen atoms to be located and refined, and it is not unusual to obtain R-factors of less that 0.05, minimizing $\Sigma w\Delta^2 = \Sigma w(F_o-F_c)^2$. Estimates of the various sources of error for each amplitude are propagated into $\sigma(F_o)$ and the weight applied to each Δ is inversely proportional to $\sigma^2(F_o)$. The underlying assumption for the validity of applying $w = 1/\sigma^2(F_o)$ is that all atoms within the unit cell which contribute to the scattering have been appropriately modeled and that errors in F_o are reflected in the $\sigma(F_o)$. Under these circumstances, average values of the goodness of fit (GOF = $[(\Sigma w\Delta^2)/n]^{1/2}$) and $|\Delta|$ in equal volume shells of $\sin\theta/\lambda$, are evenly distributed. While the goodness of fit should be equal to unity, this is seldom observed as the standard deviations in even the most careful experiments are typically underestimated; as a result, the goodness of fit is typically found to range between 1.5 and 2.0. A more sensitive way in which to examine the agreement of the entire set of data is through the use of a δ(R) plot (Abrahams & Keve, 1971; Howell & Smith, 1992). In this technique, the ranked δ(real) [δ(real) = $(F_o - F_c) / \sigma(F_o)$] are plotted against δ(expected), where the latter is calculated on the basis of a normal distribution of errors. Assuming that the amplitudes (F_o) do not contain a systematic error and that the $\sigma(F_o)$ have been correctly estimated, then the δ(R) plot should be linear with a slope of unity and an intercept of zero. In practice, the slope is found to range between 1.5 and 2.0 and is comparable to the goodness of fit. While deviations from linearity may be due in part to errors in the observed amplitudes or in the model, a δ(R) plot does allow one to assess the validity of the weights or weighting scheme.

In macromolecular crystallography, the situation is considerably different. Atomic resolution data are rarely available, there may be fewer than three observations per parameter, contributions from hydrogen atoms are seldom included, and there are only a few examples of structures for which the entire contents of the unit cell have been modeled. The first two problems can be

overcome through the use of a restrained refinement (Hendrickson & Konnert, 1980; Finzel, 1987; Sheldrick, 1993) or a simulated annealing procedure (Brünger, Kuriyan & Karplus, 1987). However the inability to model the entire contents of the unit cell means that differences between F_o and F_c, particularly in the lower resolution shells, are due not only to errors in the structure, but also to an incomplete model. The use of an 8 or 10Å lower resolution cutoff compensates in part for the incomplete model, but is not sufficient to eliminate all bulk solvent effects. Under these circumstances, what is an appropriate weighting scheme to apply in macromolecular refinements?

An examination of a plot of $\langle|\Delta F|\rangle$ against $\sin\theta/\lambda$ in equal volume shells for a well refined structure at reasonably high resolution provides examples of less than optimal weighting schemes and also suggests alternate schemes which may be more appropriate. As seen in Figure 1, the magnitude of $|\Delta F|$ decreases significantly as $\sin\theta/\lambda$ increases. This behavior is due primarily to our inability to model the bulk solvent and the fact that the largest amplitudes usually are found in the lowest resolution shells. Thus, any weighting scheme which applies equal weight across the entire resolution range will strongly bias the refinement towards the low resolution data. In this example, an average low resolution reflection will contribute 40 times more to the function minimized than a high resolution reflection. The use of experimental weights presents a similar problem as the ratio of high to low resolution $\sigma(F_o)$'s is smaller than the ratio of high to low resolution amplitudes, resulting in a similar bias towards the low resolution data. Considerable effort is expended in the growth of high quality crystals which diffract to high resolution and in the measurement and processing of data. However, there seems little point to acquire this data and then minimize its effect upon the refined structure by using unit or experimental weights.

