DL/SCI/R16

REFINEMENT OF PROTEIN STRUCTURES

Proceedings of the Daresbury Study Weekend
15-16 November, 1980

Compiled by P. A. Machin, J. W. Campbell and M. Elder

Science and Engineering Research Council

DARESBURY LABORATORY

Daresbury, Warrington, WA4 4AD

ISSN 0144-5677

REFINEMENT OF PROTEIN STRUCTURES

Proceedings of the Daresbury Study Weekend

15-16 November, 1980

Compiled by

P.A. Machin, J. W. Campbell and M. Elder

Daresbury Laboratory

Science and Engineering Research Council

DARESBURY LABORATORY

1981

"That which is not restricted will take its liberties"
Wayne Hendrickson

"There are no general rules for refining protein structures"
Eleanor Dodson

"Some of you might ask if I cheated with the contour levels. I didn't, because it wasn't required"
Bhat

"There is no general recipe for protein refinement. It's a bit like bringing up a child, each has its own problems, and needs a lot of attention and a lot of love".
Mike James

FOREWORD

The refinement of protein structures has recently become one of the developing areas of protein crystallography. During the last ten years there has been a large expansion in the field of protein crystallography and, with improvements in structure solving techniques, a rapidly increasing number of structures have been solved. In particular a detailed interpretation of enzyme function may be deduced from such studies, but usually only when refinement has been carried out to a stage which allows a description of the molecule in atomic detail. It is therefore clear that the development of reliable refinement techniques is of great importance at this time.

Throughout the history of protein crystallography the parallel development of computer technology has been most significant. The availability of fast computers today is vital both in the initial data processing and structure solving and in subsequent refinement work. Consequently the availability of vector processing machines has enormously improved the possibilities of refinement work.

A Study Weekend was held at the Daresbury Laboratory on 15-16 November 1980, to discuss the present status of refinement techniques. Constrained and restrained methods of refinement, with and without fast fourier techniques, were discussed, consideration being given both to the theory of the methods and to practical experiences of refinement. The programme for the weekend was structured in such a way as to leave an adequate amount of time for discussion and these sessions were particularly lively, stimulating an exchange of ideas and generating a range of points which will be the subject of much future work.

We believe that the meeting was most successful and we would like to thank all the participants for contributing to it. In particular we wish to thank the invited speakers for their valuable contributions both in the formal talks and in the discussion periods and not least for their co-operation in the preparation of these proceedings. We are indebted to Professor Tom Blundell for his help in organising the meeting and for his enthusiastic support.

We thank the Daresbury Laboratory and its Director, Professor A. Ashmore for permission to hold the meeting and for financial support. The meeting was arranged and partially financed in association with the Collaborative Computational Project in Protein Crystallography. We are very grateful to IBM (UK) Limited and to CRAY Research (UK) Limited for financial support for several foreign speakers. It is a pleasure to thank Mrs. Shirley Lowndes and her staff, and Mrs. Christine Thompson for their great assistance with the planning and organisation of the Study Weekend.

Pella Machin
John Campbell
Mike Elder

CONTENTS

PRACTICAL ASPECTS OF STEREOCHEMICALLY RESTRAINED REFINEMENT OF PROTEIN STRUCTURES

by

Wayne A. Hendrickson
Laboratory for the Structure of Matter, Naval Research Laboratory
Washington, D.C. 20375, U.S.A.

1. INTRODUCTION

We have developed a least-squares method for the refinement of atomic models for crystalline macromolecules.[1-3] It simultaneously drives the model to a fit with the diffraction data and with "observations" associated with known stereochemical features. This is done by minimizing a composite observational function,

$$\Phi = \phi_{\text{diffraction}} + \phi_{\text{bonding}} + \phi_{\text{planarity}} + \cdots , \quad (1)$$

which has terms such as

$$\phi_{\text{diffraction}} = \sum_{\text{refns}} \frac{1}{\sigma_F^2} \left(|F_{\text{obs}}| - |F_{\text{calc}}| \right)^2 \quad (2)$$

and

$$\phi_{\text{bonding}} = \sum_{\text{dists}} \frac{1}{\sigma_D^2} \left(d_{\text{ideal}} - d_{\text{model}} \right)^2 \quad (3)$$

The stereochemical observations restrain the model to be compatible with prior knowledge regarding the distributions about "ideal" values for particular features. Presently, we include restraints related to bonding distances, planarity of groups, chirality at asymmetric centers, nonbonded contacts, restricted torsion angles, non-crystallographic symmetry and thermal parameters. Many of these yield observational functions that are equivalent to terms in typical potential energy descriptions.

We have described elsewhere the theoretical basis for this method, its relationship to other refinement procedures, details of individual terms in the observational function, and the use of conjugate gradient techniques in solving the normal equations.[1-5] These descriptions of stereochemically restrained refinement also refer to several applications; many other applications are in progress and some are recently published. I do not propose to repeat here what is given in earlier reports nor will I describe particular applications. Rather, I will focus on some practical considerations in the use of stereochemically restrained refinement as we have it implemented.

2. PROGRAM DESIGN

The initial application of restrained refinement was not to a protein but to the mineral tridymite (a silica structure that has 240 atoms per asymmetric unit and is twinned four ways). Konnert[1] adapted ORFLS[6], a standard crystallographic refinement program, to include twinning, distance restraints and conjugate gradients for use in the tridymite refinement. When we decided to test this new refinement method on proteins, we found adaption of the refinement program itself to be quite straightforward but soon realized the need for a general program to identify restraint distances and specify ideal values. It was also evident that restraints on features other than distances would eventually be needed in protein refinements. Hence we designed a set of protein refinement programs

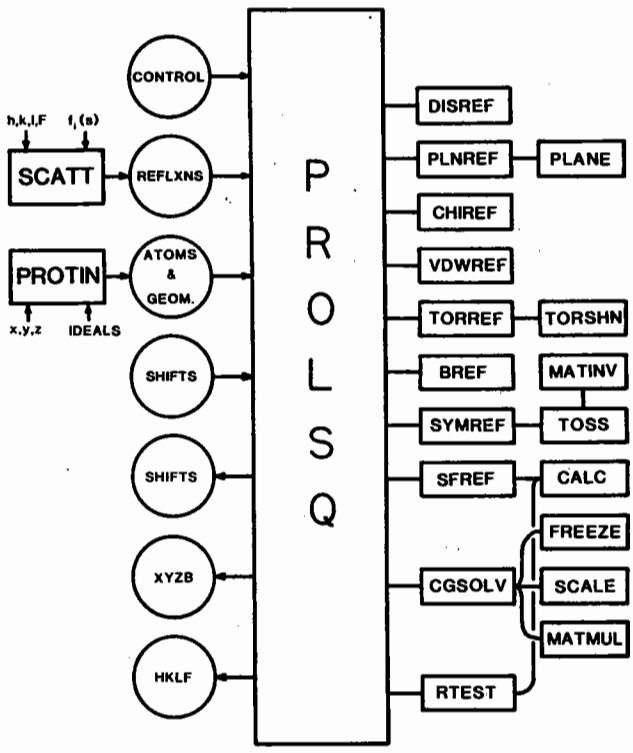


Fig. 1. Schematic structure of the PROLSQ (PROtein Least SQuares) refinement programs.

that we could quickly implement for our specific test problem, parvalbumin, but ones that could readily be extended to include other restraints and generalized to apply to other molecules.

The parvalbumin test clearly showed the need for planarity restraints. This capability was then added to the rudimentary program set. Experience gained in further applications (both at NRL and elsewhere) has fed back other enhancements. These have been incorporated in turn until the exported programs now have the overall structure shown in Figs. 1 and 2. The evolutionary process continues; a new set of programs will soon supplant the present export versions.

PROLSQ (PROtein Least SQuares) is the actual refinement program. It reads diffraction data and scattering factors prepared by SCATT (SCATTering data), initial atomic coordinates and restraint specifications prepared by PROTIN (PROTein model INput), parameter shifts from previous refinement cycles, and control card-images. It then augments the normal-equation elements pertinent to each of the stereochemical restraints and the structure factor observations. Fractional atomic coordinates are used in order to speed the rate-limiting calculations concerning structure factors. For the same reason, a highly optimized space-group specific routine, CALC, is used for computing structure factors and their derivatives. Elements of the resulting sparse normal-matrix are stored in a singly dimensioned array that is indexed by pointers. Next, PROLSQ uses a conjugate-gradients procedure to solve the new parameter shifts. Finally, it tests the expected impact of the new shifts on the R-value. An optimal shift damping-factor is searched for in trials against a sample of the data.

PROTIN is run once before a series of refinement cycles to prepare the atomic coordinate data needed by PROLSQ and to identify the atoms and ideal values involved in the individual stereochemical restraints. It incidentally also performs a useful verification function for initial models. Ideal values for the various stereochemical features are taken from those in particular small-molecule crystal structures of constituent parts of the macromolecule. This is patterned after Diamond's early model-building program.[7] Group dictionaries specifying ideal values and the atoms involved are compiled for each category of restraint. These dictionaries are then consulted to produce the specifications for a particular polymeric structure. PROTIN was written with proteins specifically in mind, but it could quite readily be adapted for other polymers. Indeed, Gary Quigley has written a corresponding NUCLIN for nucleic-acid refinements (private communication).

We have written the programs in FORTRAN for the TI-ASC (a vectorizing machine based on 32-bit words) at NRL. Although the code is optimized for the ASC, care has been taken to avoid non-standard features that might seriously impede transportability. The programs have in fact been implemented on quite a variety of computers. Execution times of course depend on the problem and on the machine. Isotropic refinement of crambin at 1.5A resolution (414 atoms and 5638 reflections) consumes 35 sec. for each PROLSQ cycle on the ASC whereas a cycle of refinement on β_4 hemoglobin at 2.5A resolution (4664 atoms and 16918 reflections) takes 751 sec. Nearly 94% of the latter time is spent in SFREF and CALC (see Fig. 1). The same refinement of β_4 hemoglobin (which includes individual temperature factors and non-crystallographic symmetry restraints) uses 4.8 hours of cpu time per cycle on a DEC VAX 11/780 that has a floating-point accelerator. This is with a version modified by Pat Briley of Iowa to minimize page faults on the VAX; the standard export version would have taken much longer (A. Arnone, private communication). Briley's modification splits PROLSQ into two parts. Gerson Cohen of NIH (private communication) has achieved similar page-faulting economy on the VAX without dividing the program.

It is clear that the use of Fourier transformations to compute structure factors and gradient vectors might greatly improve speed for large problems.[8,9] However, on vectorizing machines, such as the ASC or Cray, speeds for smaller proteins are already so good as to provide little incentive for FFT acceleration. The use of array processors on other machines can also greatly increase the speed of structure factor related computations (Bill Furey of the VA Hospital in Pittsburgh, private communication).

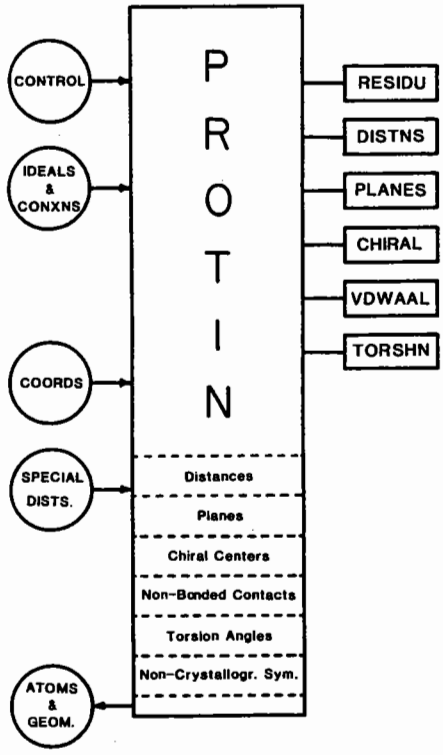


Fig. 2. Schematic structure of the PROTIN (PROTein INput) program.

3. WEIGHTING

The theory for weighting the observations in a least-squares optimization of the fit of a mathematical model to a set of observations is well established for linear problems and at the absolute minimum in non-linear problems. In the case of independent observations the appropriate final weights are inverses of the variances as indicated in eqns. 2 and 3. If there are correlations between observations then added cross terms will arise from a weight matrix which is properly the inverse of the variance-covariance matrix for the observations. Particularly among some of the restraint observations there clearly is interdependence which we ignore in order to avoid the complication of formulating the problem with a non-diagonal weight matrix. But a more serious affront to the conditions of this weighting theory is the highly non-linear nature of the crystallographic refinement problem for which, in the case of proteins, we almost never reach the absolute minimum.

In practical terms, the optimal weighting strategy for a non-linear least-squares problem is that which gives the fastest convergence to the deepest accessible local minimum if not to the absolute minimum. There is no established theory to govern this weighting. Yet experience shows that progress in protein refinements is quite sensitive to weighting. Thus it may be useful here to recount a weighting scheme that we find to be reasonably effective.

In keeping with the theoretical basis to be approached as a refinement nears completion we cast the weights in terms of variances ($W = 1/\sigma^2$). Each class of observation is assigned a "standard derivation", σ, related to the desired breadth of the distribution for features of this class. Generally there is little difficulty in driving the local geometry of a model to an excellent fit with the ideal. What proves to be extremely troublesome is the bringing about of a match between calculated and observed diffraction data. The essential problem in weighting reduces to balancing the weights for diffraction terms against those for stereochemical terms.

It is convenient to maintain an internal balance among restraint weights by placing these on a quasi-absolute scale. But in the early stages of a protein refinement the discrepancy between $|F_{obs}|$ and $|F_{calc}|$ is generally much greater than the counting-statistics error in $|F_{obs}|$. Sigmas from counting statistics will then vastly overweight the structure factor terms and thereby cause ill-conditioning and deterioration of model ideality. However, good behavior generally follows if structure-factor sigmas are set at a uniform value related to the average discrepancy between $|F_{obs}|$ and $|F_{calc}|$. As a rule of thumb, $\sigma_F \simeq 1/2 < \left| |F_{obs}| - |F_{calc}| \right| >$ is a good starting point. The value of σ_F is then used to control the progress of refinement whereas the restraint sigmas are varied primarily to fine-tune the model. Second moments of the distributions in each of the restraint classes are computed during each PROLSQ cycle and compared with the weighting sigmas to monitor progress. A set of typical weighting parameters for the current version of PROLSQ are given in Table 1.

Table 1

Typical weighting parameters for various observational classes

Bonding Distances			
Bond length	(1-2 neighbor)		σ_D = 0.02 A
Angle-related distance	(1-3 neighbor)		0.04
Intraplanar distance	(1-4 neighbor)		0.05
Hydrogen bond or Metal coordination			0.05
Planar Groups			
Deviation from plane			σ_P = 0.02 A
Chiral Centers			
Chiral volume			σ_C = 0.15 A^3
Non-bonded Contacts			
Single torsion		σ_V = 0.50A	Δd_{VDW} = -0.30
Multiple torsion		"	0.00
Possible hydrogen bond (X....Y)		"	-0.20
Possible hydrogen bond (X-H..Y)		"	-0.90
Torsion Angles			
Specified	(e.g. helix ϕ and χ)		σ_T = 15°
Planar	(e.g. peptide ω)		3
Staggered	(e.g. aliphatic χ's)		15
Transverse	(e.g. aromatic χ_2)		20
Thermal Factors		anisotropic	isotropic
Main-chain bond	(1-2 neighbor)	$\sigma_{\Delta d}$ = 0.05 A	$\sigma_{\Delta B}$ = 1.0 A^2
Main-chain angle	(1-3 neighbor)	0.10	1.5
Side-chain bond		0.05	1.0
Side-chain angle		0.10	1.5
Non-crystallographic Symmetry		positional	thermal
Tight class		σ_{SP} = 0.05 A	σ_{SB} = 0.5 A^2
Medium class		0.50	2.0
Weak class		5.0	10.0
Restraints Against Excessive Shifts		idealize	refine
Positional parameter		σ_{EP} = 0.1 A	σ_{EP} = 0.3 A
Thermal parameter			σ_{EB} = 3.0 A^2
Occupancy parameter			σ_{EQ} = 0.2
Diffraction Data			
Structure factor modulus		$\sigma_F \approx \frac{1}{2}\langle \lvert \lvert F_{obs} \rvert - \lvert F_{calc} \rvert \rvert \rangle$	

4. REFINEMENT STRATEGY

There can be no single prescription for the refinement of protein structures -- each problem is idiosynchratic. The range of starting situations includes questionable models based on poorly phased maps, models developed by molecular repositioning from a related but appreciably different structure, and excellent models based on several good derivatives. Obviously, different courses will be followed in the refinement and revision of such disparate models. Nonetheless there are many common considerations in devising a strategy for conducting a refinement.

4.1 Resolution limits

One general area of concern relates to resolution limits. A typical starting model might be based on a 2.5A resolution map phased with several derivatives to 3 or 4A resolution but only one that goes to 2.5A. There might also be a full 2A set of native data. What data should be included at the start of refinement and when should others be added? Clearly the radius of convergence is greater at lower resolution; discrepancies between model and true positions must be within a quarter wavelength of the Fourier wave correspondent to a given reflection for shift indications even to have the right sign. On the other hand, if high-angle terms used in the map are excluded from the refinement

one then runs the risk of squandering properly interpreted details. As a rule it is wise to start refinement with data corresponding to the resolution of the strong features in the map to which the model was fitted. At the other end of the scale, very low order terms (spacings greater than 10A) are greatly influenced by the solvent continuum and those in the 5-10A shell are usually heavily affected by partially ordered solvent molecules that are omitted in early models. It will often be wise to exclude these data until solvent is properly treated.

Our typical refinement problem might start with a few cycles of refinement against the 10A - 3A shell of data. One might then exclude the 10A - 5A shell as the R-value in this shell becomes relatively worse than in the 5A - 4A shell. When progress slows at 3A resolution, higher angle data might be included in a few shell-wise expansions (e.g. 2.5A, 2.2A and then 2.0A) with three additional cycles of refinement per expansion.

4.2 Manual revision

Except in circumstances so rare that they have yet to be encountered, an initial atomic model of a protein cannot be brought directly by automated refinement to a satisfactory match with the diffraction data. The refinement invariably becomes stuck in one of the many false minima that abound in such marginally overdetermined problems. Manual intervention is then essential to further progress. Refinement usually proceeds in stages of automated cycles followed by gross revision motivated by inspection of Fourier syntheses based on model phases.

Especially in the early stages and at moderate resolution (2-3A), the $(2|F_{obs}| - |F_{calc}|) \times \exp(\alpha_{calc})$ synthesis (or one of its generalization) is particularly convenient for identifying major imperfections and incompleteness in a model. The $(|F_{obs}| - |F_{calc}|) \exp(\alpha_{calc})$ synthesis often provides definitive clues for a revision, especially at higher resolution. Another useful synthesis has the coefficients $(|F_{obs}| - |F'_{calc}|) \times \exp(\alpha_{calc})$ where F'_{calc} is the structure factor of a partial model from which a fragment has been deleted. We have found it advantageous to examine a systematic series of such fragment ΔF-maps showing segments about ten residues in length.

The initial refinement usually greatly improves the quality of maps based on model phases, often to the point of making them superior to the initial map based on experimental phases. However, there may sometimes, particularly for incomplete models, be considerable advantage in combining the model phase information with that from the MIR or other phasing experiment. One then uses a synthesis based on coefficients of $m|F_{obs}|\exp(\alpha_{combine})$ where the figure-of-merit, m, and centroid phase, $\alpha_{combine}$, are computed from the combined phase probability distribution for the two kinds of information.

It is important to be able to superimpose the model on an image of an electron-density distribution. This can be done on paper sheets or transparency stacks. Of course, computer graphics systems for molecular modelling not only solve the display problem exceptionally well but they also greatly facilitate the rebuilding process.

4.3 Thermal parameters

Strategic decisions for the course of a refinement must also be made regarding the treatment of thermal parameters. There usually is substantial and meaningful variation in vibrational amplitudes within protein structures. This must be expressed to effect a good match with the diffraction data. Yet on the other hand, one must worry that at moderate resolution freely varying thermal parameters can take on meaningless values as they simply absorb errors in a problem that approaches underdetermination. The high correlations between thermal parameters and scale or occupancy factors must also be taken into account.

The approach to thermal parameters that we have commonly taken in refinements at NRL goes as follows: Intensity statistics from all available data are used to place the data on an approximately absolute scale and to estimate an overall thermal parameter (B value). In the early, low resolution cycles of refinement only the scale factor is varied and the overall B is held fixed. (If B is allowed to vary at 3A resolution it typically takes on unreasonably low values.) After refinement has been extended to somewhat higher resolution and has reduced R to the vicinity of 0.30, then individual isotropic thermal parameters will be varied but with relatively tight restraints. The intent here is to permit a smoothly varying expression of dif-

ferences in thermal factors. Later, if there are sufficient data, we will release restraints on B values. This permits the effective elimination of atoms that are grossly misplaced, it often identifies wrongly interpreted regions, and at higher resolution it can be used to discriminate between atom types (e.g. N vs O in amide side chains). Finally, highly restrained anisotropic thermal parameters might be used.

Of course, nearly endless variations are possible in the approach to thermal parameter refinement. As with weighting, facilitation of the process and avoidance of false minima are guiding principles during intermediate stages whereas conformity with stereochemical rules is the criterion for a final model.

4.4 Solvent structure

Although much of the large solvent fraction in protein crystals is essentially fluid, usually sufficient of it is well enough ordered to make inclusion of the solvent structure necessary in an advanced refinement model. Properties of the solvent structure will often be of interest in their own right. In addition, frequently, correct modeling of the protein structure can only be completed after waters and other solvent molecules have been included.

The most tightly bound of the solvent molecules can be readily identified. However, much of the solvent structure cannot be located until after initial refinement has reduced the R-value to 0.25 or less. It is often helpful to include the solvent affected 10A - 5A shell of data in refinement cycles leading to difference maps for solvent interpretation. One should be mindful of hydrogen-bonding preferences when ascribing density features to solvent molecules.

Many solvent sites are only partially occupied. Thus occupancy parameters are essential variables in the solvent model. However, occupancy factors are highly correlated with thermal parameters. In lieu of a full-matrix treatment of these covariations, for those atoms with variable occupancy factors we simply alternate the application of occupancy and thermal parameter shifts in successive cycles. In order to minimize the inclusion of meaningless variables we generally eliminate solvent sites that refine to very low occupancy factors (e.g. $Q<0.3$) or very high thermal values (e.g. $B>50A^2$).

4.5 Conformational restraints

It will be desirable in the early stages of some refinement problems to reduce the conformational freedom in a model. If the model is sufficiently restricted, meaningful refinement is then possible against low resolution data and this can yield a large radius of convergence. Refinement of a model composed of linked rigid groups[10] affords a direct and fruitful approach. It is also possible to effect a reduction of conformational freedom by imposing certain tight restraints.

One option is to restrain the torsion angles of the model to remain very close to those in the initial model (e.g. $\sigma_T = 2^o$). This is particularly sensible in the case of a model derived from molecular replacement of a related and well known structure. Another option is to restrain the backbone torsions in elements of known secondary structure to be those expected in ideal helices, β-sheets, etc. The inclusion of special distance restraints related to the hydrogen bonding within structural elements also rigidifies these units. Of course, as the refinement proceeds to higher resolution it will usually be wise to relax special conformational restraints.

A special conformational restraint that we have found generally useful concerns peptide planarity. In the initial stages we usually restrain five atoms of a peptide unit (C_α^i, C^i, O^i, N^{i+1}, C_α^{i+1}) to be coplanar. Later, after high-angle data have been included and R is below 0.20, we will drop to four-atom peptide planes (C_α^i, C^i, O^i and N^{i+1}) and rely on torsion restraints to maintain reasonable ω angles.

4.6 Convergence

Refinements at low to moderate resolution, particularly in the early stages, are prone to ill-conditioning. That is, the marginal state of overdetermination leads to near-singularity in the normal matrix and consequent instability in the solution for parameter shifts. The conjugate-gradients (c-g) method is relatively insensitive to ill-conditioning, but such as exists manifests itself by slow and irregular convergence, or even

divergence, of the c-g iterations. PROLSQ monitors the c-g shifts for a sample of parameters. If these are not monotonically convergent a change in the relative weighting between diffraction and stereochemistry observation might be needed or restraints against excessive shifts (pre-conditioning) should be invoked.

Restraints against excessive shifts are simply tethers to current parameters. In the absence of significant pressure for movement, parameters will then be held near their starting values. A suitable choice of the weighting sigmas for these restraints permits model idealization by PROLSQ even without diffraction data. These restraints also control the behavior of groups that are poorly specified by the data (e.g. some lysine side chains).

Generally, continued refinement after the first few cycles (5-10) on a given model leads to diminishing returns and manual revisions are soon in order. Sometimes, though, the progress can be rejuvenated by a sudden relaxation of stereochemical or thermal restraints or by dropping back to lower resolution. This can relax barriers between local minima and increase the radius of convergence. However, such a process also runs the risk that properly fitted features might escape and be trapped into wrong positions when tight restraints and high resolution are reimposed.

5. PROBLEM AREAS

The ultimate goal of a protein refinement is a structural model which reproduces the diffraction pattern to within the accuracy of the data and which is compatible with prior stereochemical knowledge about the molecule. To my knowledge, in no case has this yet been accomplished. The major impediment is motion and disorder.

Vibrational amplitudes and structural variability tend to be large in protein crystals. This causes diffraction intensities to diminish rapidly with scattering angle and greatly limits the extent of measurable data. The distribution functions required to model the highly anisotropic and anharmonic characteristics of such large displacements must have many variables. Yet it is just in the case of large displacements that data are few. Hence these interesting and necessary parameters tend to be indeterminate. The challenge is great to produce adequate but economical models for atomic displacements in proteins.

Another unsolved problem relates to the specification of errors in atomic positions and derivative quantities such as bond distances. In principle the method of least squares gives these values. However this formally depends on having reached the absolute minimum for a given model with correctly weighted observations.

In the case of stereochemically restrained refinement there are special difficulties in error analysis. If this is to be based on the fit to the entire minimization function, eqn. 1, then the proper relative weighting of terms is crucially important and interdependence among certain stereochemical observations should be taken into account. On the other hand, one may view the stereochemical terms merely as a computational device for restraining parameters and wish to judge errors solely from the fit to diffraction terms, eqn. 2. In doing so, one would expect that the reduced freedom in the model imposed by the restraints should be reflected in the analysis. However, it is not clear how an "effective" number of parameters of degrees of freedom could be arrived at.

There are related difficulties of comparing results from constrained, restrained (loosely or tightly) and free refinements. For example, it can be argued that restrained refinement results are less significant than those from constraints because of added degrees of freedom or that an appreciably lower R-value makes a free atom model better. Attempts to apply significance tests to resolve such questions are frustrated by difficulties in counting truly independent observations and actually effective parameters. It would help to have a good overall measure of the stereochemical reasonableness of a model -- one as comprehensive and comprehensible as is the R-value for diffraction data.

ACKNOWLEDGEMENTS

I thank John Konnert for his collaboration in the development of our refinement program and several users for feedback that has led to improvements in program capabilities and in strategies for implementation.

REFERENCES

1. J.H. Konnert, Acta Cryst. A32, (1976) 614-617.

2. W.A. Hendrickson and J.H. Konnert in Biomolecular Structure, Function, Conformation and Evolution (R. Srinivasan, ed.), Vol. 1 (Oxford: Pergamon, 1980) 43-57.

3. W.A. Hendrickson and J.H. Konnert in Computing in Crystallography (R. Diamond, S. Ramaseshan and K. Venkatesan, eds.) (Bangalore, Indian Institute of Science, 1980) 13.01-13.23

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

5. W.A. Hendrickson and J.H. Konnert, Biophys. J. 32, (1980) 645-647.

6. W.R. Busing, K.O. Martin and H.A. Levy, ORFLS, ORNL-TM-305 (Oak Ridge National Laboratory, Oark Ridge, Tennessee, U.S.A., 1962).

7. R.Diamond, Acta Cryst. 21, (1966) 253-266.

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

9. A. Jack and M. Levitt, Acta Cryst. A34, (1978) 931-935.

10. J.L. Sussman, S.R. Holbrook, G.M. Church and S.-H. Kim, Acta Cryst. A33, (1977) 800-804.

WEIGHTING IN THE RESTRAINED LEAST SQUARES REFINEMENT OF PROTEIN STRUCTURES

by

D.S. Moss
Department of Crystallography, Birkbeck College, Malet Street, London, WC1E 7HX, England

1. INTRODUCTION

The purpose of this paper is to outline a strategy which may be used for statistically weighting the least squares refinement of the crystal structure of a protein.

Techniques for the weighting of structure amplitudes in small molecule least squares refinements have been discussed by several authors[1,2,3] and a good summary has been provided by Hamilton[4]. However the weighting of least squares refinements of macromolecules presents extra difficulties associated with the increased disorder and larger unit cells which usually occur in these crystal structures. The resulting poorer observation-to-parameter ratio would lead to stereochemically unacceptable features in an unrestrained structure amplitude refinement. This problem may be partially overcome by the use of prior knowledge of sterochemistry or energy criteria in order to restrain the refined parameters. It is necessary to know how to weight this prior knowledge relative to the diffraction data.

A second problem is related to the disorder itself. The interpretation of electron density maps in disordered regions is difficult and simple structure factor models may only roughly account for the time and space averaged distributions in these regions. For medium or strong reflections the errors due to the functional form of the structure factor may be more significant than the errors in experimental measurement and must be taken into consideration in the weighting process. It is for this reason that weights derived from counting statistics alone seldom give satisfactory results in least squares refinement.

There are thus two important problems encounted in the proper use of restrained least squares refinement.

(i) How to weight the restraints relative to the structure amplitudes.

(ii) How to weight the reflections relative to each other.

2. PURPOSE OF WEIGHTING

Weighting may assume two distinct pruposes in the refinement of a protein structure. Firstly it may be used to drive the refinement down the correct minimum in as few cycles as possible. This may be called convergence weighting. Secondly in the later stages of refinement, the weights may be used to reflect the expected discrepancies between observations or target values or functions and the corresponding quantities calculated from the model. The weighting strategies to be adopted in the two cases may be quite different and this paper is solely concerned with the second case which may be called statistical weighting because the weights are chosen according to statistical criteria. In simple least squares theory it is assumed that the weights are known at the beginning of the analysis and that the proposed model is correct. Under these assumptions, an unrestrained least squares refinement yields minimum variance unbiased estimates of the crystallographic parameters. However when the proposed model contains errors which result in unpredictable effects on the agreement between observed and calculated quantities, the weights must be partially determined from the data itself. Furthermore when prior information is included in a least squares refinement, the price paid for the lower variances of the parameter estimates is that they become biased. The relative weighting of restraints versus observations governs the balance between the conflicting requirements of minimum bias and minimum variance. The balance may be struck by using the well-known statistical criterion of maximum likelihood[5].

3. MAXIMUM LIKELIHOOD ESTIMATION

Let us assume that n observations $(f_1^o, f_2^o, \ldots f_n^o)$

have been made. These observations will typically be structure amplitudes but could also be phases from isomorphous replacement or anomalous scattering. They should be corrected for systematic errors which could be simulated by adjustment of the refined parameters during refinement. Absorption effects, for example, will cause perturbations in the thermal parameters of the atoms. We now assume that the true values $(f_1^o, f_2^o, \text{---} f_n^o)$ are normally distributed about the observations.

$$P_1(f_1, \text{---} f_n) = \frac{1}{(2\pi)^{\frac{n}{2}} \prod_{i=1}^{n} \sigma(f_i)} \exp\left[-\sum_{i=1}^{n} \frac{(f_i^o - f_i)^2}{2\sigma^2(f_i)}\right] \quad (1)$$

The variance of f_i is denoted by $\sigma^2(f_i)$ and we have assumed that the errors in the observations are uncorrelated.

The restraints which are prior information may also be represented by a probability distribution.

$$P_2(d_1, \text{---} d_p) = \frac{1}{(2\pi)^{\frac{p}{2}} \prod_{i=1}^{p} \sigma(d_i)} \exp\left[-\sum_{i=1}^{p} \frac{(d_i^T - d_i)^2}{2\sigma^2(d_i)}\right] \quad (2)$$

The $(d_1, d_2, \text{---} d_p)$ represent functions of the geometrical or thermal parameters and the d_i^T are their target values. For simplicity the d_i are here assumed to be uncorrelated and this assumption is also made in most restrained least squares programs. However it should be noted that this assumption is not really valid. From a statistical point of view the errors to be feared in bond lengths, bond angles, chiral volumes etc are not independent. From the energy standpoint the potential energy of a molecule cannot be accurately represented as a sum of squares in any coordinate system employing simple geometrical parameters.

The assumption of normality may also be questioned but it must be remembered that most statistical tests which might be made on the least squares parameters require the form of a probability density function to be assumed at some stage during the analysis.

The probabilities associated with the observations and prior information may be multiplied together to give a joint probability.

$$P(f_1, f_2, \text{---} f_n, d_1, d_2, \text{---} d_p) = P_1(f_1, f_2, \text{---} f_n) P_1(d_1, d_2, \text{---} d_p) \quad (3)$$

In order to estimate the variances (the reciprocals of the weights), the observations and restraints may be divided into classes such that within each class it can be assumed that the variance is approximately constant. Structure amplitude classes may each contain reflections of similar amplitude and Bragg angle while geometrical restraints may be classified according to whether they correspond to bond distance, bond angle or van der Waals restraints. If Y_{ij}^o is the j th observation or restraint in the i th class, there are N_i observations in this class, a_{ijk} is an element of the observational equations, then the joint probability P may be written as a log likelihood function of the M parameters θ_k and the N class variances σ_i^2.

$$\log L(\theta_1, \theta_2, \text{---} \theta_M, \sigma_1, \sigma_2, \text{---} \sigma_N) = -\tfrac{1}{2} \sum_{i=1}^{N} Ni \log(2\pi\sigma_i^2) - \tfrac{1}{2} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \sum_{j=1}^{Ni} (Y_{ij}^o - \sum_{k=1}^{M} a_{ijk}\theta_k)^2 \quad (4)$$

To obtain maximum likelihood estimates (MLE) $\hat{\theta}_k$ and $\hat{\sigma}_i$ we require that the appropriate derivatives of the log likelihood function should be zero. This gives;

$$\hat{\sigma}_i^2 = \frac{\sum_{j=1}^{Ni} (Y_{ij}^o - \sum_{k=1}^{m} a_{ijk} \hat{\theta}_k)^2}{N_i} \quad (5)$$

where $\hat{\theta}_k$ are the values of θ_k which minimise,

$$R = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \sum_{j=1}^{Ni} (Y_{ij}^o - \sum_{k=1}^{m} a_{ijk} \theta_k)^2 \quad (6)$$

Equation (6) is the function minimised in restrained least squares while eqn. (5) expresses the condition that the mean weighted residual squared should be unity in each class. The MLE of σ_i given by eqn. (5) is biased but consistent, ie it tends to be a better estimator of σ_i as the resolution increases. It also reflects those errors in the functional form of the model (ie the coefficients a_{ijk}) which cannot be compensated by adjustment of the parameters θ_k. Missing parameters (eg missing atoms) will also contribute to the estimate of σ_i.