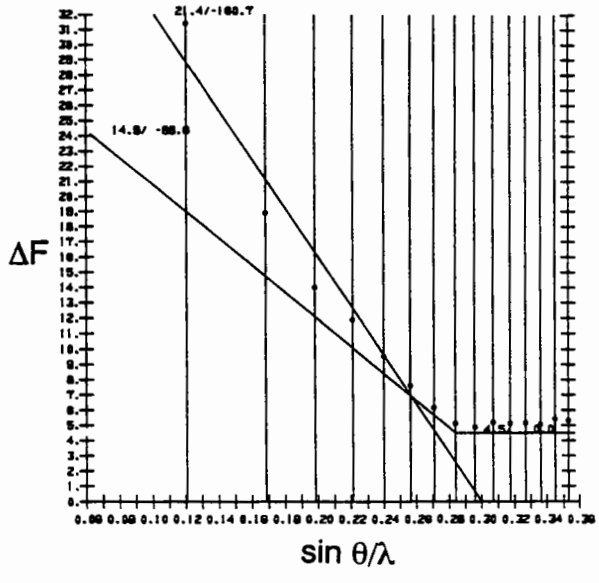


Figure 1. Plot of $|\Delta F|$ versus $\sin\theta/\lambda$ for the insulin data between 8 and 1.4Å resolution. The single straight line is the least-squares line through data between 8 and 1.76 Å; the two connected lines describe the Sigma(appl) from which weights were calculated.

The plot of $\langle|\Delta F|\rangle$ against $\sin\theta/\lambda$ does suggest an empirical weighting scheme which would be appropriate. In the original version of PROLSQ (Hendrickson & Konnert, 1980), an optional weighting scheme as a function of $\sin\theta/\lambda$ was provided [Sigma(appl) = A + B($\sin\theta/\lambda$ - 1/6)]. If we consider the $\langle|\Delta F|\rangle$'s to be equivalent to Sigma(appl), then a proper choice of the constants A and B will describe a straight line which is proportional to that obtained from the plot of $\langle|\Delta F|\rangle$ against $\sin\theta/\lambda$. For higher resolution cases, a single straight line is not adequate to model the entire resolution range, and for these cases a two-line empirical weighting scheme has been devised.

In order to assess the effect of various weighting schemes upon the refinement, a series of restrained (PROFFT and SHELXL-93) and simulated annealing (XPLOR) refinements, shown in Table I, were performed on a structure of a complex of insulin with a phenolic derivative. A total of 159339 data from crystals grown in microgravity were scaled and merged to yield 18076 independent data with an $R_{merge}(F^2)$ of 0.066 and are 99% complete (77% for $F \geq 2\sigma(F_o)$) to a resolution of 1.4Å. The starting model in all cases was a fully refined dimer (R = 0.153) consisting of 100 of a possible 102 residues (807 atoms) and 139 water molecules. Overall results are summarized in Table I and plots of the goodness of fit versus $\sin\theta/\lambda$ are illustrated in Figure 2. With the exception of XPLOR for which the R-factor and goodness of fit are somewhat larger, the overall statistics would suggest that there is little difference in the results of the refinement using the different weighting schemes in the restrained refinements. However, examination of Figure 2 clearly shows that several of the weighting schemes produce an uneven distribution of the goodness of fit, and hence $\Sigma w\Delta^2$, as a function of resolution.

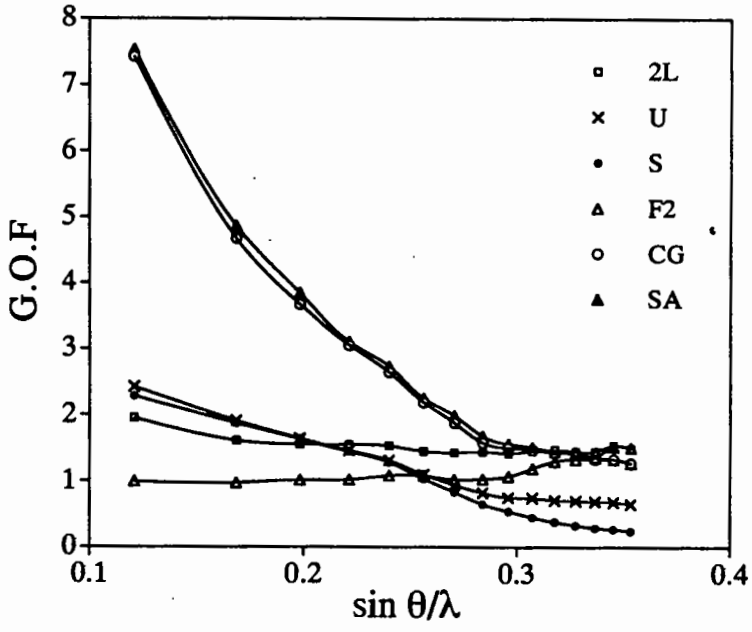


Figure 2. Plot of the goodness of fit versus $\sin\theta/\lambda$ for the six refinements on the insulin data. Codes are given in Table I.

Table I

Refinements and Overall Statistics

Code	Method	Residual	GOF
U	Unit Weights PROFFT Refinement $w = 1.0$	0.153	1.332
S	Experimental Sigma PROFFT Refinement $w = 1 / \sigma^2(F_o)$	0.159	1.235
2L	Two-Line Weighting Scheme PROFFT Refinement $\sin\theta/\lambda \leq 0.284$: $w = 1 / [A + B (\sin\theta/\lambda - 1/6)]^2$ $\sin\theta/\lambda > 0.284$: $w = 1 / [C + D (\sin\theta/\lambda - 1/6)]^2$	0.154	1.538
F2	Default Weights* SHELXL-93 Refinement (F^2) $w = 1.0 / [\sigma^2(F_o^2) + (aP)^2 + bP]$ where $P = (F_o^2 + 2F_c^2) / 3$	0.158 (0.273)	1.365 (1.122)
CG	Default Weights XPLOR Conjugate Gradient Refinement $w = 1.0$	0.177	3.285
SA	Default Weights XPLOR Simulated Annealing Refinement $w = 1.0$	0.183	3.371

* Statistics in parentheses are compiled on the basis of F^2.

The larger values of the goodness of fit at low resolution show that these data dominate the refinement, particularly for the XPLOR refinement. The smallest variation in the goodness of fit is seen for the two-line empirical weighting scheme as well as for the default weighting scheme employed in SHELXL-93. Additional information is provided by the $\delta(R)$ plots for each of the refinements, shown in Figure 3. The linearity of the empirical two-line and the default weighting scheme used in SHELXL shows that these weighting schemes quite adequately reflect the expected normal distribution of errors. In contrast, the sigmoidal shape of the other four $\delta(R)$ plots strongly suggest that the weighting scheme is inappropriate.

Currently XPLOR may be the most widely used program for refinement of macromolecular structures, but the majority of crystallographers are using the default weighting scheme of unity. As noted earlier, this tends to minimize the contribution of the higher resolution data to the refinement. However, it is a relatively simple matter to use either a one- or two-line weighting scheme as