In order to apply eqns. (5) and (6), minimisation of R is carried out with adjustment of the weights after each cycle so that they confirm to (5). Experience shows that at a resolution of about 2.5Å or better this process is convergent. At lower resolutions the convergence may not be achieved.

Table 1

Ribonuclease (1.4Å)

Reflections - Average $w(|F_o| - |F_c|)^2$ with no. of reflections in brackets

RESOLUTION (Å) →

$\|F_o\|$	INF - 2.6	2.6 - 2.0	< 2.0	Totals
0 - 50	1.657(122)	0.776(193)	0.518(4064)	0.568(4611)
50 - 100	0.922(343)	0.450(632)	0.605(5756)	0.583(7479)
100 - 200	0.627(839)	0.654(814)	0.793(1269)	0.681(4374)
> 200	0.731(699)	1.255(155)	0.939(42)	0.745(2493)
Totals	0.699(2003)	0.632(1749)	0.591(11131)	0.622(18957)

Restraints - Average $w(d_T-d_C)^2$ with no. of restraints in brackets

Distance (Å)	
< 2.1	0.661(887)
2.1 - 3.0	0.882(1165)

This is because of the correlation of least squares residuals becomes quite significant when the number of degrees of freedom of R is low relative to the number of parameters. The bias on $\hat{\sigma}_i^2$ also becomes prominent under such conditions. The possibility of gross errors in the model is also greater at lower resolutions and thus altogether the difficulty of proper weighting is considerably increased.

4. PRACTICAL CONSIDERATIONS

In the earlier stages of the least squares refinement of a protein molecule the crystallographer will usually want to ensure the widest range of convergence for the technique and thus to correct automatically as many errors as possible in his initial model of the crystal structure. This is achieved by convergence weighting where the higher angle reflections are down weighted or zero weighted. In the early stages many gross errors are usually present and inspection of $(2|F_o| - |F_c|)$ maps and manual corrections are always required during the refinement process. As the model improves, the higher angle data may be given higher weighting and the transition to statistical weights may take place by adjusting the weights after each cycle until eqn. (5) is obeyed. As the progress of a least squares refinement is unaffected by the absolute overall scale of the weights, it is sufficient to adjust the weighted mean square residuals so that they have a constant value between classes. In the usual rotation this means that the averages $\langle w(|F_o| - |F_c|)^2 \rangle$ and $\langle w(d_T-d_C)^2 \rangle$ taken over each class should be constant.

Table 1 shows mean square residuals which are taken from a computer output of a 1.4Å refinement of ribonuclease-A. It indicates that the strong reflections ($|F_o| > 200$) and non-bonded distances are rather overweighted relative to the other observations and restraints. The weights of the reflections were generated by use of a modified version of a formula due to Cruickshank[1] :

$$\omega = \frac{0.2(\sin\theta/\lambda)^{1.0}}{200 + |F_o| + 0.002\,|F_o|^2}$$

It is important to note that the weakest reflections did not have zero weight. They contain important information for the least squares process provided that they are derived from regions of reciprocal space where there are significant intensities.

The two classes of restrained distance used in this refinement were each allocated a constant weight set by the user. The weights for these restraints could have been chosen from spectroscopic force constant data or from the dispersion of observed inter-

atomic distances in structures of small molecules. In this case MLE's would be used to establish the relative scales of the reflection weights and their weighted mean residuals would be scaled to the corresponding restraint values.

Table 2

Ribonuclease: Area showing false minimum

Residue	Atoms	$\lvert\sqrt{w}(d_T - d_C)\rvert$
ASP 14		
	CA-C	2.32
	CA-O	1.72
	CA-CA	2.34
SER 15		
	N-CB	2.88
	CA-CA	2.10
	C-N	2.55
	O-CA	2.57
SER 16		
	N-CB	3.92
	Overall $< \lvert\sqrt{w}(d_T - d_C)\rvert >$ = 0.89	

Another important use of weighted residuals is to focus attention on aspects of the observations, the prior information or the molecular model that require further investigation. An example is shown in table 2. If the target distances are correct then the individual residuals weighted on an absolute scale will be asymtotically distributed with zero mean and unit variance. Table 2 shows a region of the structure where several weighted residuals are about three times the overall average. There are three possibilities which may be considered :

(i) The refinement has converged to a local minimum.

(ii) The target distances are significantly incorrect in the region and unusual stereochemistry is present.

(iii) The region is disordered in a way that has not been correctly modelled in the structure factor formula.

Reason (iii) might be indicated by large temperature factors. Reason (iii) applied in this particular case and although the atoms were not out of electron density in a $(2|F_o| - |F_c|)$ map, a gross error of interpretation had been made.

In examining least squares results it must also be remembered that among the thousands of weighted residuals occurring in a protein refinement, it is statistically to be expected that there will be some outliers which do not indicate anything amiss with the refinement.

5. CONCLUSION

The statistical weighting of restrained least squares may be accomplished by analysing batches of weighted residuals provided that the resolution of the diffraction data is sufficient for the correlation between the residuals to be neglected. Such weighting yields maximum likelihood estimates of the crystallographic parameters provided that the errors are normally distributed.

The usual assumption that the residual minimised in least squares should be a simple sum of squares is particularly unsatisfactory in the case of restraints. Restraints involving common atoms may be significantly correlated and therefore off-diagonal restraint terms should be included in refinement with weights that could be estimated from force constant data.

REFERENCES

1. D.W.J. Cruickshank in Computing Methods in Crystallography. Ed. J.S. Rollett, Oxford: Pergamon, (1965).

2. J.S. Rollett in Crystallographic Computing. Ed. F.R. Ahmed, Copenhagen: Munksgaard, (1970).

3. K. Huml in Computing in Crystallography. Ed. R. Diamond, S. Ramaseshan and K. Venkatesan, Bangalore: Indian Academy of Science, (1980).

4. W.C. Hamilton in International Tables for X-ray Crystallography, IV, Ed. J.A. Ibers and W.C. Hamilton, Birmingham: Kynoch Press, (1974).

5. N.R. Draper and H. Smith, Applied Regression Analysis, New York: John Wiley, (1966).

USE OF CORELS FOR THE REFINEMENT OF BIOLOGICAL MACROMOLECULES

by

Joel L. Sussman
Department of Structural Chemistry, Weizmann Institute of Science, Rehovot, Israel

1. INTRODUCTION

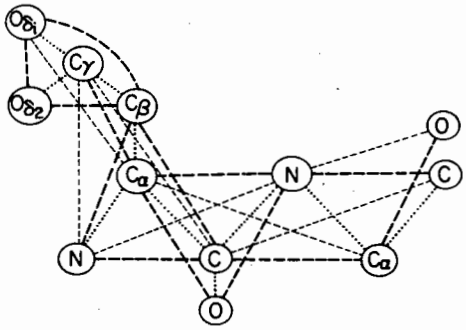
Model building and refinement programs for macromolecules fall into two extreme categories in terms of the way in which the structural parameters can be manipulated (see fig. 1). At one end are the programs where the cartesian coordinates of each atomic position can be varied independently, while the stereochemistry is restrained to standard values by spring-like energy terms between atoms[1,2,3,4]. Using these programs it is easy to move any particular part of a structure to a nearby region of space, but it is difficult to move a large portion of a structure, e.g. an entire domain, and treat it as a unit. At the other extreme are programs where the bond lengths and bond angles are strictly constrained and the only variable parameters are the dihedral angles of the backbone and side chains[5,6,7,8]. The strongest argument in favor of these latter procedures is that except where extremely high resolution data are available, protein X-ray data can not give better estimates of bond lengths and bond angles than those obtained from small molecule crystallography. It therefore seems more reasonable to hold these values fixed, and thus greatly reduce the number of degrees of freedom to be varied in either model building or refinement.

However, in attempting to manipulate molecular structures using only dihedral angles as the variables, with either physical models or on a real-time computer graphics system, one quickly sees that these are not the most convenient parameters to vary in all situations. Often it is difficult to move particular parts of a structure to specific regions in space with only these degrees of freedom. In fact the usual way that physical molecular models are fitted to an electron density map, in an optical comparator[9], consists of grabbing a hold of a fragment of the structure, e.g. a few amino acids or a base-pair, and fitting this portion as a unit, rather than manipulating only dihedral angles along the whole chain. Following the initial fitting, it is possible to idealize the stereochemistry of the connections of the fragment to the rest of the chain by relatively small adjustments. Schemes like this have now been incorporated into several computer graphic systems[10,11].

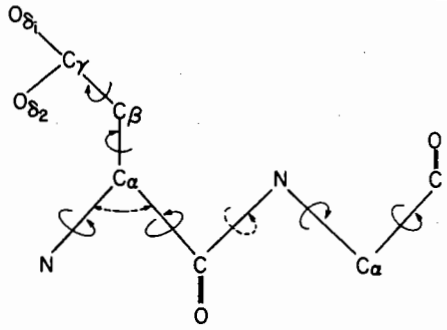
When we began to develop the CORELS (COnstrained-RElstrained Least-Squares) refinement program[12], we were guided by these real-time model building systems as well as by the principle that the amount of a structure to be constrained should be a function of the resolution of data availabe. If one had only extremely low resolution data, then large portions, even domains, could be treated as rigid bodies. At intermediate resolution some secondary structural features such as α-helices in proteins or double-helices in nucleic acids might be treated as separate groups. At still higher resolution, individual amino acids, prosthetic groups or base-pairs might be treated as discrete groups. Finally, if atomic resolution data were available then individual atomic parameters could be varied.

To accomplish this we wrote a computer program to treat a structure as a series of discrete units. Each unit we called a constrained group and it is defined as a molecular moiety where all bond lengths and bond angles have been fixed to respective canonical values but which can have any number of easily defined rotatable bonds. Examples of such constrained groups are shown in fig. 2. It is important to note that in no way is the procedure restricted to any particular set of constrained groups, rather it is up to the user to decide what to constrain for the particular structure, stage of refinement and resolution of data available.

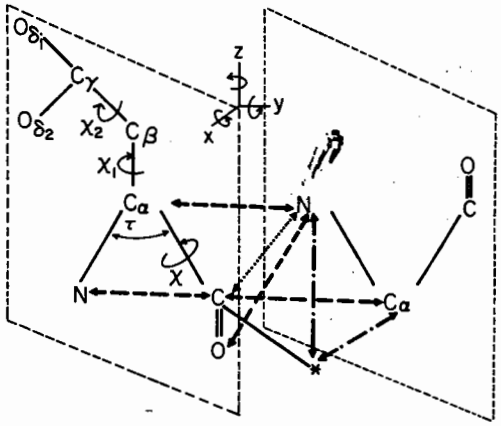
Our first attempt to implement these ideas was in the early stages of the refinement of the yeast $tRNA^{Phe}$ structure. There we refined the structure as a series of unconnected rigid groups made up of phosphate, ribose and base moieties using a program that had been developed for the reciprocal space least-squares refinement of small molecules[13]. We found that although the R factor dropped somewhat there were severe problems: 1) The stereochemistry at the connections between the discrete rigid groups



(a)



(b)



(c)

Fig. 1 Schematic illustration of the ways in which structural parameters can be varied while maintaining stereochemistry in three different kinds of model building/refinement programs. (a) A restrained refinement procedure[1,2,3,4]. The cartesian coordinates of each atom are the variables. The stereochemistry is maintained by specific restraints that correspond to either distances with spring-like connections[1,2,4], or specific bond lengths (dotted lines), bond angles (heavy dashed lines), torsion angles (light dashed lines) or non-bonded contacts (not shown)[3]. (b) A constrained refinement procedure[5,6,7]. Here the variables consist only of the backbone torsion angles and selected bond angles, while all bond lengths are strictly constrained. (The torsion or bond angles indicated with dashes are optional degrees of freedom). (c) A constrained-restrained refinement procedure[12,14]. Here two constrained groups are illustrated. Each is free to move with 6 degrees of freedom (translation & rotation) as well as any number of internal torsion or bond angles. The bond lengths within any one group are strictly constrained. The stereochemistry between groups is restrained by specific distances that correspond to bond lengths (dotted lines), bond angles (dashed lines), torsion angles and non-bonded contacts (not shown). In order to maintain the planar peptide bond, a dummy atom (with zero atomic number) was attached to the carbonyl carbon about 10 Å above the peptide plane as was first suggested in ref. 4. The distance between this dummy atom and the N & Cα of the next amino acid is restrained to a specific value.

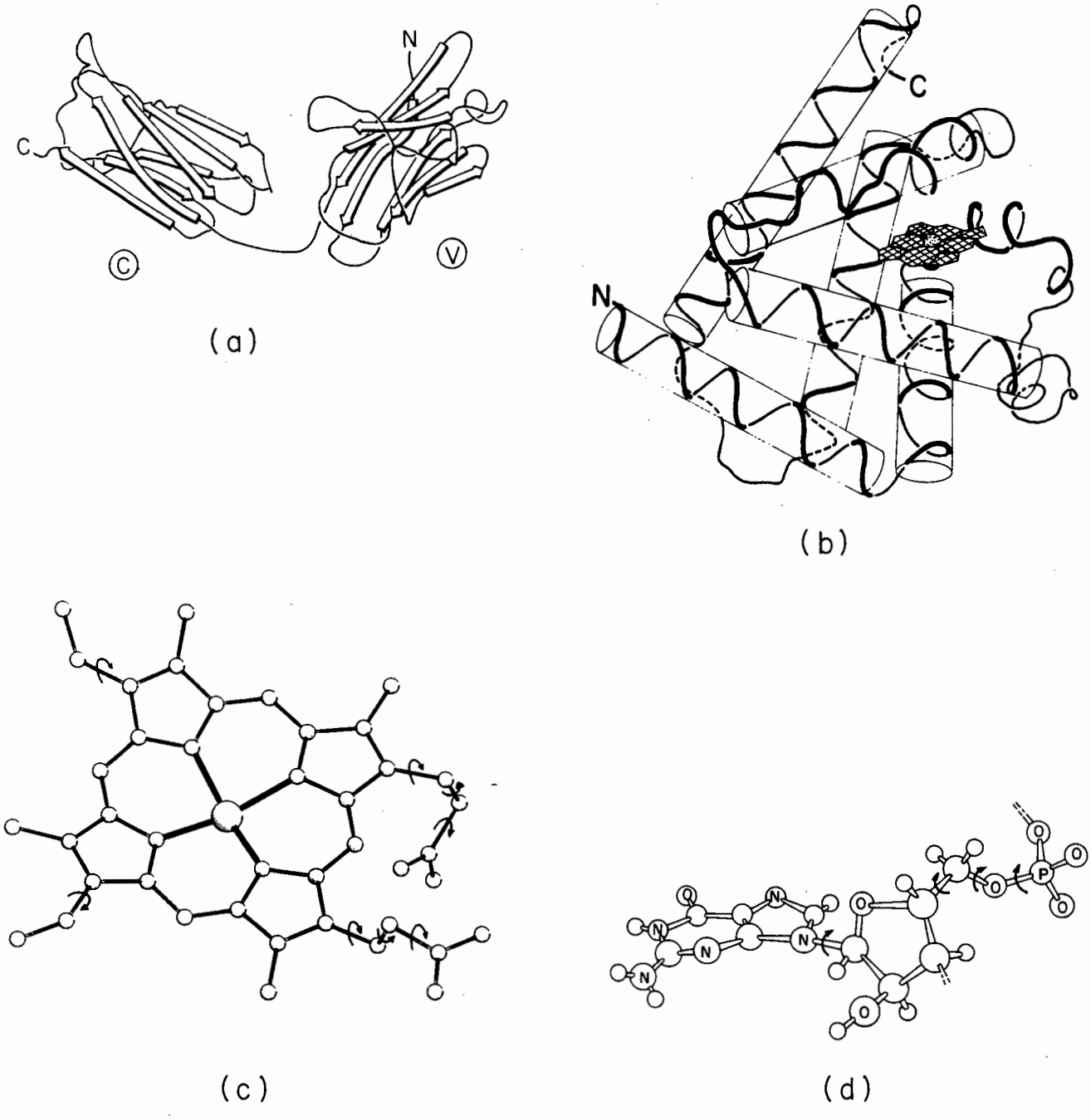


Fig. 2 Examples of different kinds of CORELS groups. (a) two domains of an immunoglobin structure (see e.g. ref 15). (b) six major α-helices and heme group of the myoglobin structure. (c) an enlarged view of the heme group showing the various torsion angles that are additional degrees of freedom[16]. (d) a nucleotide unit showing variable torsion angles.

was very unsatisfactory. 2) There were many unfavorably close contacts between non-bonded atoms. 3) Although the number of degrees of freedom was greatly reduced, the full matrix that was used to solve the normal equation matrix was still too large and thus needed an enormous amount of computer memory and time.

At about this time Hendrickson and Konnert[2] were developing a restrained reciprocal space refinement procedure using a sparse normal equation matrix and the conjugate gradient method for its solution. We incorporated some of these ideas into our rigid group refinement program in order to solve several of the above problems. Specifically, we added harmonic spring-like restraints between the rigid groups in order to maintain acceptable stereochemistry, and the sparse matrix conjugate gradient procedure to reduce computer time and memory. In additon we introduced additional restraints to prevent non-bonded atoms from moving too close together.

The advantages of this procedure are:

1) A large increase in the data/parameter ratio as compared to the restrained refinement methods.
2) Automatic maintainance of group stereochemistry, i.e. up to 80-90% of the bond-lengths and bond angles are strictly constrained.
3) An increased range of convergene as is often seen in small structural work where rigid groups are used.
4) Reduced computing time and memory by combining rigid groups and a sparse matrix.
5) Applicability to low, intermediate as well as high resolution data.

Recently a similar rigid group restrained refinement program using the Gauss-Seidel least-squares procedure was described[14].

2. MATHEMATICAL DESCRIPTION

A detailed mathematical derivation of the equations and derivatives used in CORELS has been given in ref. 12. A brief summary of this is presented here. The quantity to be minimized, Q, in the least-squares procedure consists of the sum of three terms:

$$Q = w_F DF + w_D DD + w_T DT \qquad (1)$$

where w_F, w_D, and w_T are overall weights for each term. The first term, DF, is the usual structure-factor differences summed over all or part of the reflections, h:

$$DF = \sum_h w_h(|F_{o,h}| - |F_{c,h}|)^2 \qquad (2)$$

The second term restrains the stereochemistry and is the sum over all subsidiary distance restraints, d:

$$DD = \sum_d w_d(D_{o,d} - D_{c,d})^2 \qquad (3)$$

where $D_{o,d}$ is the 'ideal' distance between specified pairs of atoms (which may correspond to a bond length, bond angle, torsion angle or a non-bonded close-contact distance) and $D_{c,d}$ is this distance calculated from the model. The third term restrains the structure from moving away from a specified set of target coordinates. Here the sum is over all atoms, i, and over the three axial components, j, of each atom:

$$DT = \sum_i w_i \sum_j (X_{T,i,j} - X_{i,j})^2 \qquad (4)$$

where $X_{T,i,j}$ is the axial coordinate (orthogonal and in Angstroms) of the target atom, while $X_{i,j}$ is the corresponding coordinate of the model.

For restrained-constrained structure-factor least-squares refinement, we set w_T=0, while for distance-target idealization (model building), w_F=0. The relative magnitudes among w_F, w_D and w_T were discussed earlier[2,17].

The quantity Q in eqn. (1), which is to be minimized, is explicitly a function of all the group positional parameters and the thermal parameters of the groups of atoms:

$$Q = Q(t,R,\psi,B) \qquad (5)$$

where t,R,ψ and B refer to all group translation vectors, rotation vectors, dihedral angles and thermal parameters. Normally, the group coordinates are chosen so that all angular parameters, R, and ψ are initially set to zero.

The group derivatives with respect to the positional parameters are obtained by differentiating eqn. (1) and application of the chain rule[12,13]. Within any group, a subgroup of atoms can be constrained to

have the same temperature factor. The subgroup derivatives are solely a function of F. The least-squares normal equations follow directly[12]. In order to reduce computer time the program was written with space group specific subroutines for the calculation of structure factors and derivatives. Takusagawa[18] has written a space group independent subroutine for CORELS to do these calculations, which is especially useful for the refinement of high symmetry space groups. Recently, using this feature, we have begun the refinement of the cubic form of yeast $tRNA^{Phe}$ in space group $I4_132$(ref. 19).

3. IMPLEMENTATION OF CORELS

In CORELS a structure is explicitly described by coordinates in Å of standard ideal groups in arbitrary initial orientations, i.e. the ATOMS file. The coordinates of the ATOMS file for each group are rotated and translated as rigid bodies and any number of dihedral angles within them are rotated by values found in the PARAMETERS file. In addition a TARGETS file containing the fractional coordinates of the initial model can be used for model building in a way similar to the guide coordinates of other model building programs[1,5]. This option was used in the model building of the way in which DNA might be smoothly deformed in wrapping around the histone core in a nucleosome[20,21]. An alternative use of the TARGETS file is simply as a reference, as to how far the structure has moved during the reciprocal space refinement, with no attempt to restrain the structure to these targets.

Specifically, each file for a particular CORELS group contains the following information:

A) ATOMS file -
   1) Atom names
   2) Atomic number (0 for dummy atoms - see fig. 1c)
   3) x,y,z Angstrom coordinates of each atom in a standard orientation. (Usually the center of mass of the group is at 0,0,0.)

B) PARAMETERS file -
   1) Name of CORELS group (e.g. amino acid name & number).
   2) Number of atoms, total number of parameters and number of temperature factors for the group.
   3) x,y,z fractional coordinates of the origin of the group (i.e. translation vector of group to its position in the unit cell).
   4) Rigid body rotation angles of the group.
   5) Any number of lists of atoms to be constrained to a single temperature factor and its respective value.
   6) Any number of dihedral angles. (A dihedral angle is defined by a pair of atoms specifying a rotation vector, together with a list of atoms which are to be rotated.)

C) TARGETS file -
   1) Atom names
   2) x,y,z fractional coordinates of each atom.
   3) Weight - used for model building option.

Any two atoms in the same or different groups can be restrained by harmonic spring-like connections. These are defined by a RESTRAINTS file which gives the pair of atoms which are to be restrained and a pointer to a DICTIONARY file which contains specific distances and weights for this kind of restraint.

In order to increase the radius of convergence, during the course of the refinement of a structure, we have found from several crystal structures, that it is best to start with rigid bodies, relatively low resolution data and initially vary the least number of degrees of freedom.

4. APPLICATION OF CORELS TO SPECIFIC PROBLEMS

In this section we describe briefly two different examples of the use of CORELS. The first, yeast $tRNA_f^{Met}$, is an application at extremely low and intermediate resolution data where CORELS helped to refine a nucleic acid structure[22,23,24]. The second, demetallized concanavalin A, is an example of the refinement of a protein structure at higher resolution[25].

4.1 Yeast initiator tRNA

The initial model of the crystal structure of yeast initiator $tRNA_f^{Met}$ was based on the interpretation of a rather noisy electron density map at 4.5 Å resolution prepared by the method of multiple isomorphous replacement (MIR) and augmented by direct methods[22,23,26]. (The poor quality of this map was subsequently found to be due primarily to a misassignment of the z-coordinate of one heavy atom deri-

vative[27], however, this was noticed only after the refinement of the entire structure.) The interpretation of this MIR map was aided by locating four covalently bonded heavy-atom markers (found in extreme positions in the structure) by MIR-phased difference Fourier maps, which helped to place the structure in the unit cell. The use of the PACKGRAF computer program[28] as adapted for a static TEKTRONIX computer graphics terminal by Podjarny & Honig[29] enabled us to manipulate large segments of the structure (e.g. helical arms) as rigid bodies.

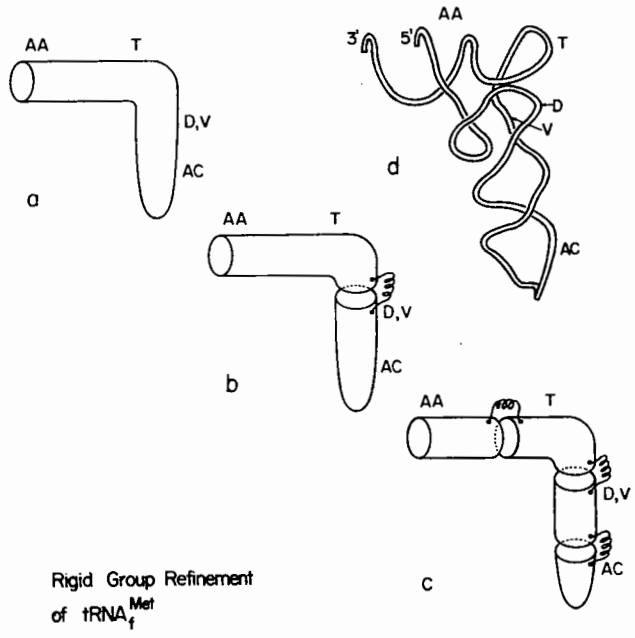


Fig. 3 Schematic representation of how the structure of tRNA$_f^{Met}$ was successively refined as rigid groups at low resolution. (a) The entire "L" shaped structure was treated as a single rigid group (6 degrees of freedom, i.e. 3 rotational and 3 translational). (b) The structure was divided into the two major domains, i.e. one consisting of the AA-T arms and the other consisting of the D-V-AC arms (12 degrees of freedom). The spring-like connections between groups represent chemical restraints used to maintain reasonable bond lengths and bond angles at the arbitrary division between groups. (c) The structure was divided into four rigid groups, i.e. the AA stem, T arm, D arm & V loop and the AC arm (24 degrees of freedom). (d) For comparison a schematic trace of the backbone of tRNA^{Phe} in approximately the same orientation is shown.

The model of tRNA$_f^{Met}$ built on the computer graphics system was similar to that of the structure of tRNA^{Phe}, although due to limitations in the model building procedure it had somewhat poorer stereochemistry. To correct this we decided to continue with a more reasonable stereochemical model. Specifically the molecular structure of tRNA^{Phe} (ref. 30) was fitted by a least-squares technique[31] matching up the phosphorus coordinates to those of the tRNA$_f^{Met}$ as obtainerd from the computer graphics fit. The RMS distance between the 75 phosphorus atoms in common for the two coordinate sets was 5.4 Å. However the R-factor,

$$R = \sum_h \left| |F_{o,h}| - |F_{c,h}| \right| / \sum_h |F_{o,h}| \qquad (6)$$

was essentially random except at very low resolution, being 58% for the range of data 12.5-20 Å, vs. a random R-factor of 65% expected for the particular combination of centric and non-centric reflections in the space group $P6_422$.

The refinement proceeded in steps, initially with the smallest possible number of degrees of freedom (the entire structure treated as one rigid group) and only the lowest resolution data (12.5-20 Å) in order to maximize the radius of convergence. After the refinement first converged in each previous step, the number of degrees of freedom was successively increased by dividing the structure into two and four rigid groups, corresponding to the different domains of the structure (see fig. 3). This procedure caused the R-factor to drop from 58% to 33% for the 12.5-20 Å data (see fig. 4). More data were then included and the structure refined as four rigid groups till at 6 Å resolution the R-factor was 42% (see fig. 5).

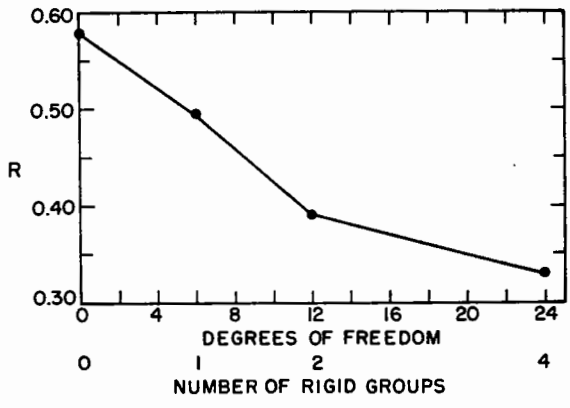


Fig. 4 R-factor for the low resolution (12.5-20 Å) refinement of tRNA$_f^{Met}$ as the structure was successively divided into 1, 2 and 4 rigid groups.

The most striking result of this procedure was a shift of the center of mass of the whole structure by almost 5 Å (see fig. 6). This is a large movement and could be obtained only because it is small compared with the resolution used in the initial stages of the refinement (12.5-20 Å).

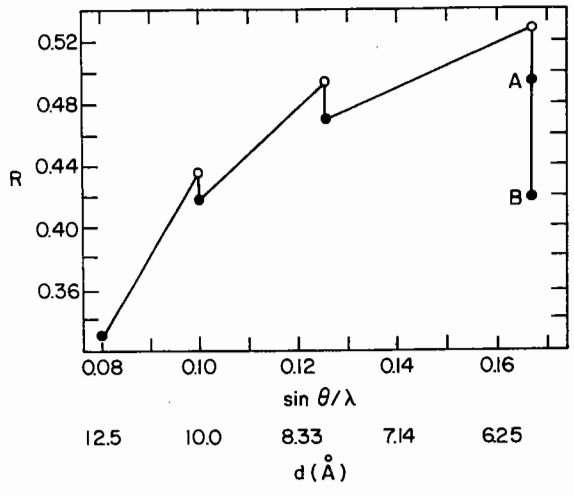


Fig. 5 R-factor for the intermediate resolution refinement of $tRNA_f^{Met}$. Each open circle corresponds to the R-factor as higher resolution data were included and the closed circles to the R-factor at the same resolution after the structure was refined to convergence as four rigid groups. The vertical drop at 6 Å resolution corresponds to the constrained temperature factor refinement of each of the 4 rigid groups (i.e. an additional 4 degrees of freedom).

At this stage more data were used (first to 4.5 Å and then to 4 Å resolution) and the rigid group constraints were relaxed, such that for the loop regions the groups were separate phosphates and nucleosides while for the double-helical stem regions they were phosphates and base-paired nucleosides. This is the same kind of scheme that was used in the $tRNA^{Phe}$ refinement[30] and is shown in fig. 7. Restraints were imposed between the constrained groups to maintain reasonable stereochemistry as well as to prevent unacceptably close contacts between nonbonded atoms. No restraints were imposed on the tertiary base-base interactions (until the very final stages of refinement after the least-squares had converged), i.e. each residue not in a helical stem was refined independently. At virtually all stages of the CORELS refinement we examined difference electron density maps to be certain that the structure had not fallen into a false minimum. Based on these maps we refitted the structure on the static computer graphics system. This was especially important in the single-stranded 3' end where the largest differences in conformation from $tRNA^{Phe}$ are found. Examples of two such maps at an early stage and near the end of the refinement are shown in fig. 8.

The R-factor at the present stage of refinement is 26% based on all 3302 reflections to 4 Å resolution (about 75% of the theoretically possible total number), with a total of 1096 degrees of freedom (rotation, translation, torsion angles and temperature factors for the 129 groups, and a single overall scale factor) and 2033 chemical restraints between the groups. Thus based solely on X-ray data, i.e. excluding the chemical restraints, the ratio of the number of observables to degrees of freedom is about 3.0. This is clearly an underestimate, as it is difficult to meaningfully compare X-ray reflections and chemical restraints, both of which are observables.

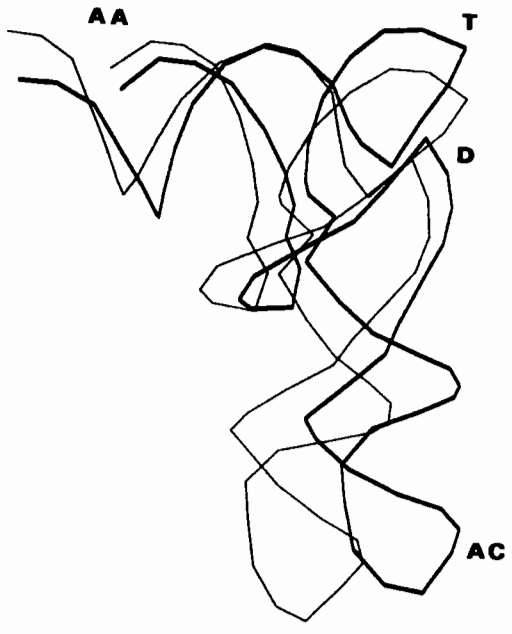


Fig. 6 Change in the $tRNA_f^{Met}$ structure after low and intermediate resolution refinement as four rigid groups (see fig. 3c). The light trace represents the starting coordinates (R = 58%, 12.5-20 Å) and the heavy trace the structure after the initial group refinement (R = 42%, 6-20 Å).

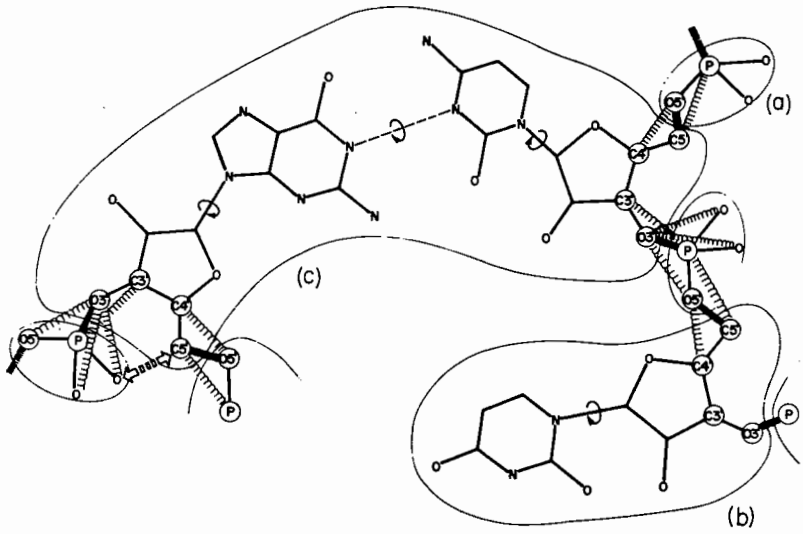


Fig. 7 The three different kinds of constrained groups used in the higher resolution refinement of $tRNA_f^{Met}$ are shown. (a) The smallest is a phosphate group with 6 positional degrees of freedom and a single temperature factor (a total of 7 degrees of freeedom). (b) The next largest group is a nucleoside with torsional flexibility of the base relative to its ribose, as well as separate subgroup temperature factors for each moiety, i.e. ribose and base (a total of 9 degrees of freedom). (c) The largest group is a constrained base-pair (for the double-helical stem regions). In addition to the torsional flexibility of each ribose relative to its respective base, one of the nucleosides is permitted to twist as a unit about a vector (shown as a dashed line) between the N1 of the purine to the N3 of the pyrimidine (for a total of 13 degrees of freedom). Restrained distances corresponding to bond lengths are shown by dotted lines, distances corresponding to bond angles by dashed lines, while non-bonded contacts with a double-headed arrow.