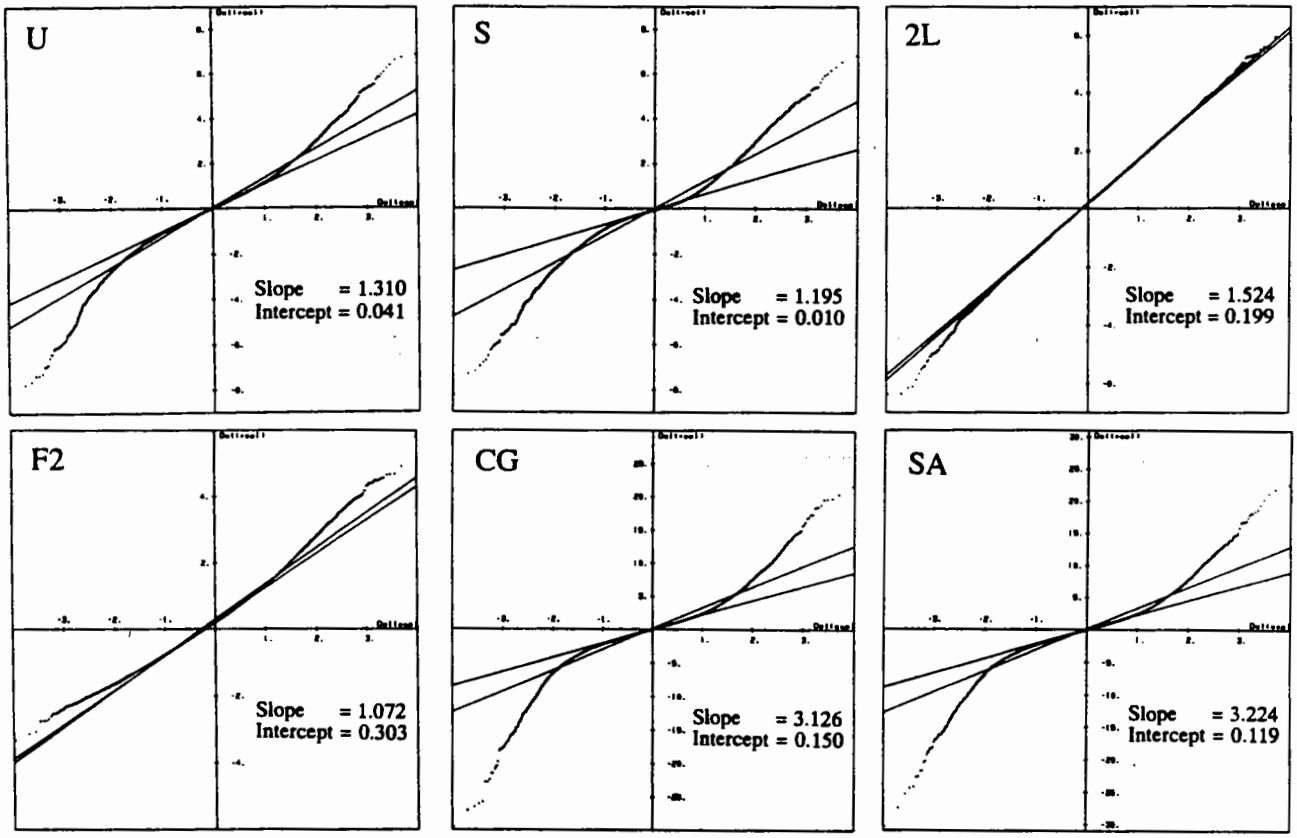


Figure 3. $\delta(R)$ normal probability plots for the six refinements on the insulin data. Codes for the refinements are given in Table I.

described above. On page 172, Section 12.5.4 of the XPLOR manual for Version 3.1 (Brünger, 1992), a procedure is described to employ the one-line empirical weighting scheme [Sigma(appl) = A + B($\sin\theta/\lambda$ - 1/6); w = 1.0 / Sigma(appl)^2]. A conjugate gradient refinement was performed on the data and structure described above using 16.91 and -66.92 for the constants A and B, respectively. While these values are similar to that used in the PROFFT refinement, values of A and B may be chosen to best match the distribution of $|\Delta F|$ as a function of $\sin\theta/\lambda$ since the value of WA obtained from the CHECK protocol in XPLOR will proportionately scale the individual weights $[(WA/\Sigma w F_o^2)\Sigma w\Delta^2]$ in the crystallographic target function. There is a marked improvement in the plot of the goodness of fit against $\sin\theta/\lambda$ (Figure 4a) and the $\delta(R)$ plot (Figure 4b) is now reasonably linear. These plots can be contrasted to those obtained with unit weights, illustrated in Figures 2 and 3.

Alternatively, weights can be directly input via the diffraction data file. In

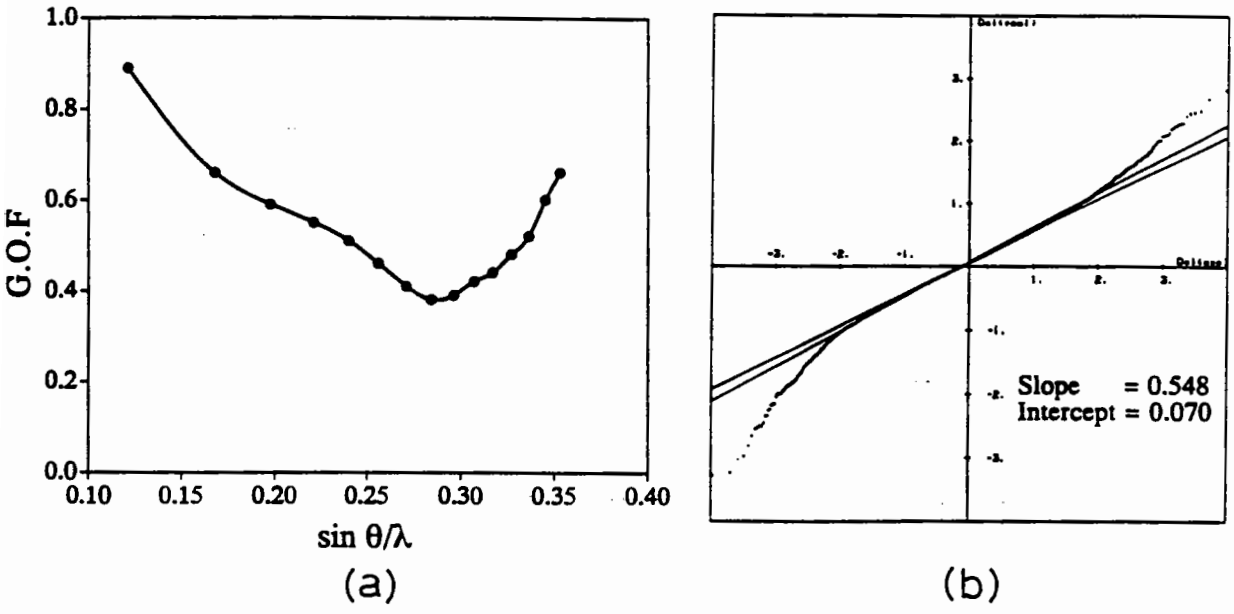


Figure 4. (a) Plot of the goodness of fit versus $\sin\theta/\lambda$ and (b) a $\delta(R)$ normal probability plot for the XPLOR conjugate gradient refinement.

Table II

Residual and Goodness of Fit for the Tox-II Structure

		Unit Weights		Two-Line Weights	
Resol.	n	R	GOF	R	GOF
2.44	2102	0.194	0.744	0.233	0.366
1.95	2072	0.187	0.459	0.209	0.290
1.71	2040	0.185	0.309	0.199	0.228
1.55	2016	0.182	0.239	0.191	0.208
1.44	1987	0.186	0.204	0.190	0.208
1.36	1959	0.188	0.174	0.190	0.214
1.29	1952	0.208	0.163	0.202	0.242
1.23	1934	0.202	0.150	0.192	0.265
1.19	1893	0.212	0.150	0.200	0.272
1.14	1888	0.221	0.146	0.206	0.262
1.11	1841	0.232	0.142	0.214	0.252
1.08	1746	0.276	0.152	0.252	0.268
1.05	1677	0.300	0.144	0.268	0.249
1.02	1603	0.322	0.140	0.303	0.253
1.00	1629	0.339	0.136	0.321	0.244

another test, a two-line weighting scheme was used for a simulated annealing refinement on a partially refined 64 residue toxin structure. The results for the two-line scheme are compared to that obtained using unit weights in Table II. Again, it can be seen that there is a considerable difference in the distributions of the GOF and residual as a function of $\sin\theta/\lambda$. Unlike unit weights, the two-line weighting scheme produces an even distribution of the GOF as a function of resolution, resulting in an equal contribution of all data to the function minimized. The smaller residual in the lower resolution ranges for unit weights might be expected, since these data make the largest contribution to the function minimized. As bulk solvent does make a contribution to the lower resolution data, but is not included as part of the model, one might expect the residual in these shells to be somewhat larger, as observed for the two-line scheme. A decrease in the residual for the higher resolution shells is noted for the two-line scheme. As observed in the other example, the $\delta(R)$ plot is sigmoidal for unit weights (Figure 5a) but linear for the two-line weighting scheme (Figure 5b). Similar results have also been obtained for a 1800 residue protein which diffracts to 2.0Å resolution (P.L. Howell, private communication).