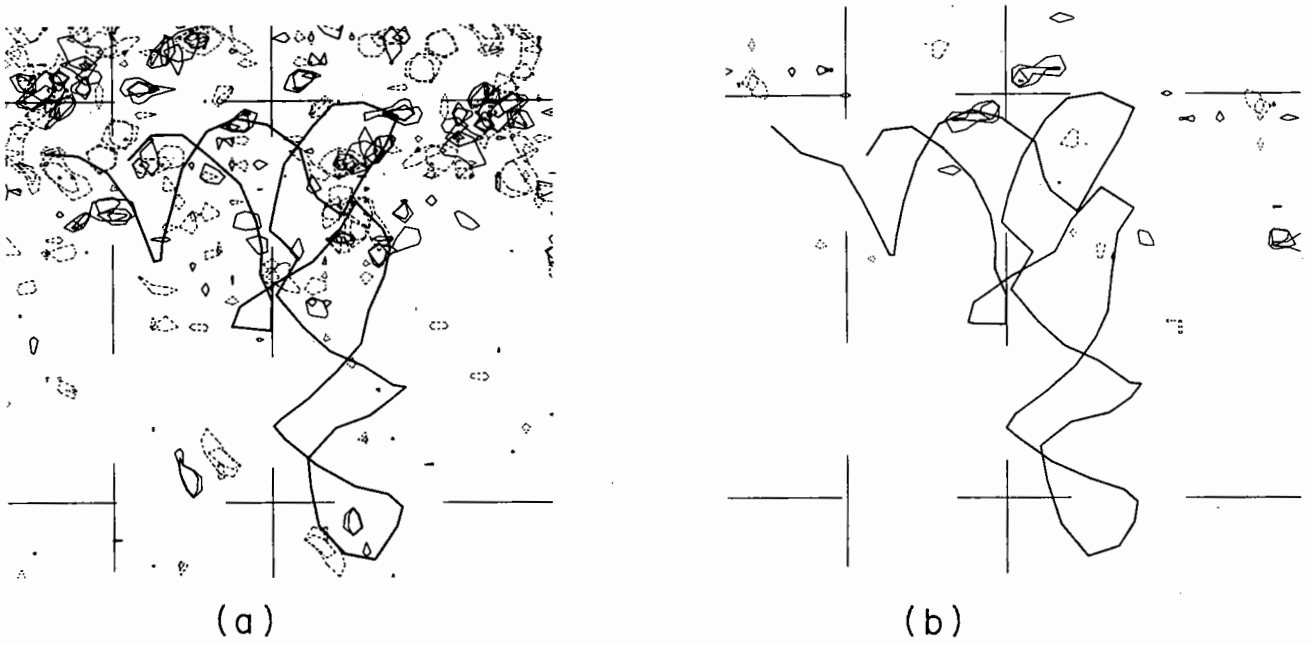


Fig. 8 Difference electron density maps at two different stages in the refinement of the crystal structure of $tRNA_f^{Met}$ together with a superimposed trace of the backbone. (a) Difference map at 4.5-20 Å resolution, R = 40.5% where only residues in the four double helical stems were allowed to vary freely during the refinement (see fig. 7c), but the loop regions were constrained to the $tRNA^{Phe}$ conformation. There is a large amount of unaccounted for electron density near the CCA end, in the region of the T & D loops as well as in the middle of the AC loop. (b) The same resolution map (contoured as in (a)) at the end of the refinement treating the entire structure as composed of the groups shown in fig. 7, R = 27%. Here the map as a whole is much cleaner than in (a) showing how the conformation of the structure changed during the refinement to fit better the observed X-ray data.

4.2 Demetallized Concanavalin A

Concanavalin A is a saccharide-binding protein of the Jack bean[32]. Its saccharide-binding properties in solution have been shown to depend on the presence of two metal ions in the protein[33,34]. The details of the structural changes in the protein on successive occupation of the two metal-binding sites can be elucidated by comparisons between X-ray structures of metal-free concanavalin A, its transition metal complex, and the native protein containing both metals. Of these, the X-ray structure of the native form has been described in some detail[35,36,37].

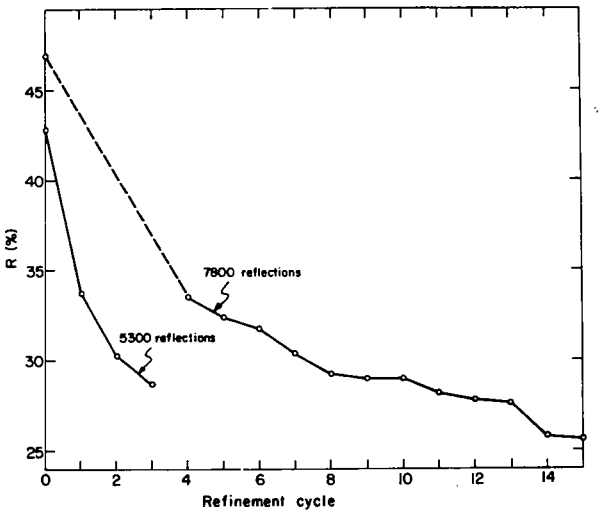


Fig. 9 Decrease of the R-factor during the reciprocal-space refinement of demetallized concanavalin A. In the first 3 cycles a partial data set of 5300 reflections was used. From cycle no. 4 the entire data set of 7800 reflections was used. The initial R-factor for the full data set was 47.8% as indicated by the dashed line. For more details see Table 1 of ref. 25

We determined the crystal structure of demetallized concanavalin A, at a resolution of 3.2 Å by the molecular replacement method using the known structure of native concanavalin A[25]. After the orientation and translation search, the R factor for data to 3.2 Å was 0.47. As might be expected from such a starting model, the electron density maps were unclear and a detailed interpretation would be unreliable (see fig. 2 of ref. 25). To overcome this difficulty we refined the structure away from the starting model by treating each amino acid as an independent unit with specific spatial restraints to neighbouring residues in the chain in order to maintain acceptable stereochemistry (see fig. 1c). Within each unit, dihedral and main-chain bond angles as well as the composite temperature factors of groups of atoms, were refined. All restraints were implemented by means of flexible "springs" between pairs of atoms. By suitably varying the "stiffness" of these springs it was possible to make the distances between pairs of atoms as close to standard lengths as desired.

The strategy we used in refining the structure was initially to vary only a minimal number of parameters and gradually to allow more to vary in the following order.

1) Rigid-body movements (rotation and translation) of individual amino acids, with restraints between adjacent residues in the sequence to maintain proper inter-residue geometry.
2) Variation of the side-chain dihedral angles.
3) Introduction of "repulsive" springs between non-bonded atoms to minimize repulsive interaction due to close van der Waals' contacts.
4) Refinement of two subgroup temperature factors per residue, one for main-chain atoms (N, C_α, C_β, C', and O) and one for the remaining atoms.

This refinement sequence was designed to allow the biggest movements of large groups of atoms to occur first. The asymmetric unit refined consisted of 3612 non-hydrogen and non-metal atoms of the dimer, and the behaviour of equivalent residues in the 2 monomers provided a check of the accuracy of the results.

Figures 9 and 10 summarize the refinement steps. As we let more parameters vary, the ratio of observations to parameters decreased from 2.4 to 1.9. The refinement remained reasonably well over-determined because of the approximately 6000 distance restraints. It should also be noted that when non-bonded restraints were introduced in refinement cycle 10, the R factor did not increase. The average change in phase angles was 45° (r.m.s. 60°), which is about the same as has been found in other protein structures during the course of refinement[38]. The final difference electron density maps for the refined structure were much cleaner than those at the initial stages of refinement (cf. fig. 7 and fig. 3 in ref. 25) and allowed us to compare the structure of native and demetallized concanavalin A in detail.

One very interesting conclusion that emerged from this work was that it was possible to rule out a cis/trans isomerization of the peptide bond between ala 207 and ala 208 upon the removal of the metal

ions when going from the native to demetallized structure. Such a model was proposed[39] to explain a kinetic analysis of NMR data of a 22 kcal/mole energy barrier in the activation of the demetallized protein[40].

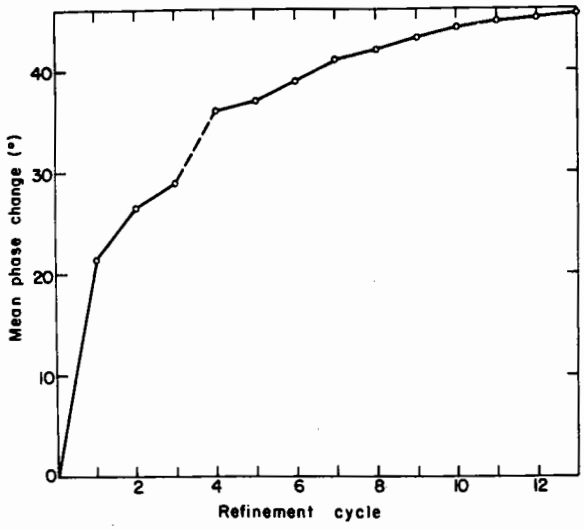


Fig 10 Mean phase change of the calculated structure factor from those of the starting model, during the reciprocal-space refinement of demetallized concanavalin A. The dashed line indicates the incorporation of the entire data set.

5. CONCLUSIONS

The use of a constrained-restrained least-squares procedure has proven to be extremely useful in refining macromolecular structures, especially when the initial model has severe errors. This method inherently has many fewer degrees of freedom than restrained refinement procedures and therefore is applicable at extremely low resolution with a very large radius of convergence. It is superior to a strictly constrained procedure as the restraints between groups reduce the influence of linked neighbors on positional shifts. The program is suitable for either structure factor least-squares refinemnt (for any space group) with subsidiary distance restraints and/or model building to guide coordinates.

ACKNOWLEDGEMENTS

I wish to thank Drs Michal Harel, John Moult, Osnat Herzberg and Ada Yonath for reading this manuscript and giving useful comments, and Donna Goldberg for help in its preparation.

REFERENCES

1. J. Hermans and J. McQueen, Acta Cryst. A30, (1974) 730.

2. J.H. Konnert, Acta Cryst. A32, (1976) 614, W.A. Hendrickson and J. Konnert in "Biomolecular Structure: Conformation, Function and Evolution" Vol. 1, (ed. R. Srinavasan), pp. 43-57 (Pergamon, Oxford, 1979).

3. A. Jack and M. Levitt, Acta Cryst. A34, (1978) 931.

4. E.J. Dodson, N.W. Issacs and J.S. Rollett, Acta Cryst. A32, (1976) 311.

5. R. Diamond, Acta Cryst. 21, (1966) 253.

6. R. Diamond, Acta Cryst. A27, (1971) 436.

7. S. Fitzwater and H.A. Scheraga, Acta Cryst. A36, (1980) 211.

8. S. Arnott, S.D. Dover and A.J. Wonacott, Acta Cryst. B25, (1969) 2192.

9. F.M. Richards, J. Mol. Biol. 37, (1968) 225.

10. D. Tsernoglou, G.A. Petsko, J.E. McQueen and J. Hermans, Science 197, (1977) 1378.

11. T.A. Jones, J. Appl. Cryst. 11, (1978) 268.

12. J.L. Sussman, S.R. Holbrook, G.M. Church and S.H. Kim, Acta Cryst. A33, (1977) 800.

13. R. Doedens in "Crystallographic Computing", pp. 198-200, (Copenhagen: Munksgaard, 1970).

14. L.G. Hoard and C.E. Nordman, Acta Cryst. A35, (1979) 1010.

15. M. Marquart, J. Deisenhofer, R. Huber and W. Palm, J. Mol. Biol. 141 (1980) 369.

16. R. Haser and J.L. Sussman, to be submitted

17. J. Waser, Acta Cryst. 16, (1963) 1091.

18. F. Takusagawa, private communication.

19. J. Nachman, J.L. Sussman, R.W. Warrant and S.H. Kim, Abstracts, XIIth IUCr Congress (1981), Ottawa, Canada (In Press).

20. J.L. Sussman and E.N. Trifonov, Proc. Natl. Acad. Sci. (USA) 75, (1978) 103.

21. E.N. Trifonov and J.L. Sussman, in "Molecular Mechanism of Biological Recognition", (ed. M. Balaban) pp. 227-232 (Elsevier/North-Holland Biomedical Press, 1979).

22. R.W. Schevitz, A. D. Podjarny, N. Krishnamachari, J.J. Hughes, P.B. Sigler and J.L. Sussman, Nature 278, (1979) 188.

23. J.L. Sussman, A.D. Podjarny, R.W. Schevitz and P.B. Sigler, in "Structural Aspects of Recognition and Assembly in Biological Macromolecules", (ed M. Balaban), pp. 597-614 (BALABAN International Science Services, Rehovot and Philadelphia, 1981).

24. J.L. Sussman and A.D. Podjarny, submitted to Acta Cryst. A. (1981).

25. M. Shoham, A. Yonath, J.L. Sussman, J. Moult, W. Traub and A. J. Kalb (Gilboa), J. Mol. Biol. 131 (1979) 137.

26. A.D. Podjarny, R.W. Schevitz, J.J. Hughes, M. Zwick and P.B. Sigler, Acta Cryst. A (In Press 1980).

27. R.W. Schevitz, A.D. Podjarny, M. Zwick, J.J. Hughes and P.B. Sigler, Acta Cryst. A (In Press 1980).

28. L. Katz and C. Levinthal, Ann. Rev. Biophys. Bioeng. 1, (1972) 465.

29. A.D. Podjarny, Ph. D. Thesis, Weizmann Institute of Science, Rehovot, Israel (1976).

30. J.L. Sussman, S.R. Holbrook, R.W. Warrant, G.M. Church and S.H. Kim, J. Mol. Biol. 122, (1978) 607.

31. S.C. Nyburg, Acta Cryst. B30, (1974) 251.

32. J.B. Sumner, J. Biol. Chem. 37, (1919) 137.

33. J.B. Sumner and S.F. Howell, J. Biol. Chem 115, (1936) 583.

34. J. Yariv, A.J. Kalb and A. Levitzki, Biochim. Biophys. Acta 165, (1968) 303.

35. G.M. Edelman, B.A. Cunningham, G.N. Reeke Jr., J.W. Becker, M.J. Waxdal and J.L. Wang, Proc. Nat. Acad. Sci. (USA) 69, (1972) 2580.

36. K.D. Hardman and C.F. Ainsworth, Biochemistry 11, (1972) 4910.

37. J.W. Becker, G.N. Reeke Jr., B.A. Cunningham and G.M. Edelman, Nature (London) 259, (1976) 406.

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

39. G.N. Reeke Jr., J.W. Becker and G.M. Edelman, Proc. Nat. Acad. Sci. (USA) 75, (1978) 2286.

40. R.D. Brown, C.F: Brewer and S.H. Koening, Biochemistry 16, (1977) 3883.

NEW RESULTS ON FAST FOURIER LEAST-SQUARES REFINEMENT TECHNIQUE

by

R.C. AGARWAL*

Centre for Applied Research in Electronics, Indian Institute of Technology Hauz Khas, New Delhi 110016, India

1. INTRODUCTION

The conventional least-squares refinement method is prohibitively expensive for large structures such as proteins. Agarwal[1] presented a least-squares refinement technique where most of the calculations were done using fast Fourier transforms. The algorithm is extremely fast and the computing required is proportional to N log N, where N is the number of reflections. Although the method is most useful for large structures, it is applicable to small structures also because of its large radius of convergence (0.75 Å) and reduced computational requirement. The method has been used to refine several proteins; among others insulin[2] and actinidin[3]. In this paper some improvements on the method in calculating derivatives and the normal matrix are presented.

2. METHOD

In the least-squares refinement of atomic parameters the function minimised is

$$P = \frac{1}{2} \sum_{hkl} W(hkl) \cdot \{|F_c(hkl)| - |F_{obs}(hkl)|\}^2 \quad (1)$$

where W(hkl) is a weighting function. This function is to be minimised with respect to atomic parameters. The corrections to the parameters are obtained from the matrix equation

$$\Delta p = - H^{-1} G \quad (2)$$

where Δp_i is the correction to be applied to the i^{th} parameter. H^{-1} is the inverse of the normal matrix whose general term is

$$H_{ij} = \sum_{s=1}^{N} W(s) \frac{\partial |F_c(s)|}{\partial p_i} \frac{\partial |F_c(s)|}{\partial p_j} \quad (3)$$

where N is the number of reflections and W(s) is the weighting function. G is the gradient vector (derivatives) of general form

$$G_i = \sum_{s=1}^{N} W(s)\, E(s) \frac{\partial |F_c(s)|}{\partial p_i} \quad (4)$$

where $E(s) = |F_c(s)| - |F_{obs}(s)|$. The size of the normal matrix is MXM where M is the number of parameters and the length of the gradient vector is M. Direct calculation of the gradient vector is proportional to NM and that of the normal matrix is proportional to NM^2.

There are three major computational steps in the refinement procedure. These are calculation of structure factors, the gradient vector, the normal matrix and its inverse. A procedure using FFT (Fast Fourier Transform) in all these steps has been given by Agarwal[1]. In this paper, a faster method of computing the gradient and normal matrix will be presented. Fast computation of structure factors will not be discussed in this paper. For this, the reader is referred to the earlier paper[1].

3. CALCULATION OF THE GRADIENT VECTOR

Agarwal[1] has derived the following expression for the gradient vector with respect to the x co-ordinate of the m^{th} atom.

$$G(x_m) = \sum_S g_m(s)\, (-i2\pi h)\, W(s)\, E(s) \exp(i\phi(s)) \exp(-i2\pi s.r_m) \quad (5)$$

where

$g_m(s) = f_m(s) \exp(-B_m s^2/4)$ = contribution of m^{th} atom to structure factors

$f_m(s)$ = form factor of m^{th} atom

B_m = isotropic thermal parameter of m^{th} atom

$|s| = 2 \sin\theta/\lambda$

$s.r_m = hx_m + ky_m + lz_m$

*This work was done while the author was at IBM, T.J. Watson Research Center, Yorktown Heights, NY 10598, USA, as a summer faculty during summer of 1980.

r_m = (x_m, y_m, z_m) - fractional cell co-ordinates of m^{th} atom

$\phi(s)$ = phase of $F_c(hkl)$

Similar expressions hold for $G(y_m)$, $G(z_m)$ and $G(B_m)$ with the term $(-i2\pi h)$ replaced by $(-i2\pi k)$, $(-i2\pi l)$, and $(-s^2/4)$ respectively.

The method suggested by Agarwal[1] required calculation of 4 FFTs for gradient calculation. Lifchitz[4] has suggested a method which requires only one FFT. In this paper, we elaborate on his method.

Let $\rho_m(r-r_m)$ be the atomic electron density of the m^{th} atom centred at $r=r_m$. Then $\rho_m(r)$ is the same electron density but centred at $r=0$ and $g_m(s)$ the fourier transform (FT) of $\rho_m(r)$. It can be easily shown that $g_m(s)(-i2\pi h)$ is the FT of $\partial\rho_m(r)/\partial x$, x derivative of the atomic electron density. To simplify equations, we introduce the following notations.

$$g'_m(s) = g_m(s)\ (-i2\pi h) \qquad (6)$$

$$\rho'_m(r) = \partial\rho_m(r)/\partial x \qquad (7)$$

Now $G(x_m)$ can be re-written as

$$G(x_m) = \sum_s D(s)\ g'_m(s)\ \exp\ (-i2\pi\ s.r_m) \qquad (8)$$

where $D(s) = W(s)\ E(s)\ \exp\ (i\ \phi(s))$. $G(x_m)$ then is the FT of the product of two functions D(s) and $g_m(s)$ evaluated at $r=r_m$ (position of the m^{th} atom). According to the convolution theorem, multiplication in reciprocal space is equivalent to convolution in real space. The FT of D(s) is simply the weighted difference density map denoted by d(r) and the FT of $g_m(s)$ is $\rho_m(r)$, the x derivative of atomic electron density of m^{th} atom, as explained above. The gradient then is computed by the summation

$$G(x_m) = \sum_r d(r)\ .\ \rho'_m(r-r_m) \qquad (9)$$

where the summation need only be carried out over the extent of the m^{th} atom. Equations for $G(y_m)$, $G(z_m)$ and $G(B_m)$ are similar with $\rho_m(r-r_m)$ referring to y, z and B_m derivatives of $\rho_m(r-r_m)$, respectively. Computation of d(r), the difference density map, requires only one FFT and is common to all derivatives. For this step, standard Fourier routines for that space group can be used without any modifications. This is in contrast to earlier work[1] where for gradient calculation Fourier routines had to be modified. This step is followed by summation of eqn.(9) for each atom. As this step is carried out, all derivatives for that atom (e.g. $G(x_m)$, $G(y_m)$, $G(z_m)$, $G(B_m)$) are computed simultaneously by using the appropriate derivative of $\rho_m(r)$ in eqn.(9).

When atomic electron density is modelled as a sum of Gaussian atoms, closed form expressions for its derivatives can be obtained in terms of various atomic parameters and distances between grid points and the centre of the atom. This simplifies gradient calculation. Another advantage of this approach is that gradient calculation becomes space group general. The only space group specific routine is the usual Fourier routine for that space group.

Since the $\rho'_m(r)$ curve is broader than $\rho_m(r)$, a larger atomic radius is required for summation in eqn.(9), but, this need not increase computation time. By judicious choice of grid spacing, BADD etc., we can reduce the calculation time. Also we are willing to tolerate more error in derivatives calculation as compared to structure factor calculation. Our experience with this method indicates that the calculation time for <u>all</u> derivatives is about the same as structure factor calculation time alone.

4. CALCULATION OF THE NORMAL MATRIX

4.1 <u>The Diagonal Terms</u>

The following expression for the normal matrix term $H(x_m,x_n)$, corresponding to interaction between x_m and x_n has been derived[1].

$$H(x_m,x_n) = H_1(x_m,x_n) + H_2(x_m,x_n) \qquad (10)$$

As shown earlier[1], the $H_2(x_m,x_n)$ term can be neglected, therefore, we shall restrict our discussion to

$H_1(x_m,x_n)$ given below.

$$H_1(x_m,x_n) = \sum_s \frac{1}{2}\ g_m(s)\ g_n(s)\ (4\pi^2h^2)\ W(s) \qquad (11)$$

$$\exp(i2\pi\ s.(r_m-r_n))$$

For the particular case of diagonal terms (m=n), this simplifies to

$$H_1(x_m,x_m) = \sum_S 2\ \pi^2 h^2\ g_m^2(s)\ W(s) \qquad (12)$$

If $g_m(s)$ is modelled as a sum of two Gaussian terms, $g^2(s)$ becomes a sum of three Gaussian terms as indicated below:

$$g_m^2(s) = \sum_{i=1}^{3} c_i \exp(-b_i s^2/4) \qquad (13)$$

The expressions for b_i's and c_i's are given by Agarwal[1]. Combining eqns.(12) and (13) we obtain

$$H_1(x_m,x_m) = \sum_{i=1}^{3} c_i \sum_S 2\ \pi^2\ h^2\ W(s)\ \exp(-b_i\ s^2/4) \qquad (14)$$

Now, we shall give an efficient and space group general method of calculating the diagonal terms.

The summation in eqn.(14) is to be carried out over all reflections used in refinement. In most cases this set is spherical with certain radius s_{max}, denoting the limits of the data. For further simplifications we can assume that all reflections within this sphere are used. In eqn.(14), all terms except h^2 are spherically symmetric in s (e.g. they have the same value on the surface of a sphere of radius $|s|$). Furthermore, the average value of h^2 on this sphere is proportional to $a^2s^2/3$, where a is the dimension of the unit cell along the x-axis. With this substitution, eqn.(14) can be made spherically symmetric. The next step is to convert the three-dimensional summation of eqn.(14) to a one-dimensional integration over s. This is not a bad approximation for large structures. The final result is as follows:

$$H_1(x_m,x_m) = V.a^2 \sum_{i=1}^{3} c_i \int_0^{s_{max}} (8/3)\pi^3\ s^4\ W(s)\ \exp(-b_i\ s^2/4)\ ds \qquad (15)$$

Where V is the volume of the unit cell. This is the space group general expression for the diagonal terms. The integral in eqn.(15) is evaluated for several dummy values of b_i. For a particular atom, the value is obtained via interpolation. The integral itself can be evaluated by one-dimensional summation over 100 or so values of s. Since, this integral is a rapidly decreasing function of b_i, it is recommended that for low b_i values, the dummy values be chosen with a small spacing and for large b_i's the spacing can be increased to reduce computation. Perhaps a spacing of 1 for b_i values up to 20 and a spacing of 4 for higher b_i values can be used.

Diagonal terms involving other co-ordinates of the same atom can be obtained from $H_1(x_m,x_m)$ using eqn.(64) of the earlier paper[1]. For B diagonal terms the equation is

$$H_1(B_m,B_m) = V \sum_{i=1}^{3} c_i \int_0^{s_{max}} (\pi/8)s^6 W(s)\ \exp(-b_i s^2/4)\ ds \qquad (16)$$

This can be calculated similarly to eqn.(15).

4.2 Off-diagonal Terms

We shall confine our discussion to orthogonal co-ordinates. For space groups which are not orthorhombic, we can still use orthogonal co-ordinates by an appropriate change of variable from fractional to orthogonal co-ordinates. While using off-diagonal terms, to minimise their number, use of orthogonal co-ordinates is recommended. However, FTs are best computed using fractional co-ordinates. Therefore, we recommend that structure factors and derivatives be calculated using fractional co-ordinates. By using the change of variable equations, the derivatives for orthogonal co-ordinates can be obtained from derivatives for fractional co-ordinates.

Let us first discuss calculation of off-diagonal terms involving the same co-ordinate e.g. $H_1(x_m,x_n)$. For the two Gaussian approximation, eqn.(11) can be re-written as

$$H_1(x_m,x_n) = \sum_{i=1}^{3} c_i \sum_S 2\ \pi^2\ h^2\ W(s)\exp(-b_i s^2/4)\exp(i2\pi s.(r_n-r_m)) \qquad (17)$$

In the inner summation, if the h^2 term was absent it would be a spherically symmetric summation and for the given weighting function and s_{max} its value would be a function of only two parameters e.g. b_i and the distance between two atoms ($|r_n-r_m|$). Because of the h^2 term, this summation would be a function of three parameters e.g. b_i, distance between two atoms along the x-axis (d_x), and the distance between two atoms in the yz plane (d_{yz}). One could possibly pre-compute a three-dimensional table

for a set of b_i, d_x, and d_{yz} values. This table need not be very large. As discussed earlier[1], the only significant off-diagonal terms are the ones involving bonded atoms. Therefore the range of d_x and d_{yz} values required is restricted to interatomic distances. Perhaps a total of a thousand entries or so may be sufficient. For in-between values, interpolation can be used. Calculation of off-diagonal terms need not be very accurate. Table entries may be calculated using integration instead of summation. If orthogonal co-ordinates in Ångstroms are used, dependencies on space group and size of the unit cell are also eliminated. Thus for a given weighting function and s_{max}, the same table could be used for all structures.

Now, we give another approach which could reduce the look-up table size. This is restricted to unity weighting function (w(s)=1). Omitting some constant multipliers, the inner summation of eqn.(17) can be written as the following integration, where the integration is carried out within a sphere of radius s_{max}.

$$f(b_i,d_x,d_y,d_z) = \iiint h^2 \exp(-b_i s^2/4) \exp(i2\pi(hd_x + kd_y + ld_z))ds \qquad (18)$$

where

$$d_x = |x_n - x_m|$$
$$d_y = |y_n - y_m|$$
$$d_z = |z_n - z_m|$$

and $$s^2 = (h/a)^2 + (k/b)^2 + (l/c)^2$$

In the integration of eqn.(18), hkl are to be treated as continuous variables. Alternatively, in eqn.(18), if we integrate within a cylinder along the h-axis, the integral can be written as a product of two integrals as shown below:

$$f(b_i,d_x,d_y,d_z) = \left[\int h^2 \exp(-b_i h^2/4a^2) \exp(i2\pi hd_x)dh\right] X \left[\int \exp(-b_i((k/b)^2 + (l/c)^2)/4) \exp(i2\pi(kd_y + ld_z))\, dkdl\right] = f_1(b_i,d_x)\, f_2(b_i,d_y,d_z) \qquad (19)$$

It can be further noted that the second integral is circularly symmetric. Therefore f_2 depends only on d_{yz}, the distance between two atoms in the yz plane. Thus, eqn.(19) can be re-written as

$$f(b_i,d_x,d_y,d_z) = f_1(b_i,d_x) \cdot f_2(b_i,d_{yz}) \qquad (20)$$

Functions f_1 and f_2 are two-dimensional tables. Therefore, they require less computation and storage. In this approach, we have approximated a sphere by a cylinder. In doing so, we have neglected some observed reflections and added some unobserved reflections. Since most of these reflections correspond to large s values, their effect on calculation should not be much. This is particularly true for atoms having large B values and structures having large s_{max} (high resolution data).

Next consider off-diagonal terms involving different co-ordinates of atoms, i.e. $H_l(y_m,z_n)$. These are somewhat more complicated. They are a function of b_i, d_x, d_{yz}, as well as the angle which d_{yz} makes with the y (or z) axis. The maximum magnitude is obtained when the angle is ± 45° corresponding to $|y_m - y_n| = |z_m - z_n|$. On the other hand if $y_m = y_n$ or $z_m = z_n$, the value is zero. This dependence on angle could perhaps be expressed as a simple function. In any case, these can also be calculated using table look-up.

Alternatively, if the data set in reciprocal space is assumed to be cubic instead of spherical, any off-diagonal term can be written as a product of three functions.

$$f(b_i,d_x,d_y,d_z) = f_1(b_i,d_x) \cdot f_2(b_i,d_y) \cdot f_3(b_i,d_z) \qquad (21)$$

Thus all tables become two-dimensional tables. As explained above, this assumption is valid for two situations, (a) when data is available to a high resolution, or (b) when B values are high.

Computation of $H_l(B_m,B_n)$ is very simple. For this the expression is spherically symmetric. Thus even for the normal spherical data set, $H_l(B_m,B_n)$ can be computed from a two-dimensional table being a function of only b_i and the distance between two atoms, because of spherical symmetry.

Off-diagonal terms involving co-ordinates and thermal parameters are perhaps of no interest, because, usually they are not refined simultaneously.

The final step in the refinement procedure is the inversion of the normal matrix. For this, some sparse matrix inversion algorithm could be used. Perhaps it could be inverted iteratively using the conjugate gradient method whereby at every iteration you improve on the estimate of the inverse.

With efficient calculation of the normal matrix and its inverse, it may be feasible to use it for large structures also. This is likely to improve refinement of atoms having large thermal parameters, because for such atoms, interatomic interactions are significant. It may also improve refinement of structures with limited data.

REFERENCES

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

2. N.W. Isaacs and R.C. Agarwal, Acta Cryst. A34, (1978) 782-791.

3. E.N. Baker and E.J. Dodson, Acta Cryst. A36, (1980) 559-572.

4. A. Lifchitz, Private Communication.

BLOCK DIAGONAL LEAST SQUARES REFINEMENT USING FAST FOURIER TECHNIQUES

by

E.J. Dodson
Department of Chemistry, University of York, Heslington, York YO1 5DD

1. INTRODUCTION

At York we have had a good deal of experience in using the fast fourier transform least squares refinement. We usually minimise the atomic parameter shifts with respect to the x-ray observational equations, and then fit our model within the expected geometric restraints for bond lengths, angles, and planarity in a separate calculation using the "Model-fit" technique[1].

The program uses space group specific fast fouriers to calculate both the structure factors and the difference map used to get the gradients (see ref.2 and Appendix 1) and is therefore relatively fast.

It is possible to have a general program which does all the fourier transforms in P1, but this would increase the time for these parts of the calculation by a factor equal to the number of symmetry operations, so we have always substituted appropriate fast fourier transforms for each space group (It is surprisingly easy to do this once you master some very simple space group theory, except for true horrors like cubic space groups). All the rest of the program is general, and at present we have versions for P1, $P2_1$, $P2_12_12$, $P2_12_12_1$, $P4_12_12$, $P3_12_1$ and $P3_1$.

The program has all been very carefully overlaid to keep core requirements to a minimum, and at present it needs core equal to approximately eight times the number of atoms to function, plus the input/output overheads. (This has been achieved at the expense of a good deal of input and output to scratch files, which may be undesireable on some computers).

A further modification has been incorporated recently following a suggestion of Alain Lifchitz, which has increased the speed per cycle by about 35%. He pointed out that $G(x_m)$, $G(y_m)$, and $G(z_m)$ which Ramesh Agarwal generated by convoluting the atomic density with three modified difference fouriers, Dx, Dy, and Dz, could be obtained by convoluting the derivatives of the atomic density with a single difference fourier. This idea was very straightforward to implement in the generalised program. A full description of his extension is given in Appendix 1, with some comments on how it affects parameter shifts.

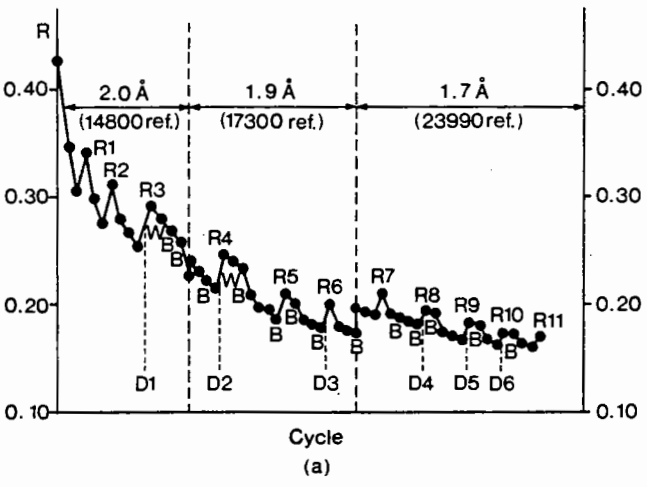
I would like to comment briefly on three structural refinements we have been concerned with.

2. ERROR DETECTION IN THE ACTINIDIN REFINEMENT

A full description of the actinidin refinement has just been published[3]. Ted Baker came to York with a set of co-ordinates derived from a 2.8 Å isomorphously phased electron density map and rapidly refined and corrected his model. The original co-ordinates had a median error of about 0.5 Å from the final set (which is accurate compared to most models at this resolution) but it turned out that 32 of 218 side chains and 3 peptides had been wrongly positioned. The greatest problem was of course to identify these errors. The course of the refinement is shown in figs.1(a) and 1(b). Typically several cycles of automatic refinement would be followed by geometric corrections, and Ted Baker was able to use the comparison of the x-ray derived shifts with the geometric corrections to pinpoint parts of the structure which needed scrutiny. The other criterion used for choosing suspect residues was high and/or inconsistent B values. Once these were identified, they were

TABLE 1

	RESLN	RESIDUES/ATOMS		CORE REQD	TIME/CYCLE	REFLNS	FFT GRID	SPACE GROUP AND CELL						
ACTINIDIN	1.7 Å	218	1821	70k	12 mins	23990	0.6 Å	$P2_12_12_1$	78.20	81.80	33.03	90	90	90
90 COENZYME-B12	1.0 Å		120	55k	3 mins	5260	0.3 Å	$P2_12_12_1$	27.93	21.62	15.35	90	90	90
DAS INSULIN	3.1 Å	102	810	70k	2.4 mins	1452	1.0 Å		83.20	83.20	33.0	90	90	120



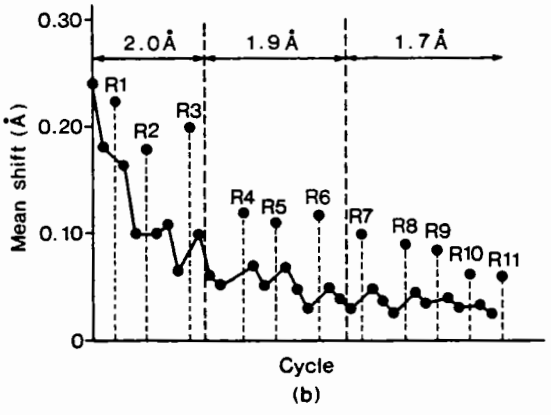


Fig.1(a) A plot of R during refinement. Regularisations of the structure are marked R; B refinement cycles designated B; steps where the only change was the introduction of further solvent molecules W; unmarked points represent xyz refinement cycles. The points where difference maps were calculated are indicated (D1, D2, etc.). Stages where the data were extended to higher resolution are also shown.

(b) Plot showing the mean shift in atomic positions in each xyz refinement cycle during the refinement. Mean shifts for regularisations (R) are also shown.

omitted from the next difference map, which was contoured with the original co-ordinates plotted onto the density. Even without a graphics system this made it fairly easy to decide which residues were in fact correct, and which were misplaced. So in this case apparent anomalies between the two least squares minimisations and between B values of adjacent atoms were actually used to correct the model. I would recommend anyone using a restrained refinement to be careful to look for such inconsistencies, and not just to accept the final set of parameters.

3. CONVERGENCE OF THE BLOCK DIAGONAL MODEL

Both x-ray and neutron data have been collected for the B12 coenzyme by Peter Lindley, John Finney and Hugh Savage of Birkbeck College. A very similar structure had been determined by Gayland Lenhert in Oxford in 1960, and his co-ordinates were used to start refinement against both data sets. The root mean square error between the Lenhert co-ordinates and the current set is about 0.5 Å with a maximum error of over 1 Å for N41. Figure 2 illustrates the difference between the initial co-ordinates, those obtained after 10 cycles of completely mindless block diagonal refinement, and those Hugh Savage obtained after three cycles of full matrix

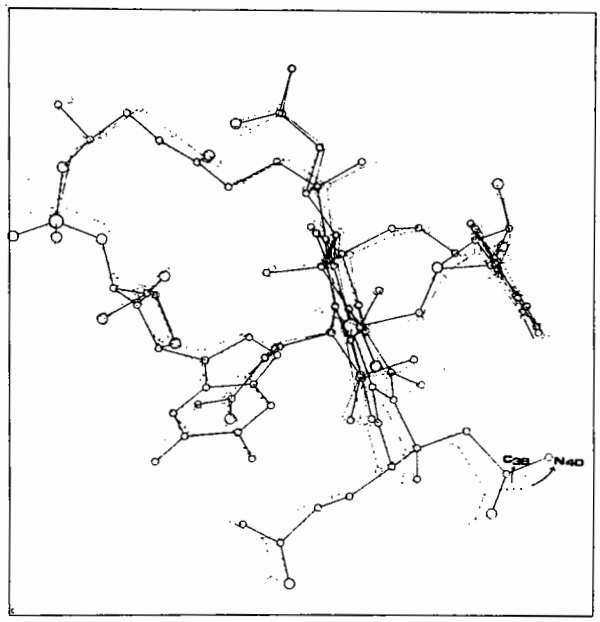


Fig.2 The projection of the co-enzyme B12 structure along the y axis.
——— Original co-ordinates (Lenhert)
----- Co-ordinates from block diagonal refinement
▬▬▬ Co-ordinates from full matrix refinement

refinement on the London CDC7600, several difference maps, and a lot of headaches. All atoms had moved towards their current positions during the block diagonal refinement except N41. The largest correc-

tion made was 0.96 Å to C38, and the root mean square difference between the block diagonal set and the full matrix set is 0.16 Å, while they have each shifted 0.35 Å on average from the Lenhert set. So here block diagonal refinement has performed as well as the full matrix one, and has been a great deal easier to use. This is encouraging for protein refinement, where we are forced to use some modification of block diagonal refinement, and I suspect that for our problems where the data is limited and poor, and our models are incomplete and somewhat disordered we would gain very little even if we could use a full matrix technique.

4. REFINEMENT WITH LIMITED X-RAY DATA

For many proteins only limited data are available. I was interested in evaluating how far refinement could proceed with such limited data sets, so I set up some tests using an actinidin model. I used as starting co-ordinates the hand built model, correcting all residues which were rebuilt durng the course of the refinement.

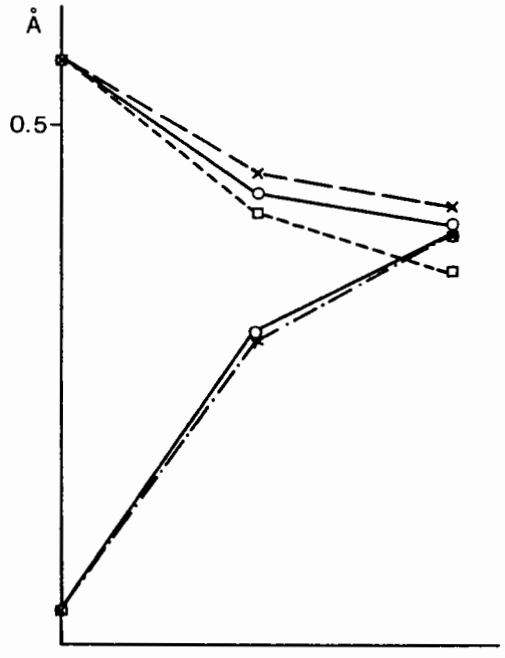


Fig.3 Root mean square differences between test refined co-ordinates: 1. from starting set. 2. from final set (R at 17.1%)
- - - 2.8 Å cycles
——— 2.5 Å cycles
----- 1.7 Å cycles

Two cycles were run using data to 2.5Å, 2.0Å, 1.9Å, and 1.7Å respectively, followed by geometric corrections. The results are illustrated in fig.3. For each calculation there was a considerable improvement in the root mean square error of the co-ordinates. As expected at lower resolution there is less to be gained in additional cycles; at this resolution all co-ordinates must have high estimated standard deviations, and there is little point in continuing refinement once the shifts fall below the e.s.d.

Encouraged by these results Colin Reynolds (York University) has been using "refinement" at 3.1 Å to improve the phases of DAS insulin. DAS insulin is a modified 2Zn insulin with a diaminosuberoyl cross link between the A1 amino group and B29 lys. It is almost isomorphous to 2Zn insulin, but the cell demensions differ by 0.7 Å, sufficient to make difference maps between DAS insulin and 2Zn insulin very noisy. Isomorphous data was collected for two derivatives to 3.1 Å, but again the isomorphous map gave an incomplete picture of the cross link, particularly for molecule 1. Figure 4 shows the current picture of the cross link obtained from an $F_{obs} - F_{calc}$ map after the DAS co-ordinates have been improved by

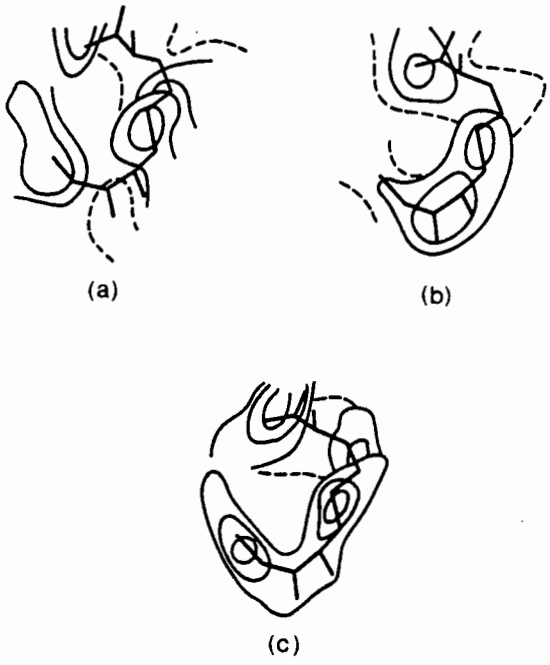


Fig.4 DAS Insulin F_{OBS} maps
(a) phased with isomorphous data
(b) phased with α_{calc} from starting co-ordinates
(c) phased with α_{calc} from refined co-ordinates

cycles of refinement followed by the application of tight geometric restraints. He feels confident that this shows that there has been an improvement in the F_{calc} phases, and that these are better than the isomorphous set.

5. EXTENSION TO BLOCK DIAGONAL LEAST SQUARES MINIMISATION

Ramesh Agarwal has reviewed the process of least squares minimisation at the beginning of his paper[2]. To quote him:

If we wish to minimise a quantity

$$P1 = \sum_S W(s) \; (|F_{obs}(s)| - |F_{calc}(s)|)^2$$

where $S = (hk\ell)$

$$= \sum_S W(s) \; (E(s))^2$$

we need to find shifts [Δu] for R parameters such that

$$[H] \; [\Delta u] = - \; [G]$$

where H is an R × R matrix,

[G] and [u] are column vectors of length R

with the (n,m) term of H, $H(n,m) = \sum_S \frac{\partial E(s)}{\partial p_n} \frac{\partial E(s)}{\partial p_m} W(s)$

and the n^{th} term, of G, $G_n = \sum_S \frac{\partial E(s)}{\partial p_n} E(s) \; W(s)$

Carrying through the differentiation, Agarwal shows

$$H(x_n, x_m) = \sum_S -0.5 g_m(s) g_n(s) (4\pi^2 \; h^2) W(s) \; \exp[i2\pi s.(r_n - r_m)]$$

and

$$G(x_m) = \sum_S g_m(s)(-i2\pi h) W(s) E(s) \; \exp[i\phi(s)] \; \exp(-i2\pi)(hx+ky+lz)$$

In restrained refinement we set out to minimise some composite function

$$P = P1 + P2$$

where for example

$$P = P1 + \sum_{bonds} W(bond) \; (B_{obs} - B_{ideal})^2$$

The least squares equations now have the form

$$[H1 + H2] \; \Delta u = - \; [G1 + G2]$$

and

$$H_{(n,m)} = \sum_S \frac{\partial E(s)}{\partial p_n} \frac{\partial E(s)}{\partial p_m} W(s) + \sum_{nb} \frac{\partial B_{(nb)}}{\partial p_n} \frac{\partial B_{(nb)}}{\partial p_m} W(nb)$$

In full matrix least squares minimisation of small crystal structures all the terms of H are computed and the whole matrix is inverted. However this procedure becomes enormously time consuming for large structures on two counts.

1. to invert a large matrix by conventional techniques is expensive and
2. it takes a long time to evaluate each $H_{n,m}$ term.

All protein crystal least squares refinements have introduced short cuts.

6. MATRIX INVERSION

The most drastic way to reduce the time needed to compute $[H]^{-1}$ is to set all off-diagonal terms to zero, that is, we ignore any correlation terms between different atoms (this is known as the block diagonal approximation). Table 2 (Ex.3) shows a sample of an H1 matrix generated from the x-ray observations, and you can see that the off-diagonal terms are smaller than those on the diagonal. All the refinements I will describe have used this approximation, and I think most Hendrickson type refinements also set the off-diagonal derivatives of the x-ray observational equations to zero. The derivatives of H2 are obtained only for neighbouring atoms, so that in restrained refinement the H matrix is still largely zero, since there are no correlations considered between distant atoms, but there are some off-diagonal elements. There is an algorithm called the conjugate gradient method of matrix inversion, which I believe is used in all the restrained refinement programs to evaluate H. This is fast even for quite large matrices; we use it on the DEC-10 for the program MODELFIT, and it takes about one minute to refine 2400 parameters using a matrix of 15870 elements.

7. EVALUATING $H_{(n,m)}$

It is straightforward to evaluate the derivatives for the restraints. They do not involve any time consuming summations[1,4]. For some time Neil Isaacs, Ramesh Agarwal, myself, and other people have discussed whether it would be worthwhile to attempt to extend our block diagonal refinement by including the $H(x_n,x_m)$ terms derived from the x-ray observational equations for neighbouring atoms, and then to use the conjugate gradient inversion technique. These terms would help to preserve sensible geometry although they would not be nearly as effective at this as the restraint terms.

We felt that this might avoid one of the difficult problems for restrained refinement, which is how to define relative weights for the two different sets of least squares equations derived from H1 and H2. If the restraints are overweighted relative to the x-ray observations, it would be possible to reach a minimum where the structure has perfect geometry, but is wrong. After all, as crystallographers, we offer the world sets of parameters which are meant to match our unbiased observations! Alternatively if the restraints are underweighted, you may be doing a lot of unnecessary computing in using the conjugate gradient method, where the block diagonal inversion would have given a very similar result at a fraction of the cost. But the snag was to evaluate the off-diagonal terms $H(x_n,x_m)$ at a reasonable cost.

All the ideas below are derived from discussion, and they lean especially on Ramesh Agarwal's letters to Neil Isaacs, in which he would clarify and assess our suggestions.

While I was in Melbourne I realised that it would be quite easy to tabulate an H map as a three dimensional table, assuming a full sphere of x-ray observations. Since $g_m(s)$ ang $g_n(s)$ are both sums of gaussian terms their product is also a sum of gaussians. So for small r_n-r_m H(xn,xm) can be seen as a type of Fourier map to be sampled at the required points r_n-r_m.

We can write

$$g_m(s)\ g_n(s) = \sum_{ngauss} c_i \exp(-b_i \sin\theta^2/\lambda^2)$$

and

$$H(xn,xm) = \sum_{ngauss} c_i \exp(-b_i \sin\theta^2/\lambda^2)\ 4\pi\ h^2\ W(s)\ldots$$

For a pair of atoms separated by the vector r_n-r_m, and with form factors defined as a sum of gaussians, the H(xn,xm) term will be a sum of functions of 3 variables

1. the projection X_{nm} of r_n-r_m in the x direction,
2. its projection Z_{nm} perpendicular to the x axis,
3. and the gaussian B_i values.

The tabulation is simplified if we consider our atom parameters (X,Y,Z), relative to an orthonormal axial system. We usually think of the atomic parameters (x_f,y_f,z_f) relative to the crystal axes, and

$$Fo(s) = \sum_{natoms} g_m(s) \exp[-2\pi\ i(hx_m + ky_m + kz_m)]$$

But we can just as well use as parameters (X_o,Y_o,Z_o), where

$$\begin{bmatrix} Xo \\ Yo \\ Zo \end{bmatrix} = [Ro] \begin{bmatrix} x_f \\ y_f \\ z_f \end{bmatrix}$$

If we define $(H_o,K_o,L_o) = (h\ k\ l)\ [Ro]^{-1}$, then

$$(h\ k\ l) \begin{bmatrix} x_f \\ y_f \\ z_f \end{bmatrix} = (H\ K\ L) \begin{bmatrix} X_o \\ Y_o \\ Z_o \end{bmatrix}$$

$$G(X_n) = \sum_S g_m(s)(-i2\pi H_o)W(s)E(s)\ \exp[i\phi(s)]\ \exp(-2\pi i(hx+ky+lz))$$

This means we can calculate $G(x_m)$ by convolution of the difference fourier with the derivative of the atomic density just as before, and then substitute

$$G(X_m),G(Y_m),G(Z_m) = [G(x_m),G(y_m),G(z_m)]\ [R_o]^{-1}$$

$$H(X_n,X_m) = \sum_S -0.5 g_m(s) g_n(s)(4\pi H_o^2)W(s)\ \exp[i2\pi s.(r_n-r_m)]$$

$$= \sum_S \sum_{ngauss} c_i \exp(-b_i(\sin\theta^2/\lambda^2))\ (4\pi H_o^2)W(s)\ \exp[i2\pi s.(r_n-r_m)]$$

$$= \sum_{ngauss} H(X_{nm}\ Z_{nm}\ B_i)$$

Figure 5 shows several of these $H(X_{nm},Z_{nm},B_i)$ maps, calculated from data generated for different unit cells, but all with the same weighting factor W(s)

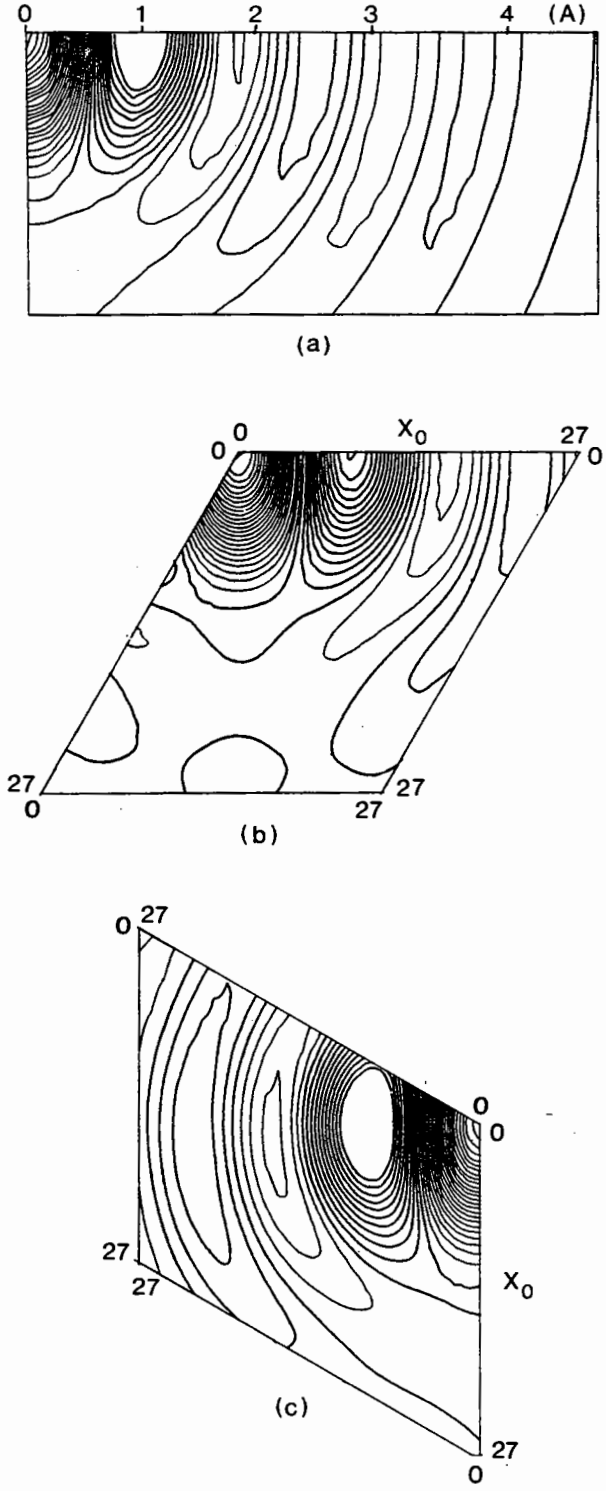


Fig.5 Normalised H maps for Bi = 5, data resolution is 1.5 Å, weighting factor is $\exp(-2.0 \sin\theta^2/\lambda^2)$. The H map function is $H(X_{nm},Z_{nm},Bi) =$

$$2/k \; \Sigma \; 4\pi \; H_o^2 \; W(s)\exp(2\pi s.(rn-rm)) \; \exp(-5 \sin\theta^2/\lambda^2)$$

where k = number of reflections.
a) $H_x(X_{nm},Z_{nm},Bi)$ - cell 50 50 50 90 90 90
b) $H_x(X_{nm},Z_{nm},Bi)$ - cell 30 40 35 90 120 90
c) $H_z(X_{nm},Z_{nm},Bi)$ - cell 30 40 35 90 120 90
Note the identical form about the Z axis.

and with the same resolution limits. It is clear I hope that all these maps are virtually identical, so that when we want to estimate values of $H(x_n,x_m)$ for some problem, we can first calculate a H table with the appropriate weights and resolution limits; (these take a minute or so to generate using a modified fft program) and then simply look up the required values for each atom pair which is close enough to warrant inclusion. You can see from the plots that the magnitude of terms falls off very quickly with increasing r_n-r_m. The origin peak gives the value of the $H(X_n,X_m)$ term, the diagonal of the H matrix. Figure 6 shows how the relative magnitudes of the diagonal and off-diagonal terms alter with resolution. At low resolution there is much more correlation between different atoms over a greater range of separation.

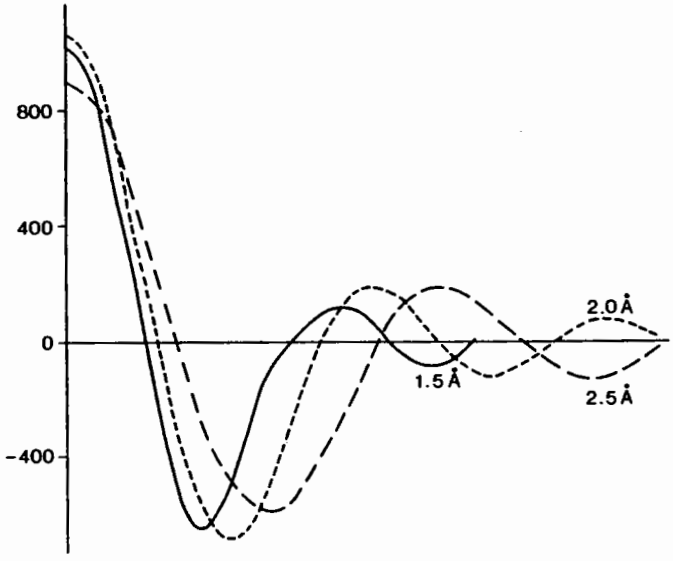


Fig.6 The line at $H_x(O,Z_{nm},2O)$, resolution limits 2.5 Å, 2.0 Å, and 1.5 Å. The scale is arbitrary.

It is too soon for me to judge how useful it will be to include these off-diagonal terms. Obviously they will help to prevent atoms coming too close together, and their contribution will be relatively stronger at lower resolution as we would wish, but they do not represent any extra "observations" as the restraints do. Table 2 shows three sections of different H matrices, the first is from an unrestrained Hendrickson refinement of Tortoise Lysozyme at 3 Å resolution, the second shows the extra terms added in when restraints were included, and the third is the one I would generate at 2.5 A.

Table 2

Example 1: Diagonal elements of H1 for Tortoise Lysozyme from X-ray observations 3A data, conventional summation from Konnert, Hendrickson program.

	CA			N			C			O		
	x	y	z	x	y	z	x	y	z	x	y	z
	1000	0	0	0	0	0	0	0	0	0	0	0
CA	0	1020	0	0	0	0	0	0	0	0	0	0
	0	0	511	0	0	0	0	0	0	0	0	0
	0	0	0	1491	0	0	0	0	0	0	0	0
N	0	0	0	0	1584	0	0	0	0	0	0	0
	0	0	0	0	0	770	0	0	0	0	0	0
	0	0	0	0	0	0	1023	0	0	0	0	0
C	0	0	0	0	0	0	0	1011	0	0	0	0
	0	0	0	0	0	0	0	0	520	0	0	0
	0	0	0	0	0	0	0	0	0	2139	0	0
O	0	0	0	0	0	0	0	0	0	0	2223	0
	0	0	0	0	0	0	0	0	0	0	0	1107

Example 2 H2 terms matching H1 above, derived from restraints used in Konnert, Hendrickson program.

	1194	0	0	-1563	468	248	-15	-63	-8	-14	-33	-11
CA	0	732	0	412	-177	-52	-38	-1093	-648	-33	-78	-28
	0	0	1005	264	-63	-45	2	-613	-418	-11	-28	-10
	-1563	412	264	311	0	0	-22	75	23	0	0	0
N	468	-177	-63	0	-1215	0	43	-33	-27	0	0	0
	248	-52	-45	0	0	-624	28	-38	-16	0	0	0
	-15	-38	2	-22	43	208	991	0	0	-837	-880	5
C	-63	-1093	-613	75	-33	-38	0	1279	0	-880	-927	5
	-8	-648	-418	23	-27	-16	0	0	625	5	5	0
	-14	-33	-11	0	0	0	-837	-880	5	-799	0	0
O	-33	-78	-28	0	0	0	-880	-927	5	0	-845	0
	-11	-28	-10	0	0	0	5	5	0	0	0	-859

Example 3. H1 matrix terms estimated from H table for 2.5 Å resolution data.

	140	0	0	-87	0	0	29	0	0	-10	0	0
CA	0	140	0	0	-28	0	0	-37	0	0	23	0
	0	0	140	0	0	32	0	0	-21	0	0	-10
	0	0	0	190	0	0	-11	0	0	-8	0	0
N	0	0	0	0	190	0	0	-12	0	0	20	0
	0	0	0	0	0	190	0	0	-10	0	0	4
	0	0	0	0	0	0	140	0	0	-2	0	0
C	0	0	0	0	0	0	0	140	0	0	-5	0
	0	0	0	0	0	0	0	0	140	0	0	71
	0	0	0	0	0	0	0	0	0	248	0	0
O	0	0	0	0	0	0	0	0	0	0	248	0
	0	0	0	0	0	0	0	0	0	0	0	248

APPENDIX 1

by R. Agarwal, A. Lifchitz and E. Dodson

We would like to report an extension to the block diagonal least squares refinement method described by Ramesh Agarwal[2]. Agarwal shows that it is possible to use the convolution of the atomic density with a modified difference fourier (which can be calculated by the fast fourier transform) to generate the diagonal gradient terms required for least squares refinement.

We will use his notation for simplicity, and summarise his method. (Equation numbers here match those in Agarwal's paper[2]).

$r_m = (x_m, y_m, z_m)$ - fractional cell co-ordinates of the mth atom,

B_m - isotropic thermal parameter of the mth atom,

$s \equiv hkl$ - a diffraction point in reciprocal space,

$s^2 \equiv |s|^2 = (4 \sin^2 \theta)/\lambda^2$.

$f_m(s) \equiv f_m(hkl)$ - scattering factor of the mth atom at reciprocal distance s,

$$= \sum_i^n c_i \exp(-b_i s^2/4)^{(2,6)}$$

$s.r_m \equiv hx_m + ky_m + lz_m$

$F_o(s) = F_o(hkl)$ = observed structure amplitude.

With this notation the expression for calculated structure factors is

$$F_c(s) = \sum_m f_m(s) \exp(-B_m s^2/4) \exp(i2\pi s.r_m) \quad (20)$$

To simplify the equations further, we introduce the notation

$$g_m(s) \equiv f_m(s) \exp(-B_m s^2/4). \quad (21)$$

Here $g_m(s)$ represents the contribution of the mth atom to structure factors after taking into account its thermal motion. This notation greatly simplifies the expressions for gradient and normal-matrix terms. With this notation (20) becomes

$$F_c(s) = \sum_m g_m(s) \exp(i2\pi s.r_m) \quad (22)$$

In least-squares refinement of the atomic parameters, we minimise the following function

$$P = \frac{1}{2} \sum_s W(s)\{|F_c(s)| - |F_o(s)|\}^2 \quad (17)$$

where W(s) is any desired weighting function.

P is minimised with respect to x_m, y_m, z_m and B_m to give parameter shifts of the form of $\Delta\rho_m = -[H]^{-1}G$ where [H] is the normal matrix of order M, the number of parameters, and G is the gradient vector, also of length M. In the block diagonal approximation only the diagonal terms of [H] are included in the calculation. This means that the largest part of the calculation time is spent on generating the elements of G.

Agarwal shows that

$$G(x_m) = \sum_s g_m(s)(-i2\pi h)W(s)E(s)\exp[i\phi(s)] \times \exp(-i2\pi s.r_m). \quad (27)$$

Similar expressions can be derived for $G(y_m)$ and $G(z_m)$. The expression for $G(B_m)$ is as follows.

$$G(x_m) = \sum_s g_m(s)(-s^2/4)W(s)E(s)\exp[i\phi(s)] \times \exp(-i2\pi s.r_m). \quad (28)$$

Eqn.(27) can be written as follows:

$$G(x_m) = \sum_s D_x(s) g_m(s) \exp(-i2\pi s.r_m). \quad (50)$$

where $D_x(s)$ is a function common to all the atoms and is defined by:

$$D_x(s) \equiv (-i2\pi h)W(s)E(s)\exp[i\phi(s)) \quad (51)$$

The subscript x denotes gradient with respect to x parameters. Equation (50) represents the Fourier transform of the product of two functions $D_x(s)$ and $g_m(s)$, evaluated at r_m (position of mth atom). According to the convolution theorem, multiplication in reciprocal space is equivalent to convolution in real space. Let $d_x(r)$ be the Fourier transform of $D_x(s)$. Since $D_x(s)$ contains $E(s)\exp(i\phi(s))$, $d_x(r)$ can be thought of as a modified difference density function. Say $\rho_m(r)$ is the Fourier transform of $g_m(s)$. Then by the convolution theorem $G(x_m)$ is the convolution of $d_x(r)$ and $\rho_m(r)$, evaluated at $r=r_m$:

$$G(x_m) = \int d_x(r)\rho_m(r - r_m)\, dr$$

For computation purposes, the integration can be replaced by a summation over the grid r:

$$G(x_m) = \sum_r d_x(r)\rho_m(r - r_m) \quad (52)$$

To reduce the computation, Alain Lifchitz suggested grouping the terms of eqn.(50) differently.

Consider $D(s) \equiv W(s)\, E(s) \exp[i\phi(s)]$ and let $d(r)$ be the Fourier transform of $D(s)$. ($d(r)$ is in fact the weighted difference fourier).

Then $-2\pi i h\, gm(s) \exp(-2\pi i s.rm)$ can be shown to be equal to $\partial\rho_m(r-r_m)/\partial x_m$, and the gradients for x_m, y_m, and z_m can be obtained from the convolution of the weighted difference fourier map with the respective derivatives of the model electron density.

The fourier relations for a Gaussian function are as follows:

real space		reciprocal space
electron density		atomic scattering factors
$\rho(r)$	+FT+	$f(s)$
$(4\pi/B)^{3/2}(\exp(-4r^2r^2/B)$	+FT+	$\exp(-Bs^2/4)$

where r represents the distance from the centre of the atom. Put $R_m = r-r_m$, ie R_m represents the distance from the centre of atom m. So for an atom with position r_m whose scattering can be described as a sum of n gaussians

$$\sum_{i=1}^{n} c_i \exp(-(b_i + B_m)s^2)$$

the associated model electron density is

$$\rho_m(r-r_m) = \sum_i^n C_i \left(\frac{4\pi}{b_i + B_m}\right)^{3/2} \exp\frac{(-4\pi^2(r-r_m)^2}{b_i + B_m} \qquad (A)$$

$$= \sum_i^n c_i' \exp(-b_i' R_m (r-r_m)^2$$

We can also write

$$\rho_m(R_m) = \sum_s g_m(s) \exp(-2\pi i s.R_m) \qquad (B)$$

Hence

$$\frac{\partial\rho m(R_m)}{\partial x_m} = \sum_s (-2\pi i h)\, g_m(s) \exp(-2\pi i s.R_m) \quad \text{from (B)}$$

which will equal

$$\sum_1^n c'\, b' \frac{\partial(R_m^2)}{\partial x_m} \exp(-b' R_m^2) \quad \text{from (A)}$$

$$R_m^2 = (\Delta x_m)^2 + (\Delta y_m)^2 + (\Delta z_m)^2 + 2\,\Delta x_m\,\Delta y_m\,\Delta z_m \cos\gamma + 2\,\Delta y_m\,\Delta z_m \cos\alpha + \Delta z_m\,\Delta x_m \cos\beta$$

and $$\frac{\partial(R_m^2)}{\partial x_m} = 2(\Delta x_m + \Delta y_m \cos\gamma + \Delta z_m \cos\beta)$$

and so we can convolute the difference map with

$$\sum_1^n 2c'_i b'(\Delta x_m + \Delta y_m \cos\gamma + \Delta z_m \cos\beta) \exp(-b' R_m^2)$$

to obtain Gx_m, and with the other appropriate derivative terms to obtain Gy_m and Gz_m, in a single pass through the model density calculation.

That part of the program which generates atomic density and convolutes it with D(s) is the same for all space groups, as is the data input and output file handling the calculation or the diagonal terms of the normal matrix. Incorporating this idea has reduced the time for a cycle by 30% to 35%. Previously a separate modified difference fourier and convolution calculation was done to obtain each of the $G(x_m)$, $G(y_m)$ and $G(z_m)$ terms, but now these are all obtained from a single pass.

The method has been recoded to reduce the space group specified sections of the calculations to a minimum. The structure factors and difference fourier calculations can be always done in space group P1, and the convolutions restricted to an assymetric unit for the particular space group, but this increases the time required and the scratch storage used during a refinement run by a factor approximately equal to the number of symmetry operations. We use modified fast fourier routines which take advantage of the crystal symmetry where possible.

Test Calculations

There are inherent inaccuracies in this method owing to

(i) The atomic density being sampled on a grid in real space before being transformed to give the structure factors, these errors becoming more serious as the grid is made more coarse.

(ii) The restriction of the atomic radius to some r_{max}.

(iii) Restriction of the number of gaussians for each atom type to one or two.

The error from (iii) is known. Agarwal shows that with a two gaussian approximation the error at 1.5 A resolution is about 1%[2]. The five gaussian approximation gives virtually no error at all[6].

The error from (ii) is a function of b'(= bi + bm). The percentage error is illustrated for different b' in table A1. However the error from (i) is harder to estimate, and for both (i) and (ii) the errors differ for different atoms.