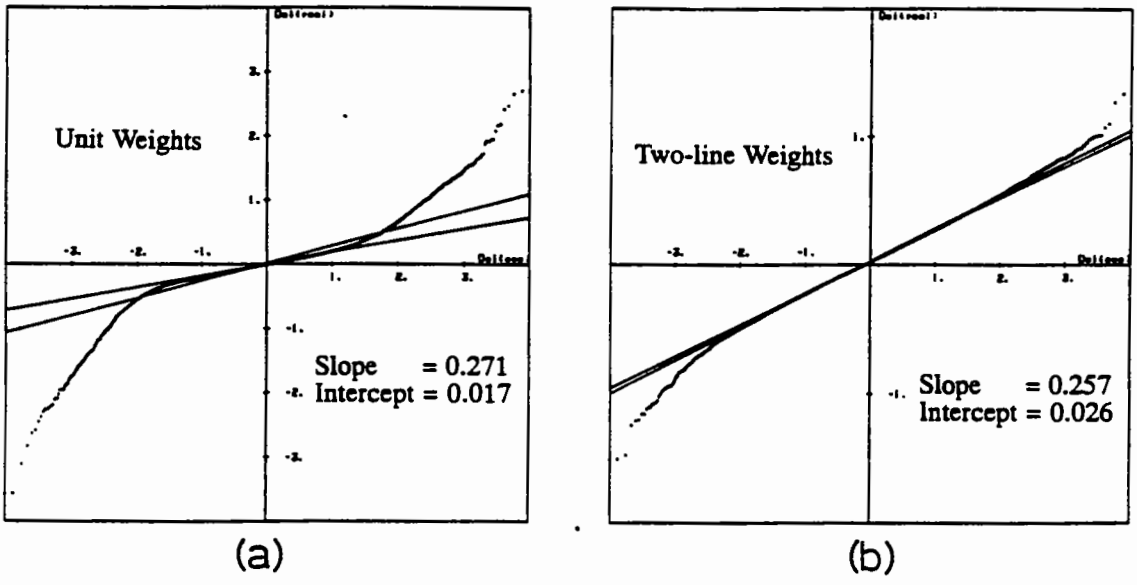


Figure 5. $\delta(R)$ normal probability plots for the XPLOR simulated annealing refinement of the toxin data using (a) unit weights and (b) the two-line empirical weighting scheme.

Except in unusual circumstances, these results should discourage the use of experimental or unit weights in a restrained or simulated annealing refinement. While most users carefully examine the distribution of the residual as a function of

resolution, equal emphasis should be given to other figures of merit, such as the goodness of fit, and the regular examination of $\delta(R)$ plots.

Acknowledgments:

The author wishes to thank Drs. B.M. Burkhart and E. Ciszak for their assistance in performing the refinements and to Drs. P. Lynne Howell and Robert H. Blessing for many helpful discussions. This research was supported by NIH grant GM46733.

References:

Abrahams & Keve. (1971). Acta Cryst., **A27**, 157.
Brünger, A.T., Kuriyan, J. & Karplus, M. (1987). Science, **235**, 458.
Brünger, A.T. (1992). *XPLOR Version 3.1: A system for X-ray Crystallography and NMR*. Yale University Press, New Haven, CT.
Finzel, B.C. (1987). J. Appl. Cryst., **20**, 53.
Hendrickson, W.A. & Konnert, J.H. (1980). In "*Computing in Crystallography*", (Eds. R. Diamond, S. Ramaseshan & K. Venkatesan), Indian Academy of Sciences, Bangalore, pp. 13.01-13.25.
Howell, P.L. & Smith, G.D. (1992). J. Appl. Cryst., **25**, 81.
Sheldrick, G.M. (1993). *SHELXL93. A Program for the Refinement of Crystal Structures*. Univ. of Göttingen, Germany.