Table A1

Percentage errors of

	$\exp(-4\pi^2r^2/b')$			$r\exp(-4\pi^2r^2/b')$		
	Rmax			Rmax		
b'	1.5	2.0	2.5	1.5	2.0	2.5
5	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.01	0.0	0.0
20	0.19	0.0	0.0	0.86	0.02	0.0
30	1.11	0.08	0.00	4.22	0.39	0.02
40	2.77	0.64	0.03	9.32	1.57	0.16
50	4.88	0.95	0.13	15.00	3.61	0.59

To test whether this modification to Agarwal's method had reduced the accuracy of the method we ran tests on gramacidin-S. This is a ten peptide structure space group $P3_12_1$, which has been refined by the Agarwal method using copper radiation data of 1.A resolution. There are 96 atoms in the structure with an average isotropic B factor of 10.24 Å[2][5]. I generated calculated F_{hkl} from the known atomic positions to a resolution limit of 1 Å; then applied random shifts to the x, y and z co-ordinates of up to 0.1 Å. This incorrect set of positions was then refined against the calculated F_{hkl}, using the original program and the modified one, testing different values of r_{max} for both the convolution step, and the generation of electron density from the atomic positions. In all cases the grid was sampled at approximately $^1/_3$ A along each axis, and the five gaussian form factors were used[6].
Four tests were made.

		r_{max} (density generation)	r_{max} (convolution)
1)	Original program	2.5	2.0
2)	Modified program	2.5	2.0
3)	Original program	2.646 = √7	2.646
4)	Modified program	2.646	2.646

For each single cycle of block diagonal least squares refinement was run on an IBM360. The results are tabulated in Tables A2, A3 and A4.

Table A2

	R Value		Time	RMS Error in Position (Å)		
	Initial	Final	(mins)	Initial	Final	Shift
1)	12.31	6.63	6.52	0.08912	0.05820	0.08167
2)	12.31	5.67	4.12	0.08912	0.05012	0.08142
3)	12.24	6.53	3.15	0.08912	0.05731	0.07660
4)	12.24	5.55	5.22	0.08912	0.04988	0.07875

Table A3

R.M.S. Differences between different solutions in Å's

	1	2	3
2	0.0270	-	0.0056
3	0.0056	0.0269	-
4	0.0269	0.0036	0.0255

Table A4

Residual Fraction of Initial Error for Carbon Atoms in Different B ranges after 1 cycle.

B Value	n atoms	1	2	3	4
6-8	10	0.53	0.36	0.53	0.38
8-10	21	0.62	0.59	0.61	0.59
10-12	10	0.67	0.39	0.64	0.38
12-14	5	0.65	0.67	0.63	0.65
14-16	4	0.54	0.66	0.50	0.63
16-18	4	0.83	0.70	0.77	0.66
18-20	4	0.72	0.78	0.70	0.78
>20	4	0.79	0.92	0.80	0.90

Note: 1) There is little gain obtained from either version by increasing r_{max}. The extra time used would probably have been better spent on beginning another cycle. 2) For atoms with low B values the modified program gives a slightly better result, but the modification does not work so well for atoms with high B values. In fact any convolution method must have problems with such atoms, which are likely to have poorly defined gradients in the difference fourier. Increasing the atomic radius for the convolution step can do little to solve this.

References

1. E.J. Dodson, N.W. Isaacs and J.S. Rollett, Acta Cryst. A32, (1976) 311.

2. R.C. Agarwal, Acta Cryst. A34, (1978) 791.

3. E.N. Baker and E.J. Dodson, Acta Cryst. A36, (1980) 559.

4. J.H. Konnert, Acta Cryst. A32, (1976) 614; J.H. Konnert and W. Hendrickson, Acta Cryst. A36, (1980) 344.

5. S.E. Hull, R.Karlsson, P. Main, M.M. Woolfson, and E.J. Dodson, Nature, 275, No.5677 (1978) 206.

6. International Tables for X-ray Crystallography (1974). Vol.4. Table 2.2B, p.99 (Birmingham: Kynoch Press).

REFINEMENT EXPERIENCES USING CHAIN CONSTRAINTS IN REAL SPACE AND ENERGY RESTRAINTS IN RECIPROCAL SPACE

by

Wolfgang Steigemann
Max-Planck-Institut fuer Biochemie, 8033 Martinsried, FRG

1. INTRODUCTION

In the field of protein crystallography initial atomic models of the biomolecules are generally obtained by the techniques of multiple isomorphous replacement and/or molecular replacement. These models are only approximate and must be improved in order to deduce sufficiently reliable results for the understanding of their biological function. The limited amount of reflection data generally prohibits the application of conventional least-squares refinement (an exception is the refinement of rubredoxin by Watenpaugh et al[1] at very high resolution) because of a poor ratio observations/parameters. This ratio can be improved either by increasing the number of observations (which need not necessarily be X-ray observations) or by reducing the number of parameters. An approach of the latter kind is Diamond's real space refinement[2]. This method has been used extensively in our laboratory in a cyclical manner introducing newly calculated phases into the electron density map after each cycle[3].

2. REAL SPACE REFINEMENT WITH CHAIN CONSTRAINTS

In this procedure[2] the bond lengths and most bond angles are kept strictly fixed as introduced into the starting model[4]. The dihedral angles in the main and side chain are the variable parameters. The refinement in real space has several advantages and disadvantages. The latter are partly due to the reduced flexibility of the polypeptide by means of imposing chain constraints.

2.1 Advantages

1. A major advantage is the possibility for the improvement of the initial model in the MIR (multiple isomorphous replacement) map. Based on a given sequence of the molecule an optimum interpretation of such a map can be obtained before the observed phases are replaced by calculated ones.

2. Any type of map can be used (except difference Fourier maps). In advanced stages these may comprise (2Fo-Fc)-maps[3] or in more difficult cases "phase-combined" maps[5] which make use of phase information from independent sources[6].

3. The refinement may be confined easily to selected portions of the structure.

4. The radius of convergence is in general larg[7].

2.2 Disadvantages

1. A principal disadvantage of the method is the restriction to structures determined at medium or high resolution. In our experience, serious problems arise in the fitting procedure at a resolution lower than 3 Å, since residues distant in sequence but adjacent in space can easily share the same electron density. The resulting close non-bonded contacts of "merging" side-chains are unacceptable.

2. The absolute rigidity of bond distances and most of the angles may lead to a serious strain in the model. The distribution for the generally flexible main chain bond angle at C^{α} becomes rather broad[8] (typically 10°). This is physically unreasonable[3].

3. The crucial step of manual intervention for the correction of gross errors (e.g., a peptide flip) poses problems of a more technical nature. The desired target positions can, in principle, be specified but are not necessarily met because of the rigidity of the model. Complicated mathematical manipulations are necessary for the projection of Cartesian coordinate space into dihedral angle space with fixed margins.

4. A related problem is that it is virtually impossible in practice to insert, delete or modify the side chains of amino acids, as is often necessary if the chemical sequence is not known. The model is not allowed to be stereochemically distorted (i.e. bond lengths and angles must be kept ideal) since no information about the stereochemistry of single amino

acids is available in the real space procedure. Perfect bonding parameters can be introduced only by tedious model-building[4] which is impracticable for frequent use.

These latter "manual" interventions were particularly time-consuming in routine application and an improvement was highly desirable. Based on this experience, a much higher flexibility of the model was allowed on the VG 3400 interactive graphics system[9] for model repair. There a modified version of Herman's & McQueen's model building procedure[10] is used. A more accurate restoration of ideal geometry and check of non-bonded contacts is then performed by Levitt's energy refinement[11].

3. RECIPROCAL SPACE REFINEMENT WITH ENERGY RESTRAINT

From our experience with the graphics system it was obvious that relaxation of geometry would also be favourable for a more rapid convergence of the refinement process in automated refinement.

The first approach towards this goal has been reported by Konnert[12] who introduced geometric restraints into the structure factor least squares procedure. A discussion about a more elaborate version of his method is described in this volume (W.A. Hendrickson).

We used simultaneous refinement of the X-ray residual $\Sigma(|F_o|-|F_c|)^2$ and potential energy as described by Jack & Levitt[13]. The main reasons for the selection of this procedure were, on the one hand, the application of physically reasonable potential parameters, and on the other hand, the use of FFT (fast Fourier transform) methods for the calculation of the crystallographic contributions. The full procedure has been implemented in our laboratory by J. Deisenhofer.

3.1 Principles

In the following, I shall briefly review some principle features of the method.

The function that is to be minimized is given by

$$R = E + k \cdot X \qquad (1)$$

where E represents the potential energy and X the crystallographic term $\Sigma(|F_o| - |F_c|)^2$. The factor k controls the relative contribution of the crystallographic residual to the total residual R. Its magnitude is of crucial importance as may be verified when it adopts the extreme values k=o and k=∞. In the first case pure energy refinement is performed, in the latter pure crystallographic refinement. Our general strategy is to choose k in such a way that both residuals decrease simultaneously. The potential energy term consists of the following contributions

$$\begin{aligned} E = {} & \Sigma\, 1/2K_b(b_i-b_o)^2 + && \text{(bonds)} \\ & + \Sigma\, 1/2K_\tau(\tau_i-\tau_o)^2 + && \text{(bond angles)} \\ & + \Sigma K_\phi(1+\cos(m\cdot\theta_i+\delta)) && \text{(torsion angles)} \\ & + \Sigma\ (A\cdot r_i^{-12} + B\cdot r_i^{-6}) && \text{(non-bonded interactions)} \end{aligned} \qquad (2)$$

The potential parameters we have used are those described by Levitt[11]. According to the atom types involved in forming bonds and bond angles different values of the force constants are applied which have been derived from vibration spectra of small molecules. The bond angle and torsion angle force constants are corrected for the omitted H-atoms. The coefficients A and B in the Lennard-Jones potential (4th term in eqn. 2) for the non-bonded interactions are of a more empirical nature. Attractive terms (negative B) are used for potential hydrogen bond partners (minimum in the energy function at 2.9 Å) and atoms pairs with more hydrophobic character (minimum energy between 4 Å and 5 Å). Apart from this only repulsive forces are applied. (The detailed values of A and B are chosen to give a 20 kcal/mol contribution at an extreme short contact which lies between 2.6 Å and 3.0 Å).

We regard it as a major advantage of this method that the parameters employed depend on the type of atoms involved in a particular interaction. Specific flexibility or rigidity in particular regions of the molecule may be introduced by use of artificial atom types and appropriately chosen potentials. This was particularly useful in the refinement of oxyerythrocruorin[18].

The crystallographic refinement is approximated by diagonal least squares refinement, the normal matrix and right hand side of which are set up by FFT methods[13]. A conjugate gradient technique is used for the subsequent minimization of X-ray and energy residual.

The type of residual (sum of independent functions in eqn. 1) chosen in this refinement procedure indicates that in this method the number of observations is increased (the target values of the stereochemical parameters and their associated energies) rather than the number of parameters decreased. The energy parameters are not used to reparametrise the least squares problem. Refinement is performed in Cartesian space. Even though, we may term this procedure for simplicity a crystallographic refinement with energy restraints, especially because of the free choice in the parameter k (eqn. 1).

When the shifts resulting from X-ray diagonal least squares are applied to the atomic coordinates with full weight and ignoring stereochemical considerations (i.e. single atom refinement with an effective $k=\infty$ in eqn. 1), the X-ray residual does not necessarily decrease in an optimum way. Depending on the resolution, a damping or enhancement factor has to be applied in order to reach a minimum R-value for a given set of shifts. Its value usually need not to be determined often, so simple "trial and error" structure factor calculations in a selected zone are sufficient. The use of a non-unit weighting factor turns out to be of major importance for an optimum convergence, even in a combined energy and X-ray refinement. It should be noted, that this factor is different from the factor k in eqn. 1.

A full cycle of combined refinement then consists of the following: Starting from a given set of coordinates, structure factors and a difference Fourier map are calculated (either by FFT or conventionally). The difference Fourier is used for the determination of gradients needed to set up the diagonal least squares system. The coordinates are then refined subject to the properly weighted X-ray and energy restraints to give a new model. With the new model the procedure may be repeated. Occasionally the optimum weighting factor for the X-ray shifts (see above) has to be determined. As a short-cut, energy parameters may be intermittently ignored during the refinement of temperature factors. We usually average the independently refined temperature factors over main and side chain in each residue subsequently.

As pointed out before, the manual correction steps, which cannot be performed automatically by any refinement procedure, are of particular importance. This step is by far the slowest in the whole refinement process, a fact which should be taken into account when one considers the investment of time to speed up the automated part of the refinement.

4. APPLICATION OF THE METHOD

The method of crystallographic refinement in reciprocal space with simultaneous minimization of energy has been used extensively for the refinement of structures over a wide range of resolution. Often the primary structure has not been known in advance, which has seriously hampered the progress of refinement.

In the following I shall focus on two examples with different characteristics.

4.1 F_{ab}-fragment of Immunoglobulin G

The first example summarizes the refinement of the antigen binding fragment of IgG1 immunoglobulin, KOL, performed by Marquart[5,14]. The general course of refinement is shown in fig. 1. The R-value and potential energy (excluding the non-bonded interactions) are drawn as a function of the cycle number. The general complexity of the diagram derives from the fact that many manual interventions were necessary. Amino acid sequence information was not known for the variable segments in the polypeptide. The only guide to the sequence in these regions was the knowledge of the peptide composition in the light chain and part of the heavy chain. In the first half of the refinement (counted in terms of cycles) RLSP (real space refinement) was used, initially with data to 2.9 Å and later extended to 1.9 Å with calculated phases. The maps used were of type $(2F_o-F_c)$ and $(3F_o-2F_c)$. The first major corrections were guided by the inspection of difference Fourier maps $(|F_o| - |F_c|)\exp(i\alpha_c)$ which were difficult to interpret, especially in the case of incorrectly recognized amino acids. After cycle 19 no improvements of the model could be derived from the (F_o-F_c)-maps. Therefore, phase information from MIR and the present models using Sim-weights[15] were combined[6] and a map with amplitudes (F_o) weighted by the derived figure of merit calculated. This led to some major revisions in the model.

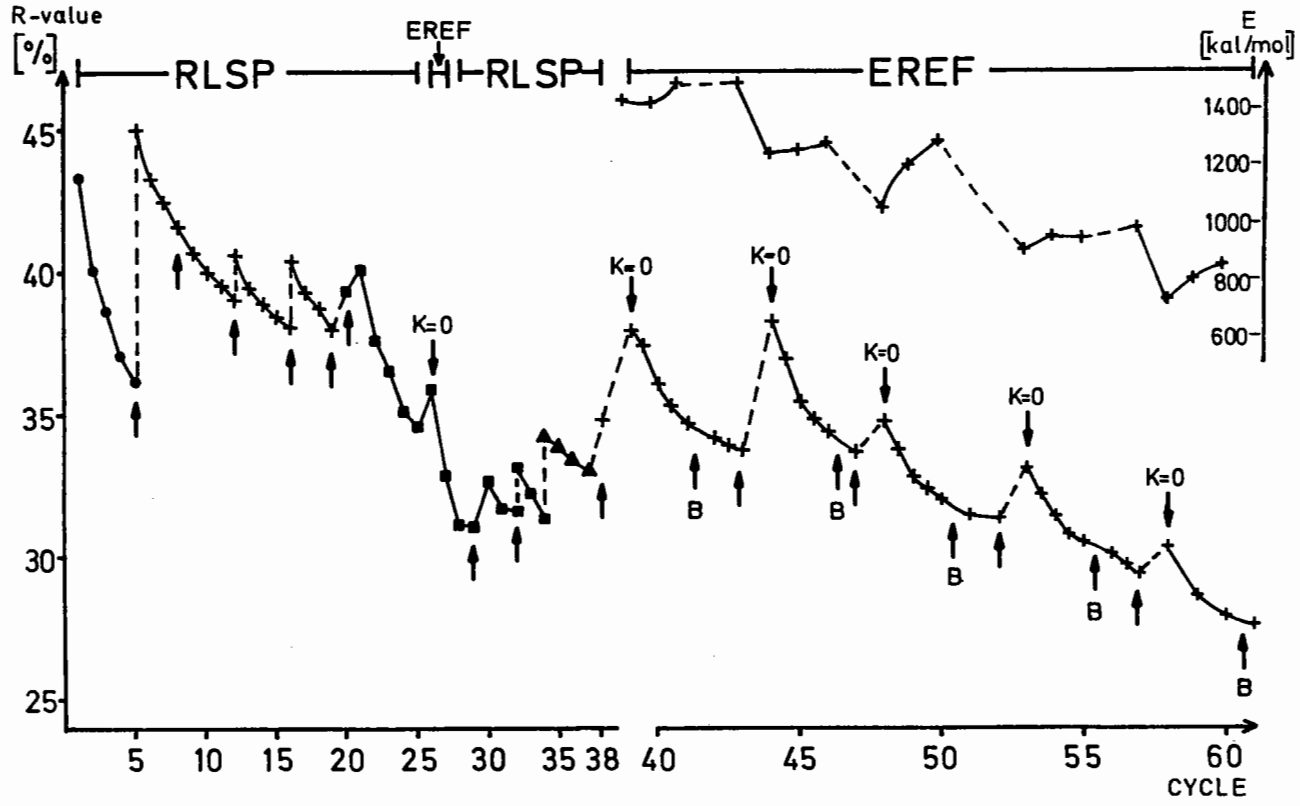


Fig. 1 Course of refinement of the Fab-fragment of IgG1 immunoglobulin KOL. The crystallographics R-value $\Sigma||F_o| - |F_c||/\Sigma F_o$ (lower curve) and the potential energy E without non-bonded interactions (upper curve) are shown in dependence of the cycle number. The various symbols for the R-value denote different ranges in resolution: 5-2.9 Å; 5-2.5 Å; 5-2.2 Å; +5-1.9 Å. The broken sections in each of the curves indicate major revisions in the model. Manual interventions at the graphics system[9] are represented by ↑. Temperature factor refinement is shown by the arrows flagged with B. Two different refinement procedures have been used: RLSP (real space refinement (2)) and EREF (combined X-ray and energy refinement[13]. The refinement cycles flagged with k=0 represent pure energy refinement.

Phase combination turned out to be an extremely useful tool and $(2F_o-F_c)$ maps with such phases, and in later stages, Sim-weighted $(2F_o-F_c)$ maps were used exclusively for model inspection and correction on the interactive graphics display[9]. In order to obtain better convergence, the resolution was decreased to 2.5 Å again and RLSP was applied to the revised model using 'normal' $(2F_o-F_c)$ maps. Cycles of EREF (k=0) using phase-combined maps were employed periodically in the later stages of the refinement purely for inspection and restoration of standard geometry.

At this point it had become clear that the slow progress in refinement was due to some extent to the strong constraints imposed by RLSP. Therefore, from cycle 39 on RLSP was abandoned in favour of EREF. At first, one EREF cycle without X-ray observations (k=0) was performed after each manual intervention in order to restore standard geometry and to detect close non-bonded contacts. These steps are reflected by an increase of the crystallographic R-value and by a decrease of the potential energy. Then several cycles of combined X-ray and energy refinement were performed, adjusting the relative scale factor k in such a way that generally both the X-ray residual (represented by R-value) and the potential energy decreased. Occasionally (e.g. in cycles 49,50) the restraints have been more relaxed by temporary increase of k to overcome local minima introduced by the energy function. Each series of automated refinement cycles included one cycle of temperature factor refinement, from which average B values were determined for the main chain and for the side chain in each residue. The final R-value is 0.276 for all reflections between 6.0 Å and 1.9 Å. Ignoring the reflections for which $2(||F_o| - |F_c||)/(|F_o| + |F_c|) > 1.2$ the value decreases to 0.256. The summed energy terms for bond lengths, bond angles and torsion angles are

56.3, 322.6 and 478.3 kcal/mol, respectively, for 440 residues in the molecule.

Summarizing, only the use of phase-combined maps in the steps of manual interventions and the application of the Jack & Levitt refinement procedure, using restraints rather than constraints, made it possible to arrive at a reasonably well refined model. (Uncertainties still exist in the sequence at several locations, especially in the hypervariable loops).

4.2 Deoxy-erythrocruorin

In the second example further improvement of the already extensively refined structure of deoxy-erythrocruorin[16] at 1.4 Å resolution was achieved. RLSP refinement had been applied resulting in a model with R-values of 0.185 and 0.170 at the resolutions of 1.4 Å and 1.2 Å respectively. The energy terms calculated from this model were 233, 246, 222 and -1301 kcal/mol for the bond lengths, bond angles, torsion angles and non-bonded contacts, respectively. Considering the small size of the molecule (136 residues), these values are rather high. This derives in part from some improbable geometric parameters in the standard groups used (e.g. systematically too small bond length for carboxyl groups, large bond angle in carboxyl of Glu, small bond angle C^{α}-C^{β}-C^{γ} in isoleucine, and slightly non-planar phenyl rings due to accumulation of errors). It should be noted that the geometric parameters for the various amino acids input to Diamond's model-building program[4] were based on rather old (mid-1950's) crystallographic data. These observations nonetheless illustrate a point made before, that restoration of distorted geometry is not possible in RLSP.

In total, four cycles of crystallographic refinement with energy restraints were performed setting the factor k to 4×10^{-5} in cycles 1 to 3 and to 5×10^{-3} in the last cycle. The shifts obtained from X-ray diagonal least squares were multiplied by a factor of 2. The RMS (root mean square) displacement between initial (after RLSP) and the last model was 0.2 Å (including 180° rotations of the far ends of Asn, Glu, His side chains). The R-value decreased to 0.169 and 0.151 for 1.4 Å and 1.7 Å data, respectively. The potential energy terms at the same time adopted values of 12 kcal/mol (bond lengths), 70 kcal/mol (bond angles), 150 kcal/mol (torsion angles) and -1470 kcal/mol (non-bonded interaction). The total energy decreased by the large amount of -1247 kcal/mol, reflecting the high weight given the energy term in the combined refinement. The effect of the weighting scheme is seen further in the observations that the final average values of bond lengths coincide with the target values and that the variances in bond lengths are less than 0.01 Å (for details see table 1).

Table 1

Bond lengths and bond angles of deoxy-erythrocruorin before and after simultaneous refinement of X-ray and energy residual (EREF).

Bond lengths (Å)

type	no.	before EREF	after EREF	target value	force const
NC (N-C^{α})	149	1.470±.008	1.469±.006	1.468	252
CC (C^{α}-C^{β})	243	1.527±.017	1.541±.005	1.540	288
CA (C^{α}-C')	161	1.526±.014	1.527±.004	1.525	468
AO (C=O)	178	1.233±.020	1.256±.003	1.257	569
AN (C-N)	153	1.320±.018	1.319±.007	1.318	230
BN (C-N His)	13	1.352±.041	1.405±.006	1.405	180
AA (C-C Phe)	67	1.390±.041	1.389±.006	1.389	288
BC (C^{β}-C^{γ}Phe)	29	1.511±.015	1.540±.004	1.540	468

Bond angles (degrees)

type	no.	before EREF	after EREF	target value	force const
ANC	138	121.0±1.4	121.0±0.4	123.8	51
CCN	133	109.7±1.7	110.4±1.3	110.1	65
CCC	145	112.3±3.5	111.9±1.4	111.9	66
ACN $\tau(C^{\alpha})$	136	113.1±4.2	112.4±1.5	113.1	63
ACC	145	109.1±2.9	109.7±1.4	109.0	67
OAO	17	122.8±4.9	119.0±0.7	119.0	107
OAN	144	123.6±1.1	123.8±0.7	124.1	57
CAO	178	120.5±2.7	119.8±0.9	119.7	69
CAN	144	115.1±1.0	116.0±1.0	115.1	56

Explanation of atom types: C tetravalent carbon; A trivalent carbon in 6-membered rings or carbonyl carbon; B trivalent carbon in 5-membered rings; O carbonyl and carboxyl oxygen; N amide and imidazole nitrogen. Force constants in kcal/mol.

As expected, the RMS deviations from the mean values are inversely proportional to the force constants of the various bond types. Similar observations are made for the other parameters (cf. bond angles in table 1). It is worth noting that the spread around the average values has been narrowed considerably during combined refinement. Despite this fact, improvement of the crystallographic R-value by more than 0.015 has been obtained. Undoubtedly, with less restraint (the present restraints must be regarded as very stringent), further improvement could be achieved. The interaction of R-factor and restraints seen here and in other examples stresses the need for the specification of a quantity describing the flexibility in the geometric parameters apart from the crystallographic R-value. The energy values are very useful, in this regard. The variation in bond lengths and bond angles and the deviation of their average values from target values seem also to be reasonable measures.

5. CONCLUSIONS

These findings underline the superiority of restrained refinement over constrained refinement, even in the case where no serious errors were present in the model.

It is clear that pure energy refinement using a given set of potential parameters will not result in a model agreeing very well with the observed crystallographic data. Nevertheless, these potentials, even in the present from (ignoring hydrogen atoms, hydrogen bonds, electrostatic interactions etc.), are good enough to act as reasonable restraints when combined with crystallographic refinement.

Summarizing, the following major advantages of the Jack & Levitt refinement have been found[17]:

1. Refinement at low resolution ($d_{min} > 3$ Å) is possible.

2. Refinement of temperature factors at medium resolution is possible.

3. Geometric restraints can be relaxed temporarily leading to better convergence without traps in local minima.

4. Distorted geometry can be easily repaired.

5. Treatment of non-polypeptide chains (sugar in glycoproteins) or any non-protein constituent (e.g. heme group and ligand in hemoglobin) is easy.

6. Computing time requirements are considerably reduced in comparison with Diamond's real space refinement.

6. ACKNOWLEDGEMENT

The author wishes to thank Dr. R. Huber for supporting this work and for his continuous encouragement. Drs. J. Deisenhofer and M. Marquart for fruitful discussions and for kindly providing their programs and results. Dr. W.S. Bennett made valuable suggestions which helped to improve the manuscript. The technical assistance of Mrs. K. Epp and Mrs. R. Sergeson in preparation of the manuscript is also gratefully acknowledged.

REFERENCES

1. K.D. Watenpaugh, L.C. Sieker and L.H. Jensen, J. Mol. Biol. 131, (1979) 509.

2. R. Diamond, Acta Crystallogr. Sect. A 27, (1971) 435.

3. J. Deisenhofer and W. Steigemann, Acta Crystallogr. Sect. B 31, (1975) 238.

4. R. Diamond, Acta Crystallogr. 21, (1966) 253.

5. M. Marquart, J. Deisenhofer, R. Huber and W. Palm, J. Mol. Biol. 141, (1980) 369.

6. W.A. Hendrickson and E.E. Lattman, Acta Crystallogr. Sect. B 26, (1970) 136.

7. R. Diamond in Crystallographic Computing Techniques, ed. F.R. Ahmed (Copenhagen: Munksgaard, 1976), 291.

8. R. Diamond, J. Mol. Biol. 82, (1974) 371.

9. T.A. Jones, J. Appl. Cryst. 11, (1978) 268.

10. J. Hermans Jr. and J.E. McQueen Jr., Acta Crystallogr. Sect. A 30, (1974) 730.

11. M. Levitt, J. Mol. Biol. 82, (1974) 393.

12. J.H. Konnert, Acta Crystallogr. Sect. A 32, (1976) 614.

13. A. Jack and M. Levitt, Acta Crystallogr. Sect.A 34, (1978) 931.

14. M. Marquart, Ph. D. Thesis, Munich (1980).

15. G.A. Sim, Acta Crystallogr. 12, (1959) 813.

16. W. Steigemann and E. Weber, J. Mol. Biol. 127, (1979) 309.

17. J. Deisenhofer, Biochemistry (in press).

18. W. Steigemann in Interaction Between Iron and Proteins in Oxygen and Electron Transport, ed. C. Ho (New York: Elsevier North-Holland Inc., in press).

ON THE RELATIONSHIP BETWEEN X-RAY AND ENERGY REFINEMENT

by

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

ABSTRACT

This paper reviews the principles involved in mixed energy and X-ray refinements and, in particular, addresses the question of their proper relative weights. The problem is viewed in terms of a priori and a posteriori probabilities and related to Boltzmannian statistics.

1. INTRODUCTION

A number of refinement procedures for macromolecules now exist which minimise some linear combination of X-ray and energy residuals. Yet others exist which minimise X-ray and energy residuals in alternation and as separate operations. What relative weights should be given to these two terms, and in the latter case, which of the two alternating procedures should be regarded as giving the final result?
The approach taken here to these questions is to view the energy function employed as being essentially a statement concerning the a priori probability distribution for the structure. Seen in these terms, it is more important that the function used should fairly represent a priori probabilities than that it should represent the highest sophistication in quantum mechanics, though hopefully, of course these two considerations would coincide. X-ray measurements then modify our original assessment of the probabilities.

1.1 Combining Probability Densities

Let x be an n-dimensional column vector of parameters describing a structure and let an a priori probability density distribution for x be

$$p_1(x) = K_1 \exp -\tfrac{1}{2}(x-x_1)^T M_1 (x-x_1) \qquad (1)$$

This density is distributed normally about $x = x_1$ with M_1 real symmetric positive definite, and the normalising constant K_1 satisfies

$$\int p_1(x)\, d\tau = 1$$

τ being the n-dimensional hypervolume.

Suppose now that an experiment is done yielding an independent estimate of the distribution of x

$$p_2(x) = K_2 \exp -\tfrac{1}{2}(x-x_2)^T M_2 (x-x_2) \qquad (2)$$

then the joint probability over both these considerations is proportional to the product

$$p(x) \propto p_1(x) p_2(x) \qquad (3)$$

$$\text{or } p(x) = K \exp -\tfrac{1}{2}\{(x-x_1)^T M_1 (x-x_1) + (x-x_2)^T M_2 (x-x_2)\} \qquad (4)$$

but with

$$K \neq K_1 K_2 \qquad (5)$$

This inequality arises because the product of two densities is not a density, and K will be found later by normalisation. The distinction between $p(x)$ and the product $p_1(x)p_2(x)$ is further discussed in an Appendix.

Equation 4 also represents a normal distribution about a point $x = x_o$ which may be found by a shift of origin. We replace x by $(x-x_o)$, x_1 by (x_1-x_o) and x_2 by (x_2-x_o), (which leaves quantities like $(x-x_1)$ unchanged) and eqn. 4 becomes

$$\begin{aligned} p(x) = K \exp -\tfrac{1}{2}\{ & (x-x_o)^T M_1 (x-x_o) - (x-x_o)^T M_1 (x_1-x_o) - \\ & (x_1-x_o)^T M_1 (x-x_o) + (x_1-x_o)^T M_1 (x_1-x_o) + (x-x_o)^T M_2 (x-x_o) \\ & - (x-x_o)^T M_2 (x_2-x_o) - (x_2-x_o)^T M_2 (x-x_o) + \\ & (x_2-x_o)^T M_2 (x_2-x_o)\} \end{aligned} \qquad (6)$$

in which the exponent now contains both linear and quadratic terms in x. However, x_o may be chosen to make the linear terms vanish, i.e. by setting

$$(x-x_o)^T M_1 (x_1-x_o) + (x-x_o)^T M_2 (x_2-x_o) = 0 \qquad (7)$$

which becomes true for <u>all</u> vectors x only if

$$M_1 (x_1-x_o) + M_2 (x_2-x_o) = 0 \qquad (8)$$

giving

$$x_o = (M_1+M_2)^{-1} (M_1 x_1 + M_2 x_2) \qquad (9)$$

With this value of x_o eqn. 6 becomes

$$p(x) = K' \exp -\tfrac{1}{2}(x-x_o)^T (M_1+M_2)(x-x_o) \qquad (10)$$

because the terms linear in x in eqn. 6 now vanish and those independent of x produce a factor which has been absorbed in the scale factor.

Evidently $p(x)$ maximises at $x = x_o$ and from eqn. 9 we see that x_o may be regarded as a weighted mean of x_1 and x_2 with the matrices M_1 and M_2 serving as weights, although, of course, they are not scalar weights. Figure 1 illustrates this situation in which the full lines are contours of $p_1(x)$ and $p_2(x)$ and the dotted line is a contour of $p(x)$. The two sets of full lines have a common tangent at x_o and at other points on a locus connecting x_1 and x_2 via x_o. Note especially that x_o is not on the straight line from x_1 to x_2 so that, in general, a set of coordinates which is a linear combination of X-ray-minimised and energy-minimised coordinates (i.e. lying on the straight line x_1 x_2) is not an optimal set.

Finally, we remark that the normalisation factor K for a distribution of the form

$$K \exp -\tfrac{1}{2}x^T Mx$$

is

$$\{(2\pi)^{-n}||M||\}^{\frac{1}{2}} \qquad (11)$$

in n dimensions.

2. APPLICATION

In order to apply the foregoing to the present problem we write

$$E-E_o = \tfrac{1}{2}(x-x_1)^T H(x-x_1) \qquad (12)$$

where H is the matrix of second derivatives of the potential energy E with respect to the components of x, x_1 being the minimum energy point at which grad E is zero. This requires that all forces be balanced out at $x = x_1$ but does not require that every constituent part should vanish. For example, a non-bonded inter-atomic separation typically produces an energy minimum at some distance, but it is not necessary that this non-bonded inter-action force should be zero at $x = x_1$ if it is balanced by a force arising, say, from a distorted bond angle. It is necessary only that there be no <u>net</u> force on any atom at $x = x_1$ and that the conformations actually encountered are sufficiently close to x_1 for <u>net</u> forces to vary linearly with displacement from x_1. I.E. eqn. 12 is Hookean.

With these assumptions x-space becomes the spatial part of a phase space and the Maxwell-Boltzman Law of Energy Distribution applies, namely that the probability that x lies within dx_1, dx_2, dx_3... dx_n is proportional to

$$e^{-(E-E_o)/kT}\, dx_1\, dx_2 \ldots dx_n$$

i.e.

$$p_1(x) = K_1 \exp -\tfrac{1}{2}(x-x_1)^T H(x-x_1)/kT \qquad (13)$$

from which also follows the equipartition law result that the mean energy $\bar{E}$ is given by

$$\bar{E}-E_o = \int (E-E_o)p_1(x)d\tau = nkT/2 \qquad (14)$$

For the X-ray case the experimental distribution $p_2(x)$ is given by

$$p_2(x) = K_2 \exp -\tfrac{1}{2}(x-x_2)^T D^T C^{-1} D(x-x_2) \qquad (15)$$

in which D contains the derivatives of the calculated values of the observables with respect to the parameters and C is the covariance matrix among the observations so that $D^T C^{-1} D$ is the correctly weighted normal matrix and x_2 is the experimental solution point. Thus the combined result, eqn. 9, may be obtained by setting $M_1 = H/kT$, $M_2 = D^T C^{-1} D$.

3. DISCUSSION

Boltzmann statistics describes the time-average of a fluctuating system, nevertheless it does, I think, provide the best relative weights of X-ray and energy considerations, and it allows us to address such questions as the extent to which a given coordinate set is dependent on X-ray measurements or assumptions concerning the energy function. It can also characterise the extent to which the two criteria are or are not in agreement.

For example, by a suitable change of variables, classical statistics tells us that for the probability density distribution

$$p(x) = \left(\frac{||M||}{(2\pi)^n} \right)^{\frac{1}{2}} \exp -\tfrac{1}{2}(x-x_o)^T M(x-x_o) \qquad (16)$$

the probability that $y = \tfrac{1}{2}(x-x_o)^T M(x-x_o)$ exceeds Y is

$$P(y>Y) = \int\!\!\int \ldots \int_{y>Y} p(x)dx_1 dx_2 \ldots dx_n$$

$$= \frac{1}{\Gamma(\frac{n}{2})} \int_Y^\infty y^{\frac{n}{2}-1} e^{-y} dy = e^{-Y} \sum_0^{\frac{n}{2}-1} \frac{Y^m}{m!} \qquad (17)$$

($\Gamma(\frac{n}{2})$ is defined as the value of this integral with $Y = 0$ and y is $\frac{1}{2}\chi^2$ in the usual statistical notation). Thus, by setting

$$Y = \tfrac{1}{2}(x_2-x_1)^T M_1 (x_2-x_1) \qquad (18)$$

eqn. 17 gives the thermodynamic probability that the

X-ray structure, x_2, should differ from the energy minimised structure, x_1, by at least the amount found, as expressed by eqn. 18. Similarly, the probability of the structure x_1 differing from x_2 by at least the amount found, given the observations, is obtained by replacing M_1 by M_2 in eqn. 18. Questions concerning the joint probability (eqn. 10) may be similarly treated.

Conversely the value of Y corresponding to

$$P(y>Y) = 0.01$$

is given [1] to a good approximation for large n by

$$Y = \tfrac{1}{2}n + 1.6447\sqrt{n} + 1.4701$$

which means that the solution, in x-space, is 99% certainly located within an ellipsoidal contour of hypervolume

$$\frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)} ||M||^{-1/2} (2Y)^{n/2}$$

for this value of Y and the n^{th} root of this would give the edge an equivalent cube. The point here is that this hyper-volume is proportional to $||M||^{-1/2}$, and it would be interesting to compare $||M_1||^{-1/2}$, $||M_2||^{-1/2}$ and $||M_1 + M_2||^{-1/2}$, the last of which may be quite small even when the other two are indefinitely large or infinite as is the case for structures which are undetermined by one criterion alone.

APPENDIX

On combining probability densities.

It was pointed out in connection with equation 4 that multiplying together the densities $p_1(x)$ and $p_2(x)$ gives a product proportional to but not equal to the combined density. Furthermore, these densities have dimension x^{-n} (so that their volume integrals are pure numbers) and their product of dimension x^{-2n} is clearly different, so what is the justification for multiplying densities?

Multiplication of probabilities (not probability densities) is appropriate for considering the simultaneous occurrence of independent events. Suppose two independent scalar variables x and y have probability densities p(x) and q(y) then the probability that x lies between x and x + dx and simultaneously y is between y and y + dy is p(x)q(y)dxdy in which p(x)q(y) is a superficial density on the x,y plane, p(x) and q(x) being themselves linear densities.

If additional information becomes available that necessarily x = y we may write

$$p(x,y)dxdy = p(x)q(y)dxdy = p(\xi,\eta)d\xi d\eta$$
$$= p(\xi|\eta)p(\eta)d\xi d\eta$$

in which $\xi = (x+y)/\sqrt{2}$. $\eta = (x-y)/\sqrt{2}$, and the linear density on the line $\eta = 0$ is given by setting $p(\eta)$ to a delta function at the origin, i.e. $\int p(\eta)d\eta = 1$ with $p(\eta) = 0$ unless $\eta = 0$. Integrating on η then gives the dependence on ξ as

$$d\xi . \int p(\xi|\eta)p(\eta)d\eta = p(\xi|0)d\xi$$

in which $p(\xi|0)$ is a linear density proportional to $p(x)q(x)$.

As an example of combining n-dimensional probability densities by multiplication and renormalising, consider the refinement of a crystal structure of n parameters with m observations, and let the observations be divided into two groups in some way. Any way will do provided that there are no covariances between observations in one group with any of those in the other. Refining the structure on group 1 alone produces a structure expressed by the n-dimensional vector x_1 and normal matrix M_1 based on these reflections only. Similarly, refining on the second group of reflections only yields x_2 and M_2. Refining on all reflections simultaneously produces a structure x_o related to x_1 and x_2 by eqn. 9 as is easy to verify. Thus the inclusion of the additional information represented by the second group of reflections is formally equivalent to multiplying probability density distributions in x-space and renormalising, as has been done in connection with eqn. 4.

REFERENCES

1. E.A. Cornish and R.A. Fisher. Moments and cumulants in the specification of distributions. Revue de l'Institut International de Statistique, iv, (1937) 1-14

or R.A. Fisher and F. Yates, Statistical Tables for Biological, Agricultural and Medical Research. Oliver and Boyd, London and Edinburgh.

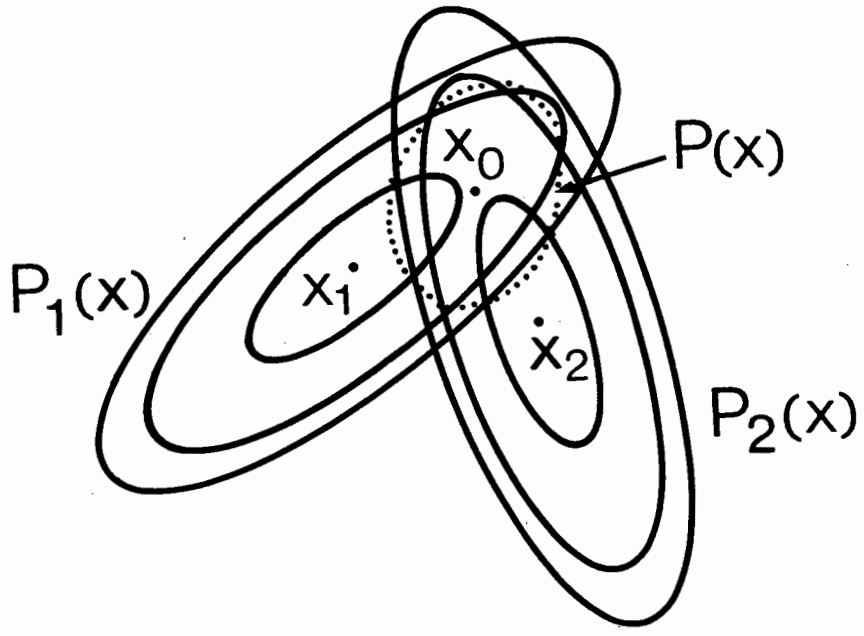


Figure 1. Combining two normal probability density distributions $P_1(x)$ and $P_2(x)$ can produce a maximum joint probability at x_o not co-linear with the maxima at x_1 and x_2.

SUMMARY OF THE MAIN DISCUSSION PERIOD

by

A.C. BLOOMER and P.R. EVANS

MRC Laboratory of Molecular Biology, Hills Road, Cambridge

1. INTRODUCTION

The main discussion period was chaired by Dr. D.S. Moss with the following topics on the agenda:-

a) What do you minimise? Constrained, restrained or...?
b) How do you calculate gradients?
c) Which weighting functions do you use?
d) What are the biases inherent in the methods?
e) How do you find whether you are right?

An introductory illustration from the Chairman summarised the various methods utilised by different procedures for the two constituent parts of any refinement process:- determination of the difference between observed and calculated x-ray terms; adjustment of the atomic model to minimise these differences. Most of the possible combinations have been used (see fig. 1) but it was generally agreed that, if one were now starting from scratch with no available programs, the preferred combination would be FFT calculation of structure factors with restrained shifts to individual atoms.

	Constd.	Restd.	Unrestd.
Real space density map	Diamond REAL SPACE		
Classical structure factor calcn.	Sussman	Konnert CORELS	Classical least squares (rubredoxin)
FFT calcn. of structure factors	Jack-Levitt		Agarwal (insulin)

Fig.1 Methods of refinement

Unrestrained shifts are only useful for structures at very high resolution (say 1.5 Å). As the resolution of a structure decreases, the restraints need to be increasingly hard until they become essentially rigid constraints. Current procedures using classical methods of structure factor calculation are generally easier to use than those using FFT methods but the advantage of the latter, arising from lower computing costs, increases markedly with the size of a structure. However, for the largest current structures, such as viruses, the cost of even an FFT structure factor calculation is such that it may prove more efficient to maximise the information derived from the initial map, as used for visual interpretation of the structure, by use of a real space procedure (Bricogne, personal communication afterwards). The case for this is greatly strengthened whenever the initial map is of particularly good quality as, for example, may result from the presence of a high degree of non-crystallographic symmetry. The practical advantage of the real space approach is the facility with which partial structures may be refined and although it has been used only with constrained parameters, this is not a necessary restriction upon the method.

2. HOW SHOULD THE RESTRAINTS BE SET?

The Chairman listed the different types of restraint for discussion in the following order:

a) Bonded distances

Exclusion of hydrogen atoms from most refinements distorts the centroid positions of atoms to which they are bonded, thus increasing the bond lengths. In a recent neutron diffraction study of myoglobin the H atoms were included producing a decrease of 0.01 Å in the mean C - C distance averaged over 150 bonds (S.E.V. Phillips).

b) Next nearest neighbour distances

c) Bond angle restraints

These two treatments of next nearest neighbour positions are strictly equivalent only when rigid constraints are applied. Hendrickson's program uses

the distances and Levitt's program uses the angles. Hendrickson expressed a preference for the use of angles, notwithstanding the distances having derivatives which are easier to handle in computing terms.

d) Torsion angle restraints

These are a computationally convenient means of ensuring planarity and of avoiding the need to include H atoms. Soft restraints should be used whenever the torsion angles are restricted, thus enabling movement away from false minima and the correction of gross errors.

Torsion angles are usefully calculated even if unrestrained. Hendrickson reported rms deviations of 12° for the peptide bond torsion angles in CRANBIN. Hermans indicated that for rubredoxin these angles deviated by 10-20° from planarity with a reasonable correlation between the values arising from an un restrained least squares refinement at very high resolution (beyond 1.5 Å) and those from an energy minimisation of the x-ray structure.

e) Planarity restraints

The possibility of real departures from planarity needs to be recognised but any relaxation of these restraints should be done with caution especially at low resolution or where the map may be distorted by thermal motion.

Methods of applying the restraints include

a) dummy atom at 20 Å
b) restrain normal to plane } restrain to be in current plane
c) stress the inter-atomic distances
d) improved force field
e) rigid body refinement (e.g. CORELS)

A suggestion attributed to Ten Eyck utilises the fact that the first derivatives of the energy are zero for all shifts normal to the plane.

f) Hydrogen bond restraints

Significant gains in enrgy may arise from H-bonds well over 3 Å in length. Thus the length should not be tightly restrained to a canonical 2.8 Å.

Referring to the work of a colleague in Groningen, Hol reported a prediction that non-planar peptides would show reduced H-bonding potential. This correlates with the resonant hybrid nature of the peptide bond involving partial charges on the N and O atoms.

g) Non-bonded restraints

Hendrickson chose to leave out the attractive potential, which could be dangerous, using a quartic approximation to the full Lennard-Jones potential which Levitt uses.

Levitt has now added to his program a protection against the ill-effects of the extremely strong repulsion at very small separations.

An electrostatic potential may be used within the Levitt program but the dielectric constant involved here is enormously uncertain.

h) Thermal parameters

However these are restrained, they must be recognised as attempts to account for a variety of inadequecies in the structural model, since no unique model can conform correctly to a dynamic density.

i) Non-crystallographic symmetry restraints (NCS)

One must beware of real departures from NCS especially in regions of contact between molecules related by crystallographic symmetry. Classical hypothesis testing allows comparison of the goodness of fits following refinement both with and without use of NCS restraints, but assumes that a correct weighting scheme is used for the observations.

3. CONSTRAINTS VERSUS RESTRAINTS AND CONVERGENCE

The evidence from small molecule studies suggests that hard constraints give a larger radius of convergence but also a tendency to stick in false minima as atoms are then prevented from moving through a bad contact to reach a correct position. The appropriate level of restraint is thus governed by the starting point which is, in turn, influenced by the following considerations. The appropriate resolution for a cycle of refinement, should be related to the features seen in the difference map and the state of the refinement. It can be estimated from the rms deviation in atomic positions. There was a strong difference of opinion as to whether any higher resolution data should be included, normally or with reduced weighting, or excluded and

the structure prevented from moving too far from the initial model by additional restraints.

The accuracy of prior fitting was also a contentious point with some supporters for adjusting the model in as much detail as possible and others for adjusting only to a point from which they think refinement can proceed, and at each cycle correcting only those features which the refinement program cannot cope with.

Regions of uncertainty within the model may be treated by a suitable choice of occupancy or temperature factor but this is not really satisfactory as one cannot properly use a single parameter to describe both the height or shape of density and its uncertainty. Omitting such regions entirely from the structure factor calculation works successfully for a few residues. However, several cycles of refinement may be needed after omitting them, in order to eliminate any compensating effects of neighbouring structure. Removing 1/8th of the structure at a time had worked well with insulin and allowed the residues causing concern to be studied in the same context as the remainder of that 1/8th of the molecule. There were no suggestions for an optimum procedure in cases where the uncertainty represents a large fraction of the model, e.g. no sequence data to provide side-chain assignments or uninterpretable density for, say, 30% of the molecule.

Difference maps should be calculated with combined phases (α_{MIR} and α_{CALC}) but various combinations of ($n\,F_o - F_c$) amplitudes are being used. For refinement of insulin ($1\frac{1}{2}\,F_o - F_c$) had been used as a compromise, giving shifts about 75% of those seen in a straightforward ($F_o - F_c$) map, compared with the 50% found with ($2\,F_o - F_c$).

The extent of self-correctingness of these refinement procedures was felt to be reasonable, but a cautionary tale was reported by Guy Dodson of an Isoleucine residue in insulin where there was no evidence of an error until two independently refined sets of coordinates were compared. With hindsight, a chiral volume check might have been useful in raising suspicion.

4. RELATIVE WEIGHTING OF X-RAY AND ENERGY TERMS

Hermans presented results of minimising R and wU for different relative weights w between the residuals of the x-ray (R) and energy (U) terms. His experience was that this function exhibited a clear elbow in the curve (see fig. 2), this point representing the optimum solution to the minimisation problem. There was no experience of predicting suitable values for the weighting parameter.

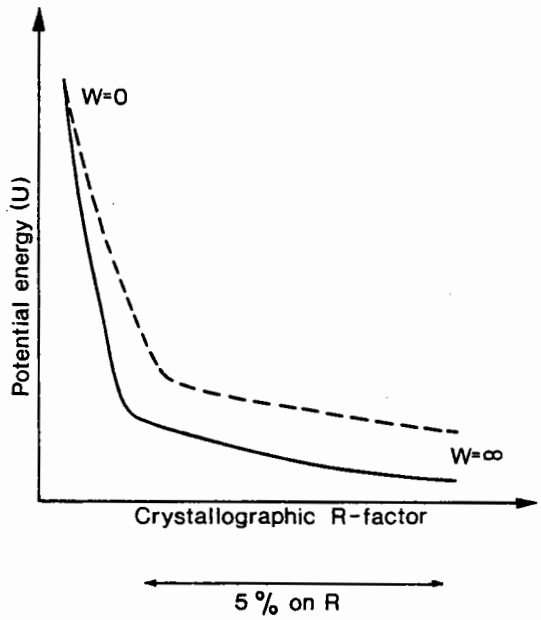


Fig.2 Effect of relative weighting parameter (w) on minimising [(x-ray terms) + w (energy terms)] is shown for two different energy functions U_A (——) and U_B (---).

MODELING DISORDERED PARTS OF PROTEIN CRYSTALS: A POSSIBLE AID IN CRYSTALLOGRAPHIC REFINEMENT

by

Jan Hermans
Department of Physical Chemistry, University of Groningen, Groningen, the Netherlands, and
Department of Biochemistry, School of Medicine, University of North Carolina, Chapel Hill, NC 27514, USA

1. INTRODUCTION

Crystallographic determination and refinement of molecular structure is commonly carried out in terms of a unique molecular model, i.e., a set of atoms, whose positions and thermal vibrational parameters (B's) are adjusted to improve the agreement between observed and calculated intensities. Such a model represents an ordered crystal, one in which all the molecules are confined to vibrations about a single conformation of minimum energy. The model can accommodate a modest amount of disorder: if (parts of) the structure correspond to a different minimum-energy conformation in some fraction of the molecules, then alternate positions may be specified, with variable occupancies (totaling unity).

This method has been applied with success to protein cyrstals. Two aspects are worth noting: (1) Except in cases where, for proteins, resolution is extremely high, success of the refinement is greatly enhanced by introduction of conformational constraints or restraints on the molecular model, and, (2) Refinement has not led to a comprehensive description of disordered parts of the crystal: solvent and some parts of the protein (usually long polar side chains) that are highly solvated[2-6].

Solvent typically occupies between one and two thirds of the volume of a protein crystal[1]; some solvent molecules are bound, in isolated groups of one or a few, inside, or between protein molecules, and these can easily be included in the unique molecular description used in the refinement. However, the greater part of the solvent occurs in volumes of several hundred $Å^3$ and more, in which (partly?) ordered water molecules are identifiable only near the surface of the protein. As a consequence, a considerable fraction of solvent, and hence of total crystal content, is not represented during the refinement.

One of the objectives of the study that we have undertaken, is to obtain a (theoretical) model of the disordered parts of the structure. Two possible uses of such a model are: (1) The properties of the model may suggest how solvent structure may best be represented and parametrized for crystallographic refinement; (2) Solvent structure in the refined model may be restrained not to differ radically from the theoretical model.

We have two methods available with which to simulate a statistical sample of conformations, from which an average model can be calculated. These are, the Monte Carlo method and the molecular dynamics method. These methods require an algorithm for calculation of the conformational energy and are related to energy minimization; however, both simulation methods perform minimization of the free energy. Use of a model of minimum free energy as a restraint in crystallographic refinement is a conceptually straightforward extension of the use of energetic (or quasi energetic) restraints, now common.

2. CRYSTAL OF PANCREATIC TRYPSIN INHIBITOR (PTI)

We have selected the crystal of this small protein for a variety of reasons: its structure has been refined[4,5,7]; it has a suitable space group; there are approximately 140 water molecules per protein molecule; results can be compared with those obtained by others in molecular dynamics simulations of the isolated protein molecule[10]. The high salt concentration of the crystals' mother liquor is possibly a drawback; salt content of the crystal is unknown, and neither do we know how the salt ions affect solvent structure. We do not represent any salt ions in the simulation.

The pH of the crystals is circa 10; the protein molecules have approximately zero net charge at this pH. We represent carboxylate ions and side chains of lysine and arginine with a full net charge of ± 1 electron, which gives a net positive

charge to the entire molecule, and hence to the contents of the asymmetric unit. In the crystal, ionization equilibria of lysine and tyrosine side chains reduce the average net charge to zero. Divergence of the calculated crystal energy, due to finite net charge of the unit cell, is avoided by use of a cutoff distance for the nonbonded interactions.

We work in the following, nearly cubic, asymmetric unit: $0 < x < a/2$, $0 < y < b$ and $0 < z < c/2$, and restrict our calculations to the contents of this volume and symmetry-related molecules surrounding it. (a=43.1, b=22.9, c=48.6 Å, volume of the asymmetric unit is 12,000 Å^3).

3. MONTE CARLO CALCULATION[8,9]

This simulation generates a set of conformations as a statistical sample of the equilibrium ensemble. Each consecutive conformation is generated from the one last added to the set by a random change; a new conformation is included in the set if its energy is lower than that of the one from which it was generated, or, if its energy is higher by ΔU, with probability equal to the Boltzmann factor $\exp(-\Delta U/kT)$.

In our application of this method to the PTI crystals[11], the conformation could change randomly in one of three ways: (1) One water molecule was moved and rotated within limits from its current position; (2) Part of a side chain was rotated about a single bond by a limited amount; (3) One water molecule was moved to a position anywhere within the asymmetric unit (jump). Initial positions of the protein atoms were those that had been obtained by crystallographic refinement. Initial placement of the 140 water molecules within the asymmetric unit was random.

4. MOLECULAR DYNAMICS CALCULATION

This simulation generates a (short) history of the movement of the system; thus, one obtains not only a sample of the equilibrium ensemble, but also information about time-dependent properties. The calculation requires the stepwise integration of the equations of motion. For any coordinate, x_{ij}, of a particle with mass m_i, the acceleration is related to the force component, F_{ij}, and the conformational energy, U, by

$$a_{ij} = d^2x_{ij}/dt^2 = F_{ij}/m_i = -(\partial U/\partial x_{ij})/m_i \qquad (1)$$

and velocities and positions are obtained by integrals

$$v_{ij} = dx_{ij}/dt = \int a_{ij}\, dt \text{ and } x_{ij} = \int v_{ij}\, dt. \qquad (2)$$

We have applied this calculation to the PTI crystal. Initial water positions for the dynamics simulation were obtained with the Monte Carlo method, modified to ensure that there is a water molecule in the model conformation, near each of the preferred water positions identified by crystallographic refinement (see below).

We have used the two simulation methods in a complementary manner, utilizing the advantages of each. With the Monte Carlo calculation an equilibrium distribution of water molecules is obtainable which depends on the coordinates of the protein but not on any information on preferred water positions obtained with x-ray crystallography. With the molecular dynamics calculation, movement of water molecules in space is restricted by time continuity and briefness of the simulated span. On the other hand, molecular dynamics is more readily utilized than the Monte Carlo method in order to simulate motion of all atoms, and, besides, produces time-dependent information.

5. ENERGY FUNCTION AND PARAMETERS

5.1 Protein.

For a computation of this magnitude, use of a simple energy function is mandated. In the function used in this study, the following terms are included: quadratic terms for deformation of bonded geometry; cosine terms to represent barriers to rotation about single bonds; Lennard-Jones and electrostatic pair potentials for nonbonded forces.

For two atoms, i and j, one has the following nonbonded potential

$$U^{nb}_{ij} = -A_{ij}/r_{ij}^{6} + B_{ij}/r_{ij}^{12} + \varepsilon_i\varepsilon_j/r_{ij} \qquad (3)$$

where r_{ij} is the interatomic distance, the ε are partial charges (if necessary, normalized for the effect of the dielectric constant). Partial charges and the attractive and repulsive parameters, A_{ij} and B_{ij}, depend on the type of the atoms i and j. Commonly A_{ij} and B_{ij} are obtained as products of two constants depending on a single atom type:

$$A_{ij} = \sqrt{A_{ii}} \cdot \sqrt{A_{jj}} \quad \text{and} \quad B_{ij} = \sqrt{B_{ii}} \cdot \sqrt{B_{jj}} \qquad (4)$$

Computation time is determined by the number of nonbonded terms. In an exploratory study, some loss of realism is acceptable, if the calculation can be shortened by a simplification. Thus, hydrogen atoms bonded to carbon have not been explicitly included, but CH, CH_2 and CH_3 groups have been represented as "united atoms".[12]

Interactions at a distance greater than a cutoff of 6 Å are neglected. The nonbonded energy is calculated by reference to a table of interactions which is not updated after every movement. The cutoff criterion is applied during construction of the table, but not when the table is used to calculate nonbonded energies. Dipolar groups (e.g., the peptide CO group) are carried into the table in their entirety. A new table is calculated every so many dynamics steps.

5.2. Water.

Simple energy functions that have been used for simulations of liquid water do not quite match the simple description of protein-protein interactions. When expressed in terms of pair interactions, the best known function, the ST2 model, used by Rahman and Stillinger[13-15], uses the positions of the oxygen masses as centers of Lennard-Jones 6-12 potentials, those of the hydrogen masses as positive charges, and those of two fictional atoms, without mass, as negative charges. The two hydrogen atoms and the two fictional atoms surround the oxygen with tetrahedral coordination, at 1.0 and 0.8 Å, respectively. The geometry of the molecule is rigid.

5.3 Modified water model.

A well-tested model of the water molecule with centers for attraction and repulsion at the three atom centers, alone, was not available at the start of this project. Thus, early calculations were done with the ST2 model of water-water interactions and a three-atom water model for calculation of protein-water interactions. Parameters for Lennard-Jones forces of the protein model had been obtained from analysis of crystals of small apolar molecules[16], while partial charges and hydrogen bond parametrization were those of Poland and Scheraga.[17,18]

Postma and Berendsen, in collaboration with the author, have recently investigated the possibility of using a three-center water model. An acceptable model must describe the interactions of a pair of water molecules in vacuo, and that of several molecules in a molecular dynamics simulation of the liquid. The simple model has two adjustable parameters: the oxygen-oxygen repulsive coefficient and the partial charge on the oxygen atom, the oxygen-oxygen attractive coefficient being known from the molecular polarizability. Furthermore, a special hydrogen-oxygen potential may be introduced (e.g., 8-12 or 8-9). Postma has obtained an acceptable model (PS1) with an 8-9 hydrogen-oxygen potential, and has recently been able to improve the model by making small adjustments of the parameters. The best model that he has so far obtained has no hydrogen-oxygen potential, other than the electrostatic (1/r) interaction (PS2).

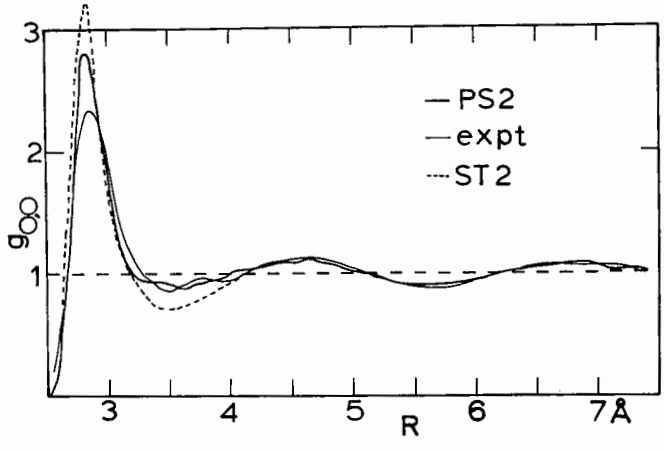


Fig. 1 Oxygen-oxygen radial distribution functions calculated, respectively, from experimental data (ref. 19), from molecular dynamics simulation with the ST2 model (Rahman and Stillinger, ref. 13-15) and from molecular dynamics simulation with the PS2 model (Postma, Berendsen and Hermans, unpublished).

Figure 1 shows the oxygen-oxygen radial distribution function obtained by molecular dynamics with the PS2 model, together with experimental results (Narten and Levy[19]) and with the simulated distribution for the ST2 model. One sees that the resulting approximation for PS2 is a good one, better than that obtained with ST2. Diffusion constant and pressure in the simulation are also good approximations. The dipole moment of the model is 2.33 Debye.

In order to model protein-water interactions in terms of the new model, it was found advisable to replace the Poland-Scheraga[17] charges with partial charges obtained by Hagler, Lifson and Dauber[20-22] in studies of crystals of small polar molecules. The force field obtained by these authors corresponds to a much more polar protein model, in line with the highly polar water model. By combination of the Hagler-Lifson-Dauber description of the protein with the new model for water, we have obtained a unified description of protein-protein, water-water and protein-water interactions.

One of the features of the new description is a high value of the repulsive Lennard-Jones parameters for oxygen and nitrogen. It is easy to see why: when the electrostatic force is high, and hydrogen bond equilibrium distance is the same, then oxygen-oxygen (or oxygen-nitrogen) repulsive force must also be high.

We have, so far, used a description based on the PS1 model, but intend to replace this with one based on the final "best" model that Postma will obtain. However, we believe that conclusions drawn so far on the basis of these results, are independent of further small adjustments of the model.

6. RESULTS

Results of a first Monte Carlo calculation, with jumps of water molecules, (which have been reported,[11]) were found to be decidely at odds with expectation. Simulated water structure was highly ordered and showed many preferred water positions which, by virtue of their confinement between protein molecules, should have been found even in a conservative interpretation of electron density maps during the crystallographic refinement process, but had, in fact, not been identified. A continuation of this Monte Carlo simulation for triple the number of steps did not alter this qualitative result. Further analysis of the results of these calculations showed that the excessive incidence of waters tightly confined by surrounding protein, was accompanied by the absence of water molecules in a portion of the largest water space.

Introduction of the new water model and the accompanying alteration of water-protein and polar protein-protein interactions, produced a very much different result. Most of the tightly confined water molecules "jumped" to new positions during a brief equilibration. Following equilibration adjustment to the new parameter set, the simulation was carried for 300,000 steps. Average water densities have been calculated for this run and for its central 100,000 steps (Figure 2).

The new results show improvement over the old ones on a number of important points. The number of water molecules in confined positions has dropped drastically; in fact, it is now too low. Many water molecules have moved to regions of less order; in particular, the large water space is now entirely filled. Over the shorter calculation of 100,000 steps, many waters appear to be ordered, according to the number of high density peaks in the simulated map. In the longer run of 300,000 steps, many of these maxima become indistinct and one observes in the larger spaces what is clearly highly disordered solvent. A number of the maxima that remain correspond to ordered water molecules identified during x-ray crystallographic refinement. However, the positions of some of the most strongly confined, and hence clearest, crystallographic water molecules are almost entirely avoided in the Monte Carlo simulation.

In Figure 3 is shown the distribution of peak heights in the simulated maps calculated over 100,000 and 300,000 steps. The corresponding scale of approximate crystallographic thermal parameters (B) shows that the distribution of maxima from the long run is qualitatively in agreement with experience from x-ray crystallography. An even longer Monte Carlo calculation is in progress, and will provide a test of statistical (in)sufficiency of these results.

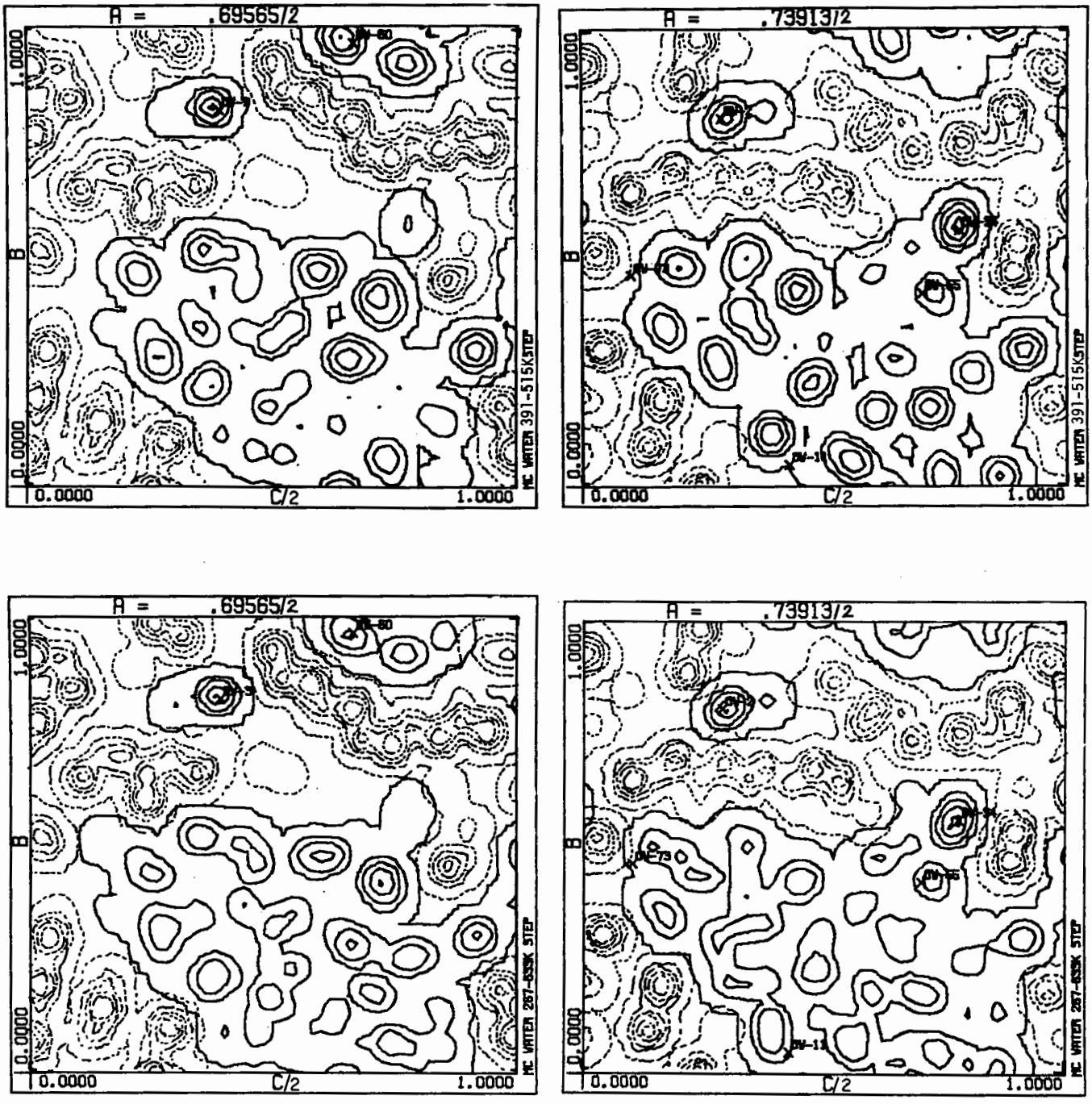


Fig. 2 Contour diagrams of density maps calculated as averages over, respectively, 100,000 and 300,000 Monte Carlo steps. (Each step consists of an average of 5 attempts at displacement.) Each non-hydrogen atom is represented as a density $\exp(-d^2/0.25)$ where d is the distance from the atom's position, and the sum of these Gaussians has been averaged over the configurations produced by the simulation. Contours were drawn at 0.0001, 0.016 (the average of the function in liquid water), 0.06, 0.17 and 0.34. The drawings are superpositions of two contour maps; solid contours represent water density, dashed contours represent protein density. (Most protein atoms do not move during the simulation, hence protein density is artificially sharp.) Crosses mark preferred solvent positions identified during crystallographic refinement (Ref. 5,7). Extent of agreement between crystallographic water positions and peaks in the simulated density observed in these two sections, is representative. (The sections are 1 Å apart; simulated density was calculated on a 0.5 Å grid.)

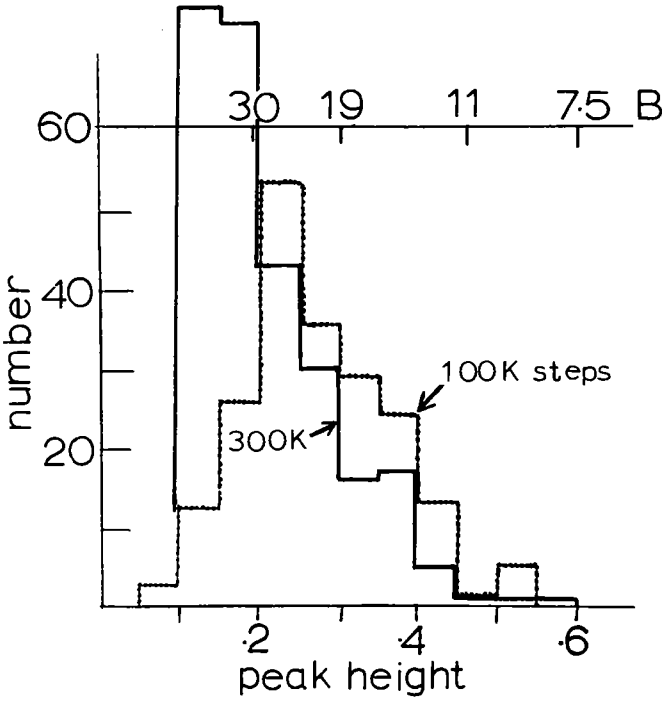


Fig. 3 Distribution of the peak heights of the maxima of the simulated densities of Figure 2. A scale of approximate thermal parameters (B) is given at top.

7. DISCUSSION

This article is in the first place a progress report. Recent progress described here, has, more than anything else, suggested a number of additional analyses and calculations. Some of these are briefly described in this discussion.

The model of water-water and water-protein interactions that we have used, appears to be biased against the occurence of water molecules in close contact with the protein. We have considered two possible reasons for this. (1) Water molecules in close contact with protein hydrogen-bonding groups are often also in close contact with apolar groups, e.g., CH_2. The large repulsive Lennard-Jones coefficients for water oxygen, introduced in order to obtain correct hydrogen bond distance, via eqn. 4 also give a high calculated energy for water-apolar group pair interactions. Analysis of crystal structures of relatively apolar molecules containing oxygen[16] does suggest use of a lower repulsive coefficent for (water) oxygen paired with non-hydrogen bonding groups. (2) One expects the distribution of water molecules within the asymmetric unit that results from the Monte Carlo simulation to be sensitive to the differences in the energies of water-water and water-protein hydrogen bonds. Since estimates of these differences that can be made on the basis of experimental data, are accurate to at best ± 1 kcal/mole, the energy model might be significantly biased in favor of water-water contacts without appearing unreasonable.

Monte Carlo simulations can be used to test both of these possibilities, and these calculations are in progress. We thus propose to adapt our energy model to fit the observed solvent distribution in one or more highly refined crystal structures, within the limits permitted by considerations of other available experimental data and of theory. Adaptation of the model will be required before Monte Carlo simulations can acquire predictive value in studies of solvent structure in the crystal of any protein.

We have begun a molecular dynamics calculation using the obtained energy model, with as a starting point the distribution of water molecules within the asymmetric unit obtained at the end of the Monte Carlo simulation. Since very large movements, equivalent to jumps of water molecules in the Monte Carlo simulation, are extremely improbable in a molecular dynamics trajectory, errors in the relative energies of different types of hydrogen bonds are <u>not</u> expected to lead to a significant error in equilibrium conformation. However, movement of all atoms of the protein (rather than movement of side chain atoms, alone) in molecular dynamics, might easily lead to undesirably large systematic displacements of protein atoms if any tightly confined water molecules were not in their appropriate places at the start of the simulation. A starting conformation with a water molecule near each crystallographic water position was easily produced by a brief continuation of the Monte Carlo simulation in which jumps were permitted only to crystallographically identified water positions. (Interestingly, this procedure proved fully effective only with use of a low repulsive coefficient for water-apolar interactions and after some adaptation of the water-protein and water-water potentials in favor of water-protein interactions).

Experience with the molecular dynamics calculation of this system has so far not led to useful results. A preliminary analysis indicates that a major problem arises from application of a distance cutoff to the non-bonded atom pairs considered, in particular, pairs of atoms (or atom groups) bearing non-zero net charge. Stratagems, some of them makeshift, that have been applied in other calculations in order to deal with charge-charge interactions, are inappropriate in a calculation in which the dielectric is represented explicitly, and, presumably, completely, in the model, in the form of polar solvent molecules. We are presently making some progress towards the development of a method for a proper calculation of the electrostatic energy and forces.

Calculation of Fourier transforms of simulated density is in progress. Results may indicate in a qualitative sense the relative importance of disordered, but not featureless, solvent areas in the Fourier transform of the entire structure. Comparison of calculated Fouriers with observed x-ray intensities will be one crucial test of the validity of simulated solvent distribution.

I am pleased to acknowledge the important contributions made during the first part of this study, at the University of North Carolina, by M. Vacatello. Current participants in this project include H. Berendsen, W. van Gunsteren, W. Hol and J. Postma, at the University of Groningen. This work has been supported by a research grant from the US National Science Foundation, a University of North Carolina Kenan leave, a fellowship from the Dutch research organization ZWO and by the University of Groningen computer facility.

8. REFERENCES

1. B. Matthews, Ann. Rev. Phys. Chem. 27, (1976) 493.

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

3. C.W. Carter, J. Kraut, S.T. Freer, W. Zuong, R.A. Alden, R.G. Bartsch, J. Biol. Chem. 249, (1974) 4212.

4. R.G. Huber, D. Kukla, A. Ruhlmann, O. Epp, H. Formanek, Naturwissenschaften 57, (1970) 389.

5. J. Deisenhofer and W. Steigemann, Acta Cryst. B31, (1975) 238.

6. K.D. Watenpaugh, T.N. Margulis, L.C. Sieker, L.H. Jensen, J. Mol. Biol. 122, (1978) 175.

7. W. Hendrickson, unpublished results.

8. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, J. Chem. Phys. 21, (1953) 1087.

9. A.T. Hagler and J. Moult, Nature 272, (1978) 222.

10. J.A. McCammon, B.R. Gelin, M. Karplus, Nature 267, (1977) 585.

11. J. Hermans and M. Vacatello, in "Water In Polymers", S. P. Rowland, ed., American Chemical Society, Washington, DC, 1980, p. 199.

12. B.R. Gelin, Ph.D. thesis, 1976, Harvard University.

13. A. Rahman and F.H. Stillinger, J. Chem. Phys. 55, (1971) 3336.

14. F.J. Stillinger and A. Rahman, J. Chem. Phys. 57, (1972) 1282.

15. F.J. Stillinger and A. Rahman, J. Chem. Phys. 60, (1974) 1545.

16. D.R. Ferro and J. Hermans, in "Liquid Crystals and Ordered Fluids", J.F. Johnson and R. Porter, eds. Plenum, New York, 1970; p. 259.

17. D. Poland and H.A. Scheraga, Biochemistry 6, (1967) 3791.

18. H.A. Scheraga, Adv. Phys. Org. Chem. 6, (1968) 103.

19. A.H. Narten and H.A. Levy, J. Chem. Phys. 55, (1971) 2263.

20. S. Lifson, A.T. Hagler and P. Dauber, J. Amer. Chem. Soc. 101, (1979) 5111.

21. A.T. Hagler, S. Lifson and P. Dauber, J. Amer. Chem. Soc. 101, (1979) 5122.

22. A.T. Hagler, P. Dauber and S. Lifson, J. Amer. Chem. Soc. 101, (1979) 5131.

A NOVEL TECHNIQUE TO IMPROVE THE QUALITY OF AN ELECTRON-DENSITY MAP

by

T. N. Bhat and D. M. Blow
Department of Physics (Biophysics), Imperial College of Science and Technology, London, SW7 2BZ

1. INTRODUCTION

In protein crystallography, the isomorphous replacement method becomes progressively more difficult to use, and the resulting phase angles less reliable, as resolution is extended. Once the point is reached where a detailed atomic interpretation of almost all the molecule can be made, more accurate phase angles may be obtained from calculated structure factors, and the way is opened for further refinement by a variety of techniques. The resolution may then be extended as far as diffracted intensities are observable, with a substantial increase in the accuracy of interpretation. Here I describe a technique for improving the existing set of phases when they fail to give a completely interpretable electron density map and hence none of the available refinement techniques can be safely used.

2. PROCEDURE

It is convenient to begin by describing the procedure used when a tentative atomic model has been made (this may well contain only main-chain coordinates without side-chain assignments). The procedure adopted without a model is described later.

The procedure includes seven steps which form an iterative cycle:

(1) Determination of 'occupancy' for each residue in the tentative model of the molecule, as observed in the current electron-density map.

(2) Calculation of electron-density for the tentative model, using the occupancies determined in step (1).

(3) Extraction from the current electron-density map of (a) features corresponding to the tentative model used in steps (1) and (2), or linked to it through regions of high electron-density (and (b), optionally, other electron-density features which are judged by their size and connectivity to correspond to real features of the structure), to form the extended model volume.

(4) Generation of the extended model density on a finer grid, including appropriate scaling of the model density and the electron-density map.

(5) Calculation of structure factors based on the extended model density.

(6) Modification of the phase information for each reflection, based on the calculated phase, pre-existing phase information derived from isomorphous replacement or any other source, and on the global agreement between observed and calculated structure factors.

(7) Calculation of a revised electron-density distribution, using these modified phases.

This new electron-density map is used for re-assignment of occupancies (step 1), and the whole procedure is iterated, using the same tentative model of the molecule. The process is found to converge within a few cycles, and then the latest revised electron density distribution forms the basis for building a new tentative model. If this model contains a significant number of new features, a further application of the whole iterative procedure may be made. When a starting model does not exist, steps (1), (2) and (3a) would be omitted. Such a method might be useful in the first interpretation of a relatively straight forward electron density map.

3. APPLICATION

As a test of the procedure, the method was applied to improve the electron density map of tyrosyl tRNA synthetase. It resulted in significantly improved phase angles, as judged by the peak/background ratio in a difference electron-density map for the binding of an inhibitor.

The procedure can be extended to include refinement of the partial structure. It involves (a) calculation of structure factors of structural features omitted from the model, but picked up in step 3(a), or 3(b), or both of these; and (b) refinement of the model after adding these structure factors to the contribution calculated from the model.

PHASE EXTENSION AND REFINEMENT AT 1.37Å RESOLUTION OF AVIAN PANCREATIC POLYPEPTIDE USING A MODIFIED TANGENT FORMULA

by

I. J. Tickle
Department of Crystallography, Birkbeck College, (University of London),
Malet Street, London W1CE 7HX

1. BACKGROUND

Avian Pancreatic Polypeptide (APP) is the third pancreatic hormone, after insulin[1] and glucagon[2], to be investigated by x-ray crystallographic techniques. The molecule consists of a single chain of 36 amino-acid residues. Zn^{2+} ion is essential for crystallisation, and is found to be incorporated into the crystal lattice. The structure has been solved at 2.04Å resolution by the isomorphous replacement technique, using a single $HgCl_2$ derivative which allowed measurement of both isomorphous and anomalous intensity differences to a resolution of 2.04Å. The details of this procedure have been reported elsewhere[3,4].

Intensity measurements were also made for the native APP crystal to a resolution of 1.37Å, and the possibility of extending the isomorphous replacement phases and figures of merit by a method not requiring prior structural information was investigated. Such methods include use of:

(i) The tangent formula[5].
(ii) The Sayre eqn.[6].
(iii) The Karle-Hauptman determinant?[7]

Initial attempts to extend the phases using the tangent formula, weighted according to the phase variance expected on the basis of independent contributors (ie as implemented in the MULTAN 78 program package[8]), led to the introduction of spurious noise peaks in the calculated electron density function, which were particularly noticeable on the crystallographic 2-fold axes. Recently, a modification of the weights used in the tangent formula, which was shown to be statistically equivalent to the Sayre equation, has been described[9], and this was made the basis of the method used for APP.

2. THE MODIFIED TANGENT FORMULA

The tangent formula can be expressed in the form:

$$\tan \phi_h = S_h/C_h$$

where $S_h = \Sigma_k K_{hk} \sin (\phi_k + \phi_{h-k})$

$$C_h = \Sigma_k K_{hk} \cos (\phi_k + \phi_{h-k})$$

$$K_{hk} = 2N^{-\frac{1}{2}} W_k W_{h-k} |E_h||E_k||E_{h-k}|$$

N = number of atoms in primitive unit cell

W_k = weight for reflection k

E_k = normalised structure factor for reflection k

The MULTAN program uses weights of the form:

$$W_h = \min (0.2\ \alpha_h,\ 1.0)$$

where $\alpha_h = (S^2_h + C^2_h)^{\frac{1}{2}}$

which being based on the assumption of independence of the contributors ϕ_k and ϕ_{h-k} to the tangent formula summations, leads to a gross underestimate of the variance, and hence an overestimate of the weight attached to ϕ_h.

It was thought that the use of the structure amplitude F rather than its normalised value E would give better convergence of the summations at a d-spacing at which the atoms are not completely resolved. (However substitution of E by F did not by itself significantly improve the results). The modification proposed involves use of a different weighting function W_h. For this purpose we define:

$$\alpha'_h = (S'^2_h + C'^2_h)^{\frac{1}{2}}$$

The prime (') will be used to indicate replacement of E by F in the tangent formula.

The basis of the method is to constrain $\alpha_h'^2$ to be equal to $\langle\alpha_h'^2\rangle$, its expected value in the absence of phase information, given, in a form suitable for computation, by:

$$\langle\alpha_h'^2\rangle = \Sigma_k K_{hk}'^2 \left[1 - \left(\frac{I_1(K_{hk})}{I_0(K_{hk})}\right)^2\right] + \left[\Sigma_k K_{hk}' \frac{I_1(K_{hk})}{I_0(K_{hk})}\right]^2$$

where I_1 and I_0 are hyperbolic Bessel functions. Note that $\langle\alpha_h'^2\rangle$ is a function of both F and E, and therefore both must be carried through in the computation of the phases.

It was shown[9] that the constraint $\alpha_h'^2 = \langle\alpha_h'^2\rangle$ is equivalent, in a statistical sense, to the constraints on the phase implied by Sayre's equation:

$$F_h = A\Sigma_k F_k F_{h-k}$$

The weighting function used is entirely empirical, giving reduced weight to reflections which deviate from the constraint. The function used is shown graphically in fig.1. and is given by:

$$f(x) = ke^{-x^2}\int_0^x e^{t^2}\,dt$$

where k is a normalising factor such that $f(x)_{max} = 1$ and $x = \alpha_h'/\langle\alpha_h'^2\rangle^{\frac{1}{2}}$. (This differs from the function originally recommended, in which $x = \alpha_h^2/\langle\alpha_h^2\rangle$. This was found to have undesirable statistical properties).

The isomorphous replacement phases were introduced into the tangent formula with weights equal to their figures-of-merit, and were kept fixed. If this was not done these phases rapidly diverged to a self-consistent but totally meaningless set with a centric distribution. (ie clustering around 2 values differing by π), corresponding to a single large maximum in the calculated electron density function, and little else.

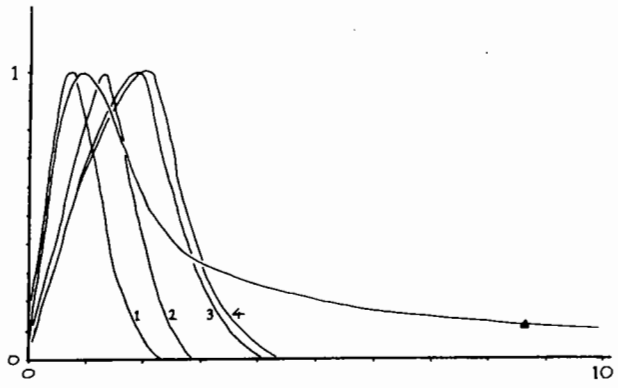


Fig.1 ▲ : Weighting function f(x) (abscissa = x)
1,2,3,4 : Normalised distribution N(x) after tangent refinement cycles 1,2,3 and 4.

Also shown in fig.1. is the progress of the phase refinement in the resolution range 2.04 to 1.37Å in terms of the normalised distribution of x.

One would anticipate that this distribution would resemble the weighting function, as indeed it does initially, but during the refinement it moves in the direction of increasing x while the phases converge smoothly after only 3 cycles. No further significant changes in the phases or weights were observed on continued recycling. This is in contrast with previous observations on tangent-formula refinement [10, 11, 12, 13], in which the phases initially converged to an essentially correct set, but subsequently 'blow up', corresponding to an accumulation of electron density at one place.

The rapid convergence was aided by a device which damped large changes in the phase or weight. The weighted structure factor for the nth cycle was computed as:

$$\left(W|F|e^{i\phi}\right)_n = \frac{\left(\left(W^2|F|e^{i\phi}\right)_{n-1} + \left(W^2|F|e^{i\phi}\right)_{calc}\right)}{\left(W\right)_{n-1} + \left(W\right)_{calc}}$$

where $\left(\ \right)_{n-1}$ and $\left(\ \right)_{calc}$ refer respectively to the value for the (n-1)th cycle and the value for the nth cycle calculated before the damping is applied. The new weight and phase for any given reflection do not take effect immediately, but only on conclusion of the current cycle.

The computation of 2738 phases with $|E|>0.7$ involved

almost 8 x 10^6 triples (h, k, $h-k$) and required approximately 12 mins cpu time for list generation and 10 mins cpu time per refinement cycle on the Daresbury Laboratory IBM 370/165.

Table 1 shows the distribution of reflections in terms of the final calculated weights, and table 2 the distribution of weights in resolution ranges. (Corresponding statistics are also shown for the 1960 phases determined by isomorphous replacement, for comparison). It is to be noted that the high-order reflections have significant weights, and will have an appreciable contribution in an electron-density function computed with Fourier coefficients $W_h|F_h|\exp(i\phi_h)$.

Table 1

Distribution of reflections in the resolution range 2.04 to 1.37Å, according to calculated weight.

Weight	No. of reflections	Mean $\|F_{obs}\|$
0 to 0.1	2	164
0.1 to 0.2	6	143
0.2 to 0.3	67	98
0.3 to 0.4	336	130
0.4 to 0.5	432	159
0.5 to 0.6	422	178
0.6 to 0.7	300	185
0.7 to 0.8	307	184
0.8 to 0.9	303	178
0.9 to 1.0	563	169

Table 2

Distribution of reflections and mean weight according to resolution range.

Resolution	No. of reflections	Mean $\|F_{obs}\|$	Mean weight
∞ to 4.75	152	565	0.83
4.75 to 3.02	465	562	0.84
3.02 to 2.51	463	374	0.82
2.51 to 2.23	441	323	0.79
2.23 to 2.04	439	267	0.82
2.04 to 1.90	274	267	0.72
1.90 to 1.80	269	242	0.68
1.80 to 1.71	301	204	0.68
1.71 to 1.64	283	173	0.68

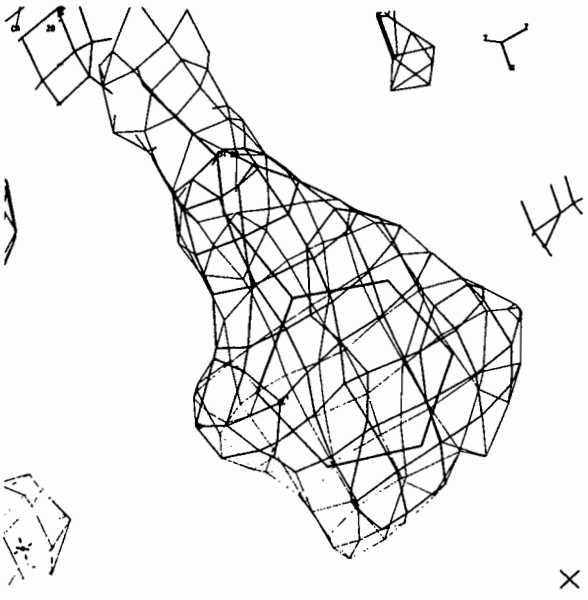
Table 2 (continued).

Resolution	No. of reflections	Mean $\|F_{obs}\|$	Mean weight
1.64 to 1.58	267	157	0.65
1.58 to 1.53	294	145	0.64
1.53 to 1.48	295	133	0.64
1.48 to 1.44	291	120	0.64
1.44 to 1.40	284	107	0.61
1.40 to 1.37	180	101	0.61

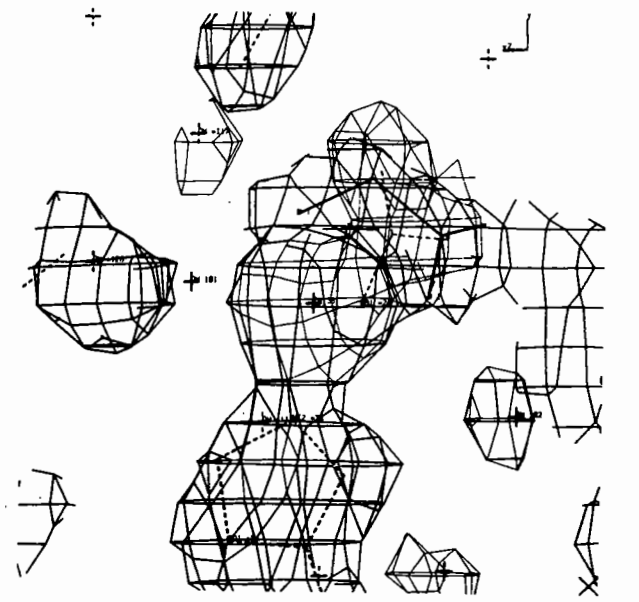
Phases and weights of reflections above the line in the table were determined by the isomorphous replacement method; those below the line (with E>0.7) by the modified tangent formula method.

RESULTS AND DISCUSSION

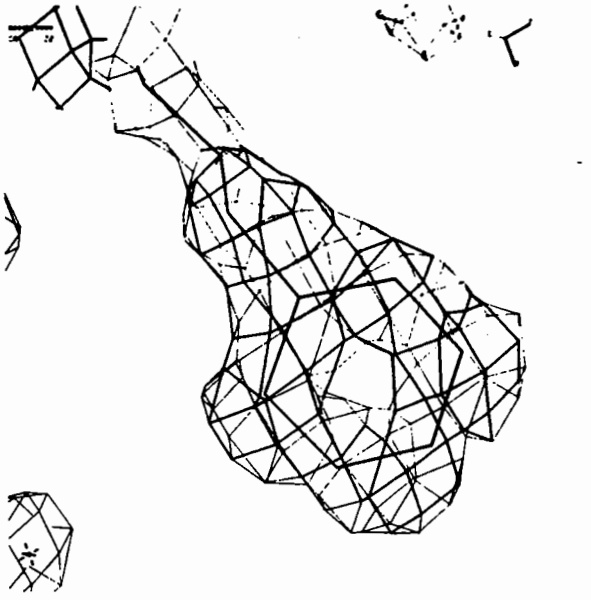
Figs.2a, 2b and 3a, 3b compare portions of the computed electron density with the original 2.04Å map. In general there is an observable improvement in resolution, without any loss of information or appearance of additional spurious density. However, one cannot say that the procedure has provided new structural information for the peptide. Significant changes in the solvent structure in the neighbourhood of the Zn^{2+} ion can be seen (figs. 3a, 3b), which are confirmed by a restrained-atom least-squares refinement currently in progress.

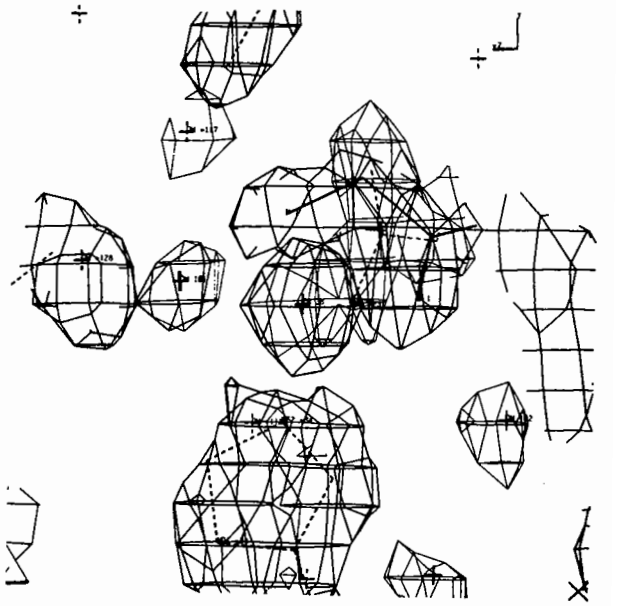


Figs. 2a and 2b Electron density corresponding to a) isomorphous replacement phases at 2.04Å (above) and b) tangent formula phases at 1.37Å, for PHE20.



Figs. 3a and 3b Electron density corresponding to a) isomorphous replacement phases at 2.04Å (above) and b) tangent formula phases at 1.37Å, in the region of Zn^{2+}





REFERENCES

1. Blundell, T. L., Cutfield, J. F., Dodson, G. G., Dodson, E., Hodgkin, D. C., Mercola, D. and Vijayan, M., Nature 231, (1971) 506-511.

2. Sasaki, K., Dockerill, S., Adamiak, D., Tickle, I. J. and Blundell, T. L., Nature 257, (1975) 751-757.

3. Wood, S. P., Pitts, J. E., Blundell, T. L., Tickle, I. J. and Jenkins, J. A., Eur. J. Biochem. 78, (1977) 119-126.

4. Pitts, J. E., Blundell, T. L., Tickle, I. J. and Wood, S. P., Proc. Am. Peptide Symp. 6, (1980) 1011-1016.

5. Karle, J. and Karle, I. L., Acta Cryst. 21, (1966) 849-859.

6. Sayre, D., Acta Cryst. 5, (1952) 60-65.

7. de Rango, C., Mauguen, Y. and Tsoucaris, G., Acta Cryst. A31, (1975) 227-233.

8. Main, P., "MULTAN 78. A Program System for the Automatic Solution of Crystal Structures from X-ray Diffraction Data." Univ. of York, England (1978).

9. Hull, S. E. and Irnin, M. J., Acta Cryst. A34, (1978) 863-870.

10. Coulter, C. L., Acta Cryst. B27, (1971) 1730-1740.

11. Weinzierl, J. E., Eisenberg, D. and Dickerson, R. E. Acta Cryst. B25, (1969) 380-387.

12. Reeke, G. N. and Lipscomb, W. N., Acta Cryst. B25, (1969) 2614-2623.

13. Blundell, T. L. and Johnson, L. N., "Protein Crystallography", Academic Press, London (1976), p. 438.

RESULTS AND COMPUTATIONAL ASPECTS OF REFINEMENT ON THE CRAY-1

by

Mr. W. Pulford

Laboratory of Molecular Biophysics, University of Oxford, South Parks Road, Oxford OX1 3PS.

ABBREVIATIONS

ρ - the electron density at Cartesian co-ordinates (x,y,z).

F_o - observed structure factor.

σ_{obs} - standard deviation of an observed structure factor derived from counting statistics.

$|F_c|$ - calculated structure factor magnitude.

α_{calc} - calculated phase of F_c.

α_{iso} - phase of F_o calculated from the isomorphous replacement method.

R-factor - the reliability factor for a specified set of data

$R = \sum_{1}^{m} \left| |F_o| - |F_c| \right| \Big/ \sum_{1}^{m} |F_o|$ m is the number of reflections in the set.

Output R-factor - the R-factor calculated on that set of data used in the refinement cycle.

B-factor - the temperature factor; defined for an atom i as $Bi = 8\pi \bar{\nu}_i^2$ where $\bar{\nu}_i$ is the RMS displacement of that atom.

CPU time - computer central processor time.

TL - Tortoise egg-white lysozyme.

1. INTRODUCTION

Many proteins have been or are currently being refined by Oxford workers. These refinements have been facilitated by the use of the Hendrickson-Konnert restrained stereochemistry structure factor least squares program. Table 1 gives details of these refinements:

Table 1

Protein	Resolution	No. Refls.	No. Atoms	Space Group	R-value now	Prole
Tortoise Egg-White Lysozyme (TL)	1.6Å	19700	1127	$P2_12_12_1$	22.3	Bill Pulford
Hot Hen Egg-White Lysozyme	2.0Å	7000	1001	$P2_12_12_1$	27.0	Dr. Peter Artymiuk
Human Leukaemic Lysozyme (HL)	1.5Å	18600	1173	$P2_12_12_1$	18.7	Dr. Peter Artymiuk
Seal Myoglobin	2.5Å	6300	1260	C2	33.6	Dr. Helen Scouloudi
Human Serum Prealbumin	1.8Å	24000	1966	$P2_12_12$	18.5	Dr. Stuart Oatley
Horse Muscle Phosphoglycerate Kinase (PGK)	2.5Å	13000	3200	$P2_1$	21.9	Dr. David Rice
E.Coli Arabinose Binding Protein (ABP)	1.9Å	24000	2300	$P2_12_12_1$	26.0	Dr. David Rice
Chicken Triose Phosphate Isomerase (TIM)	2.5Å	15000	3742	$P2_12_12_1$	27.0	William Taylor
Rabbit Muscle Phosphorylase (PPB)	3.0Å	17500	6519	$P4_32_12$	34.4	Dr. David Stuart
Hen Egg-White Lysozyme (HEWL)	2.0Å	8900	1100	$P4_32_12$	18.6	Drs. Diana Grace & Peter Artymiuk

2. METHODS OF REFINEMENT AVAILABLE AT OXFORD

The CPU times given below are for a protein crystallizing in space group $P2_12_12_1$ containing 1100 atoms, and for which there are 13500 reflections to 1.8Å, using an ICL 2980 computer (similar speed to IBM 370/165).

(i) Difference Fourier shifts (program FOURSHIFTS) involve the calculation of an $(|F_o|-|F_c|)\exp i\alpha$ map, where α may be α_{calc}, α_{iso} or some combination of both. The atomic shifts (Δx) are calculated using[1]:

$$\Delta x = -(\partial\rho/\partial x)/(\partial^2\rho/\partial x^2) \qquad (1)$$

The method is quick, requiring about 200 sec. using a reverse Fourier transform structure factor calculation (or 2000 sec. using a conventional structure factor calculation) but has been found to be decreasingly effective as the R-factor enters the 0.2 to 0.3 range. Another problem with this method is that considerable manual intervention (i.e. rebuilding the model, preparing and submitting jobs to the computer) is required in order that refinement can proceed.

(ii) The Isaacs-Agarwal program is a fast Fourier structure factor least squares program[2] whose cycle time is about 300 seconds. The method is more powerful than (i); refinements using this program have been performed to final R-factors below 20%[3].

Neither of the above methods takes account of the known stereochemistry of protein structures, and consequently they tend to produce stereochemically unreasonable models. Normally a regularizing process (program MODELFIT[4]) follows each refinement cycle using the methods described above, which despite measures like setting the weight on an atom's position inversely proportional to its B-value, reverses many of the shifts calculated in the refinement cycle. The net result of this is the necessity for a large number of refinement cycles in order to achieve a convergence.

(iii) The Hendrickson-Konnert restrained stereochemistry structure factor least squares program (referred to hereafter as RLSQ)[5] overcomes the above problem by jointly minimizing the disagreement between F_o and F_c and the deviation from ideal stereochemistry of the model, by least squares. The inclusion of stereochemical information increases the number of observations of physical quantities, thereby at least in theory, enabling the meaningful refinement of protein models at a lower resolution than hitherto possible by straightforward least squares. RLSQ uses a conventional structure factor and derivative routine at present, this results in the cycle time being 2200 seconds.

3. TYPES OF ELECTRON DENSITY MAPS USED IN REBUILDING

(i) An 'omit' map is an $(|F_o|-|F_c|)\exp i\alpha_{calc}$ map but with the contributions to F_c and α_{calc} of the atoms under investigation subtracted. The intention of the process is to produce an electron density map with minimum bias to the input coordinates, where only calculated phases are available. The resolution evident in the map deteriorates as more atoms are 'missed out', although the more correctly placed atoms included in the calculation of the map, the better this resolution at each stage. For TL, the calculation of the map was organized so that 1/8th of the molecule was missed out during the production of those sections containing the omitted atoms.

(ii) Much use of $(3|F_o|-2|F_c|)\exp i\alpha_{calc}$ maps was made during the refinement of TL. This is an example of a $(m|F_o|-n|F_c|)\exp i\alpha_{calc}$ map; the rules governing the appropriate choice of m and n are discussed by Vijayan[6].

4. FACTORS AFFECTING THE AGREEMENT BETWEEN OBSERVED AND CALCULATED STRUCTURE FACTORS

Figure 1 illustrates how the agreement between observed and calculated structure factors varies with resolution. The poor agreement at low resolution (Region A) is due primarily to the absence of an adequate description of the disordered solvent contained in most protein crystals. The gradually worsening agreement between F_o and F_c with increasing resolution (Region B) is the result of a combination of many effects: for example, the data are generally weaker at higher resolution and are consequently often unavoidably less well-measured than at lower resolution, the higher resolution terms

are also more sensitive to slight errors in the model.

5. THE REFINEMENT OF TORTOISE LYSOZYME (TL)

This protein was the first to be refined by an Oxford worker, the author, using the RLSQ program. The progress of the refinement and the speeding up of the program have been concurrent since the first cycle was run in early 1978.

The refinement progressed in the four stages set out below.

Stage I - Weighting used - $1/\sigma^2$
Map type - 'Omit'
No. parameters - 3026 (An overall temperature factor was refined)

Table 2

Progress of Refinement of TL; Stage I

No. Cycles	Resolution	No. Refls.	No. Atoms	Overall Start	R-Value Finish	Output Start	R-Value Finish
4	10.0-3.0Å	3054	1008	53.3		50.1	37.0
1	10.0-2.7Å	4165	1008			38.0	36.6
1	10.0-2.5Å	5229	1008			36.9	36.0
1	10.0-2.0Å	9760	1008		41.1	38.3	36.0

The initial information at our disposal was as follows:

(i) Model could be well-approximated by the hen egg white structure (HEWL).

(ii) A good 6Å isomorphous phased map[7]

(iii) The sequence of residues 1-46 was known[8].

(iv) The composition of the protein was known.

(v) A data set had been collected on a linear diffractometer to a resolution of 1.75Å comprising 15148 unique reflections.

The starting model was created by rotating the co-ordinates of HEWL[9] using a matrix obtained by comparing the respective 6Å isomorphously phased maps of TL and HEWL. An initial overall temperature factor ($16Å^2$) was determined from a plot of $\ln(F_o/F_c)$ against $\sin^2\theta/\lambda^2$, (cf. Wilson plot[10]) giving a starting R-factor for all terms to 1.75Å of 53.3%.

The only strong features in the starting 'omit' map were the four disulphide bridges and some large aromatic side chains (e.g. TRP 112) (see photograph 1). A map calculated after cycle 4 revealed improved electron density for just the main chain, whilst a map produced at the end of the stage (see photograph 2) revealed considerably improved density for both side chains and main chains; thereby demonstrating the beneficial effect of incorporating progressively higher resolution data in the refinement. Various quantitative methods of estimating the maximum resolution of data which may usefully be included in the refinement are available (e.g. (2)), but they were not used in this case due to difficulty in evaluating the root mean square deviation of the model atoms from their 'true' positions with sufficient accuracy.

The initial overall B-factor may be difficult to set where a significant proportion of the available data lies in Region A of fig.1. A symptom of this problem is that overall B-factors tend to assume unrealistically small values ($<8Å^2$); this was evident in the refinements of both Triose

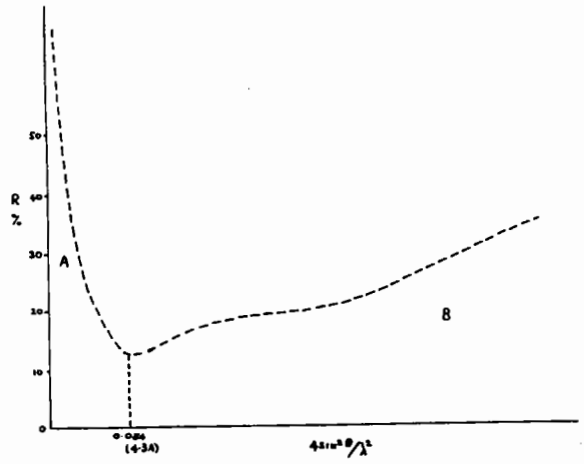
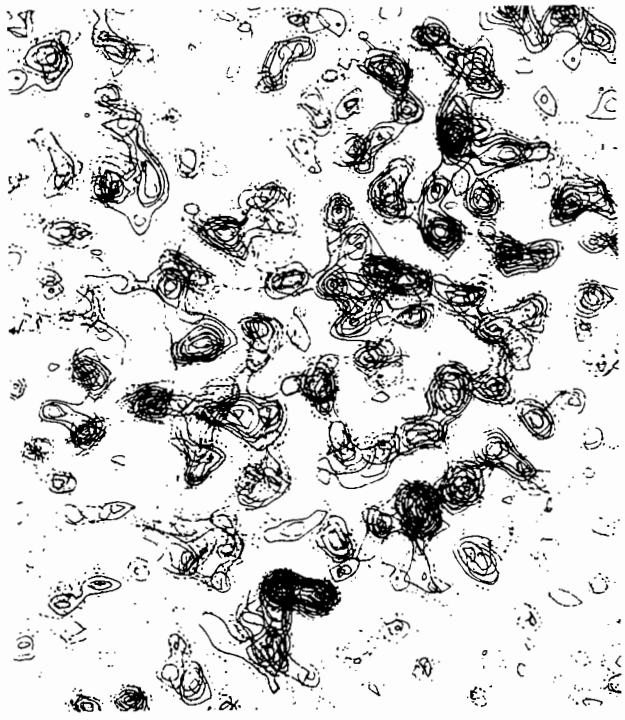
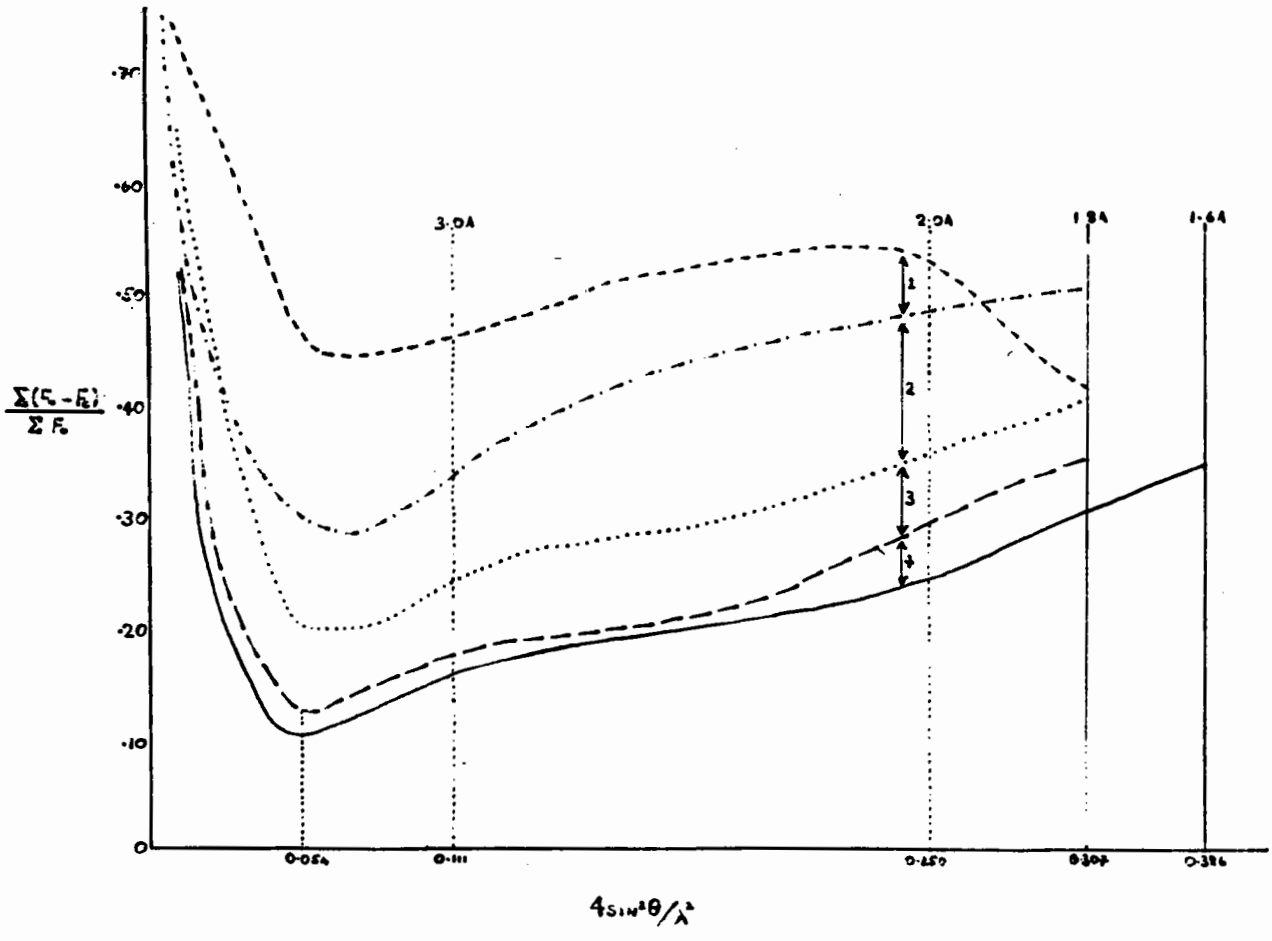


Fig.1. The R-factor plot for TL at an R-factor of 23.7%.

Fig.2. (below) A series of R-factor plots illustrating the progress of refinement of TL.



Photograph 1. Sections 9 to 13 of the initial 'omit' map of TL, R-factor 53.3%.



Phosphate Isomerase and Phosphorylase-b. The effect may be avoided by holding the overall B-factor at some arbitrary but reasonable value, and then allowing the scale factor between F_o and F_c to vary.

Stage II - Weighting used - $1/\sigma^2$
Map type - First rebuild to 'omit', subsequently to $(3|F_o|-2|F_c|)\exp i\alpha_{calc}$ maps.
No. parameters - 4000 approximately.

The 'omit' map calculated at an R-factor of 41.1% was later shown to have contained much spurious detail; this indicated that it was unwise to rebuild extensively to this map. Considerable effort may be saved by refining the protein model as far as possible before rebuilding.

The 13 difference Fourier shifts cycles were run when difficulty was experienced in obtaining a spread of B-values for the atoms in TL using RLSQ; the lack of previous experience of using the program was mainly responsible for this problem. This difficulty did not occur during the refinement of Arabinose Binding Protein by David Rice; this remains the most straightforward refinement of those carried out by Oxford workers, requiring only 25 cycles of RLSQ to refine the starting model produced by Dr. F. Quiocho. This emphasizes the desirability of a good starting model.

Table 3

Progress of Refinement of TL; Stage II

No. Cycles	Resolution	No. Refls.	No. Atoms	Overall R-value Start	Overall R-value Finish	Output R-value Start	Output R-value Finish
Rebuild							
2	10.0-2.0Å	9760	1008	44.0		39.7	36.4
Fourier Shifts - B-factors only							
3	10.0-1.8Å	13473	1003	41.0	36.5		
Rebuild							
Fourier Shifts							
5	10.0-1.8Å	13473	998	38.0	31.8		
Rebuild							
Fourier Shifts							
2	10.0-1.8Å	13473	1010	32.6	30.6		
Rebuild							
Fourier Shifts							
3	10.0-1.8Å	13473	1025	31.4	28.4		

The first two cycles of RLSQ served to further improve the R-factor and the geometry of the model before allowing individual isotropic B-factors for the atoms. Maps exhibited gradually improving definition for the atoms together with decreasing noise levels. Many difference features indicating probable errors in the model appeared during this stage.

Stage III - Weighting used - unit
Map type - $(3|F_o|-2|F_c|)\exp i\alpha_{calc}$
No. parameters - 4000

Table 4

Progress of Refinement of TL; Stage III

No. Cycles	Resolution	No. Refls.	No. Atoms	Overall R-value Start	Overall R-value Finish	Output R-value Start	Output R-value Finish
8	10.0-1.8Å	13473	1025			26.1	23.8
Rebuild							
5	10.0-1.8Å	13473	1022			25.5	23.4
28 water molecules added - now 55 in all							
3	10.0-1.8Å	13473	1060	23.8		23.6	22.1
43 water molecules added							
3	10.0-1.8A	13473	1103		22.7	22.7	20.4

Experience gained in the refinements of Prealbumin and Human Lysozyme has shown that difference Fourier shifts lead to premature convergence on R-factors between 20 and 30%. Unfortunately RLSQ using $1/\sigma^2$ weighting was also found to be ineffective over this range of R-factors. Unit weighting in RLSQ has provided a viable alternative weighting scheme.

Unit weighting has been successful in all refinements so far carried out in Oxford; the final R-factors are lower, and the adherence to accepted peptide stereochemistry is better than in $1/\sigma^2$ weighting.

Further improvement was made in the model by including a gradually increasing number of solvent molecules. The more strongly bound solvent molecules (i.e. internal ones or those bound to the protein surface by two or more hydrogen bonds) were positioned by visual inspection of maps, whilst the remainder were found by an automated difference map peak search (PEAKS) followed by a test on whether, if included, they would conflict with expected hydrogen bond behaviour (i.e. bond distances and angles) in solvent-protein systems (Program WATF3).

Stage IV

A new set of data was used from the start of this stage, comprising 19800 unique reflections to 1.6Å. These data were significantly more accurate than those used previously, having been measured on the Oxford 5-circle diffractometer[11] with long count times for the high resolution data. These data were then profile-fitted[12] and subjected to a three-dimensional absorption correction[13].

The better resolution expected in a map due to the increase in the number of data only gradually became evident as more cycles of RLSQ were run. (See photograph 3).

Figure 2 illustrates the progress of the TL refinement during the four stages described above.

6. IS THE REFINEMENT MEANINGFUL?

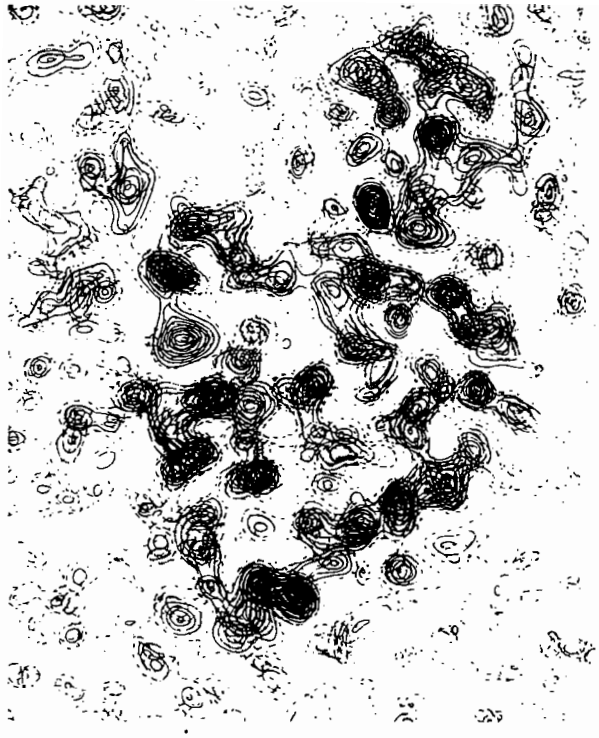
The ultimate test is whether the final electron density map reveals features not included in the refinement. These features may be subtle: for example previously unrestrained hydrogen bonds and Van der Waals contacts should assume reasonable values; or they may be more obvious like the appearance of density for side chains or substrate molecules which had not been considered previously. An analysis of TL hydrogen bonds, which were never restrained during the refinement, reveals bond distances and angles consistent with those observed in other structures. Further confidence in the refinement may be gained by noting that the unknown sequence of TL beyond residue 46 has been established solely from inspection of electron density maps (Note - there are still 5 residues unidentified). Table 6 gives details of the progress of identification of some of these residues.

The refinement of Arabinose Binding Protein provides a further example of the success of refinement; electron density was observed for arabinose in the final electron density map despite the fact that the sugar was not included in the refinement.

Table 5

Progress of Refinement of TL; Stage IV

No. Cycles	Resolution	No. Refls.	No. Atoms	Overall Start	R-Value Finish	Output Start	R-Value Finish
2	5.0-1.6Å	19037	1103			27.7	23.1
2	6.7-1.6Å	19450	1103			21.5	20.6
2	10.0-1.6Å	19673	1103			20.7	20.2
Rebuild							
25 water molecules added							
1	10.0-1.6Å	19673	1127			21.3	21.2
Diagonal Cycle							
1	5.0-1.6Å	19037	1127			21.0	20.1
7	10.0-1.6Å	19673	1127			20.2	19.7
6	5.0-1.6Å	19037	1127			19.6	17.9



Photograph 2. Sections 9 to 13 of the 'omit' map calculated after stage I.

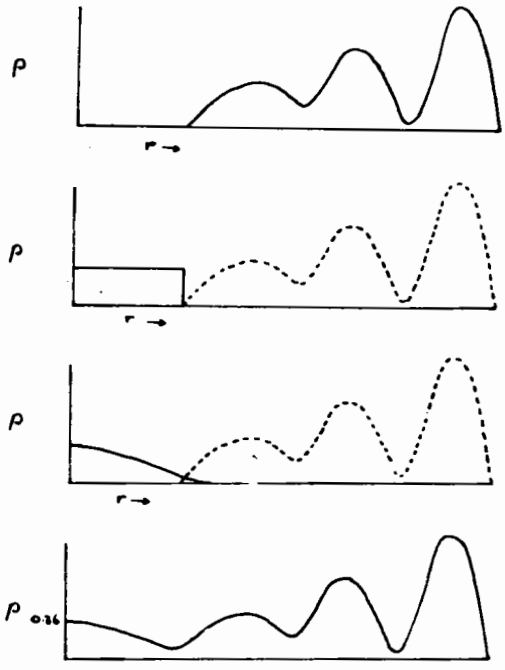
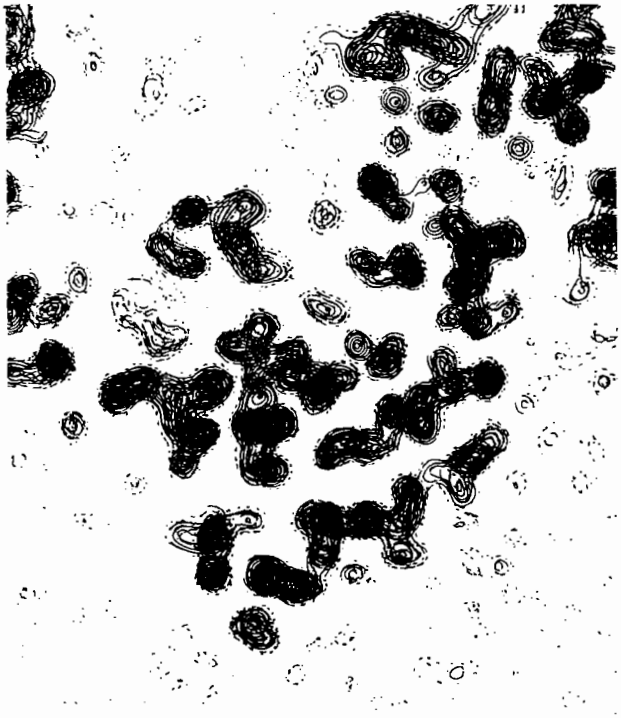


Fig.3a-d. Schematic diagrams to illustrate the incorporation of a crude model of the disordered solvent structure into the protein model.



Photograph 3. Sections 9 to 14 of the $(3|F_o|-2|F_c|\exp\alpha_{calc}$ map calculated after stage IV.

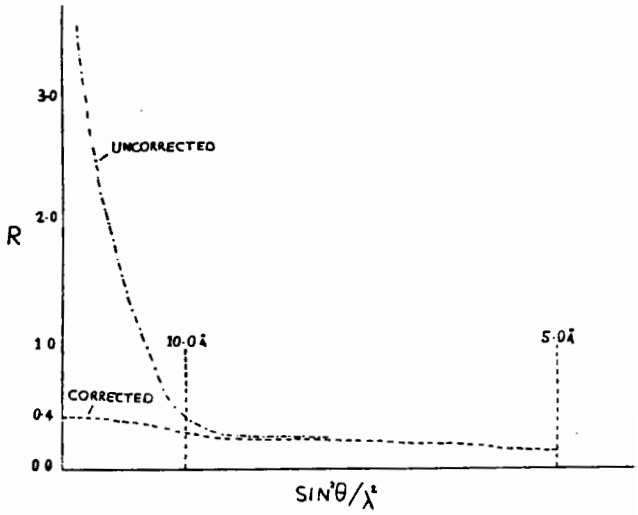


Fig.4. The dramatic improvement in the agreement between F_o and F_c at low angle.

Table 6

	Residue No.												
R-Value	27	47	48	69	73	74	76	78	80	83	92	103	107
53.3	ARG	THR	---	ARG	SER	ARG	LEU	ASN	PRO	ALA	SER	GLY	ASN
36.5	ARG	THR	---	ARG	SER	ARG	VAL	HIS	PRO	ALA	ALA	PRO	ALA
31.8	ARG	THR	---	GLU	SER	LYS	ALA	GLY	SER	ALA	ALA	PRO	GLY
30.6	ARG	PRO	GLY	GLU	ALA	LYS	ALA	GLY	ASP	ASN	ALA	PRO	GLY
23.8	ARG	PRO	GLY	LYS	ALA	LYS	ALA	GLY	ASP	THR	ALA	PRO	GLY
22.7	HIS	PRO	GLY	LYS	ALA	SER	ALA	GLY	ASP	GLN	ALA	PRO	GLY
Now	HIS	PRO	GLY	LYS	ALA	SER	ALA	GLY	GLN	GLN	ALA	PRO	GLY

7. A MODEL FOR THE DISORDERED SOLVENT STRUCTURE

The following procedure was adopted in the case of TL:-

(i) An electron density map where a sum of two Gaussians approximation[2] used to generate the density for an atom was calculated for the asymmetric unit. (Illustrated schematically in fig.3a.)

(ii) This map was modified by setting those grid points not occupied by protein density or ordered solvent density to a value corresponding to $0.4eA^{-3}$, and then resetting the protein and ordered solvent density back to zero (fig.3b).

(iii) The modified map was then Fourier transformed to calculate an intermediate set of solvent structure factors of magnitude $|F_w|$ and phase αw. These structure factors were modified by the application of a scale factor SC, to allow for variation in the disordered solvent density level, and an artificial B-factor Bw, to take account of the unrealistically sharp density edges introduced in (ii). This is illustrated schematically for real space in fig.3c.

(iv) The modified structure factors from (iii) were vectorially combined with F_c to produce a total corrected structure factor F_t. The values of SC(0.9) and Bw($80.0A^2$) were chosen to optimize the agreement between F_t and F_o. The effective electron density map transformed to produce F_t is illustrated by fig.3d.

The agreement between F_t and F_o is shown in fig.4.

The above process caused the R-factor for TL to drop from 22.3% to 19.5%.

This consideration of the disordered solvent, although crude, demonstrates that most protein models would benefit from the inclusion of some approximation to the disordered solvent structure, if only to improve the R-factor. It is interesting to speculate what the effect on a protein's refinement would be if the low angle terms did not have to be weighted down or left out of the refinement process.

8. PROPOSED IMPROVEMENTS TO RLSQ

In order to take account of Region B of fig.1, the current program allows sloping of weight w applied to a reflection at $\sin\theta/\lambda$ of s, according to:-

$$w = 1.0/(AF + BF(S-1/6))^2 \qquad (2)$$

AF and BF are user selectable parameters.

With refinements at lower resolutions (e.g. 2.5Å), a negative value of BF is often input to weight up data between 3.0-2.5Å resolution where the model is more likely to be able to agree with the observations. Therefore it is proposed to institute the slightly different weighting formulae:-

$$\left.\begin{aligned} w &= 1.0/(AF + BF(S-N))^2 \quad S > N \\ w &= 1.0/(AF + CF(N-S))^2 \quad N > S \end{aligned}\right\} \qquad (3)$$

CF is a new parameter introduced to control the weight applied to data lying in Region A of fig.1. The pivot value of the function, set to $1/6Å^{-1}$ in (1), is resettable in (3) by inputting a value for N.

9. COMPUTATIONAL ASPECTS

The versions of RLSQ implements on the CRAY-1 computer at Daresbury are now used extensively by Oxford workers, and they give at least an 18-fold improvement in speed over an ICL 2980. Table 7 gives a speed comparison between an ICL 2980 and the CRAY-1 for various protein refinements.

Table 7

	No. Atoms	No. Refls.	2980 time seconds	Cray-1 time seconds
TL	1100	13500	2200	80
PGK	3100	13000	3200	160
PPB	6520	17000	19000	770

Clearly the CRAY-1 computer enormously simplifies the task of refining a protein, and makes possible refinements which would be impossible using another slower computer. It should be stressed that the speeds quoted above are given for programs written in sensibly coded FORTRAN and not in assembler or special code. Further speed improvement, should it be necessary, may be brought about by either incorporating 'Fast Fourier' methods analogous to the Isaacs-Agarwal program to calculate structure factors and derivatives, or coding the most heavily used parts of the present program in CRAY ASSEMBLER LANGUAGE (CAL) to increase their vectorization.

At present the RLSQ program allows for space group $P2_1$, C_2, $P4_32_12$, $P2_12_12$ and $P2_12_12_1$ but there is little difficulty in expanding the program to allow for other space groups.

REFERENCES

1. H. Lipson and W. Cochran, "The Determination of Crystal Structure". (London: G. Bell and Sons Ltd., 1966).

2. R.C. Agarwal, Acta Cryst. A34, (1978) 791.

3. N.W. Isaacs and R.C. Agarwal, Acta Cryst. A34, (1978) 782.

4. E.J. Dodson, N.W. Isaacs and J.S. Rollett, Acta Cryst. A32, (1976) 311.

5. J.H. Konnert, Acta Cryst. A32, (1976) 614.

6. M. Vijayan, Acta Cryst. A36, (1980) 295.

7. R. Aschaffenburg, C.C.F. Blake, H.M. Dickie, S.K. Gayen, R. Keegan and A. Sen, Biochimica et Biophysica Acta 625, (1980) 64.

8. J. Jollés, A. Sen, E.M. Prager and P. Jollés, J. Mol. Evol. 10, (1977), 261.

9. T. Imoto, L.N. Johnson, A.C.T. North, D.C. Phillips and J.A. Rupley, The Enzymes, 3rd Edn. 7, (1972) 665 (Boyer, P.D. Ed.).

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

11. D.W. Banner, P.R. Evans, D.J. Marsh and D.C. Phillips, J. Appl. Cryst. 10, (1977) 45.

12. S.G. French, D.Phil. Thesis (1975) Oxford University.

13. R. Huber and G. Kopfmann, Acta Cryst. A25, (1969) 143.

THE IMPORTANCE OF REFINED STRUCTURES TO THE UNDERSTANDING OF ENZYME ACTION

by

A.R. Sielecki and M.N.G. James
MRC Group in Protein Structure and Function, Department of Biochemistry,
University of Alberta, Edmonton, Alberta, T6G 2H7, Canada

1. INTRODUCTION

It has long been recognized that substrates bind with high affinity to the active sites of enzymes when their structures are complementary. Empirical observations of this principle are reflected in the very tight binding of inhibitor molecules that can mimic the transition state of the chemical reaction, but can not undergo turnover themselves. Also, minor chemical modifications to substrate structures that disrupt their complementary nature, have marked effects on measured catalytic rate constants. Therefore, in order to appreciate the many intricate molecular events that lead to enzymatic catalysis, a prerequisite must be an accurate definition of the atomic positions in the active site.

Strain and induced fit are two descriptions of molecular mechanisms that make use of the enzyme-substrate binding energy to explain the lowering of the activation energy of the uncatalyzed reaction. Both of these proposals support the tenet that maximum stabilization by the enzyme is achieved only upon reaching the transition state of the reaction. In the strain mechanism which implies a rigid enzyme, geometrical distortion of the substrate towards the transition state is aided by more favorable binding interactions. An induced conformational change of the enzyme structure upon substrate binding provides the enzyme-transition state complementarity in the induced fit mechanism.

Knowledge of the precise geometry of groups at the active site will aid in elucidating the following points: the direction of flow of electrons in the covalency changes of the substrate; whether or not a specific hydrogen-bonding pattern is important in a proposed reaction mechanism; which atoms provide for enzyme-substrate stabilization by electrostatic or hydrophobic interactions; and whether or not atoms of the enzyme move and by how much during the course of the reaction.

Such questions have not been answered definitively for any enzyme. One system, that of the serine proteases, has received perhaps the greatest attention in terms of numbers of crystal structures done. We would like to describe some of the high resolution refinement studies that have been done with a member of this enzyme family, the A protease from Streptomyces griseus (SGPA).

2. SERINE PROTEINASES

There are five proposed intermediates on the catalytic pathway of serine proteases[1]. Under normal circumstances most of these intermediates are relatively short-lived and it is unlikely that one could trap them for a time sufficiently long to characterize them crystallographically. SGPA is a bacterial serine endopeptidase isolated from the commercially available protease mixture Pronase[2-4]. The three dimensional crystal structure of SGPA at pH 4.1, was determined initially at 2.8 Å resolution[5,6]. Subsequently, the structure of the native enzyme has been refined with the restrained parameter least-squares program of Hendrickson & Konnert[7] to an R-factor of 0.13 at 1.8 Å resolution[8]. More recently, and making use of the fact that the tetragonal crystals of SGPA were at pH 4.1, we have determined and refined independently the structures of the complexes that SGPA makes with two specific tetrapeptide substrates and with a tetrapeptide aldehyde inhibitor[9].

Detailed kinetic parameters, k_{cat} and K_M, have been determined for the interactions of SGPA with a large number of oligopeptide substrates[10,11]. From these studies it was shown that the value of k_{cat}/K_M depends very strongly on (a) the length of the polypeptide chain in both the N- and C-terminal

Table 1

Dependence of k_{cat} and K_M on the nature of P_1 and the chain length of substrates for SGPA†

	P_5 P_4 P_3 P_2 P_1 P_1'	$k_{cat}(s^{-1})$	K_M (mM)	k_{cat}/K_M (s^{-1} M^{-1})
I	Ac-Pro-Ala-Pro-Gly-NH_2	0.012	20	0.6
II	Ac-Pro-Ala-Pro-Phe-NH_2	5.8	0.54	10700
III	Ac-Pro-Ala-Pro-Tyr-NH_2	10.1	1.4	7200
IV	Ac-Phe-NH_2	0.007	25	0.28
V	Ac-Pro-Phe-NH_2	0.13	4.9	27.
VI	Ac-Pro-Ala-Pro-Phe-CHO	-	$K_i = 5x10^{-5}$ mM	-

† Data taken from Bauer et al., 1976[10,11]

directions and (b) the nature of the amino acid in the P_1* position. These dependencies are illustrated in the data of Table 1. In addition to the data on hydrolysis of these amides at pH 9.0, the inhibition constant for the tetrapeptide aldehyde (VI) is also given[13].

We have soaked crystals of SGPA in solutions containing the substrates described in Table 1 (kindly supplied by Dr. C. Bauer), collected the three dimensional diffraction data to a variety of resolutions, and used the traditional difference Fourier methods to determine the binding modes of these peptides. Interpretation of the initial studies was at 2.8 Å resolution from difference maps computed with the MIR phases. Gradually we have extended these studies to higher resolution and have used calculated phases obtained at different stages of the refinement of the native structure. Difference Fourier maps suffer from two serious drawbacks. Firstly, the overlap of positive and negative features as a result of displacement of bound solvent in the native structure leads to distortion of the electron density for the bound molecule. In addition, it is difficult to differentiate actual shifts in enzyme position from possible changes in atomic temperature factors. Secondly and perhaps more importantly, the subjective interpretative bias of the crystallographer is almost impossible to remove.

Some of these points are illustrated in the series of figures 1-3. Fig. 1 is a difference Fourier map (amplitudes $|F_{N+S}|-|F_N|$, where $|F_{N+S}|$ are measured amplitudes from a crystal soaked in the tetrapeptide; $|F_N|$ structure factor amplitudes from native crystals; MIR phases) at 2.8 Å resolution of the tetrapeptide AcProAlaProPhe (AcPAPF) bound in the active site of SGPA. This is a good map, even though it is at medium resolution, but its interpretation was difficult and the details limited. Primarily, as the difference electron density peak assigned to the carboxyl group of the P_1 Phe residue lay close to the O^{γ} of Ser195, we favored an interpretation of a covalent bond between these two groups. It can be seen that the remainder of the molecule fits quite well to the difference density.

The vastly improved resolution (1.8 Å) displayed in Fig. 2 allowed for a better fit of the tetrapeptide AcPAPF to the difference density (amplitudes as in Fig. 1; phases are α_c computed from the partially refined native structure at R = 0.23). However, the distortions in the density are evident

*The nomenclature of Schechter & Berger[12] has been adopted to describe the enzyme-substrate binding interactions. The substrate has amino acid side chains P_n ..... P_n' that bind to subsites on the enzyme S_n ..... S_n'. The bond hydrolyzed is between residues P_1-P_1'. The N-terminus of the substrate is P_n.

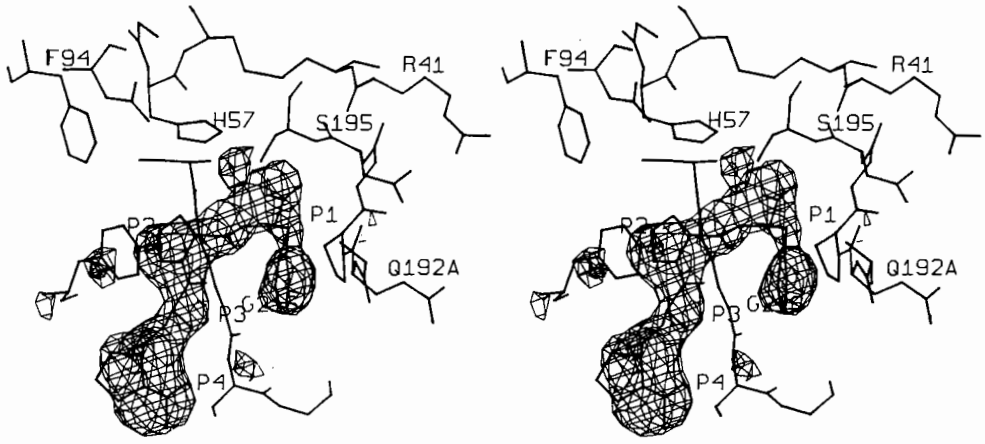


Fig. 1. Difference electron density map (2.8 Å resolution) in the region of the active site of SGPA. The structure factor amplitudes are $|F_{N+S}|-|F_N|$, phases from multiple isomorphous replacement (MIR) at 2.8 Å. The structural models represented are those for native SGPA and the bound tetrapeptide AcPAPF as determined from the Kendrew model[5].

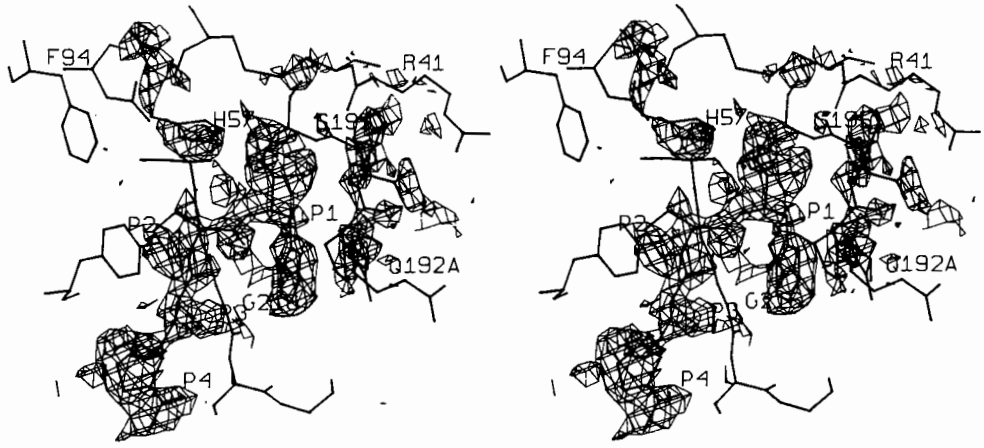


Fig. 2. Difference electron density map at 1.8 Å resolution. The SGPA model and the view angle are the same as in Fig. 1. The model for the tetrapeptide AcPAPF has been fitted on an interactive graphics system (MMS-X). Negative density contours have been omitted for clarity, the contour surface displayed is 0.09 eÅ^{-3}. Apparent conformational changes in the enzyme structure could not be separately interpreted from changes in B factors.

(especially at P_2 Pro). In addition, the electron density peak corresponding to the carboxyl group could be interpreted equally well by a planar carboxyl group or by a tetrahedral group which would include a covalent bond to O^{γ} of Ser195. Attempts to clarify this latter point at several further stages in the refinement of the native enzyme did not resolve the ambiguity. Differences in atomic position for the two models were of the order of 0.3 - 0.4 Å. One of the aims of the study of the binding of these peptides to SGPA was to provide reasons, on the molecular level, for the kinetic differences presented in Table 1. For example, the presence of a p-OH group in AcPAPY has a marked effect on both K_M and k_{cat} relative to AcPAPF, and clearly such subtle differences could not be elucidated with the ambiguities in the binding modes mentioned above.

Refinement of the intensity data measured from crystals of the complexes alone and independently of the native data was then carried out.

Figure 3 shows an electron density map in the same

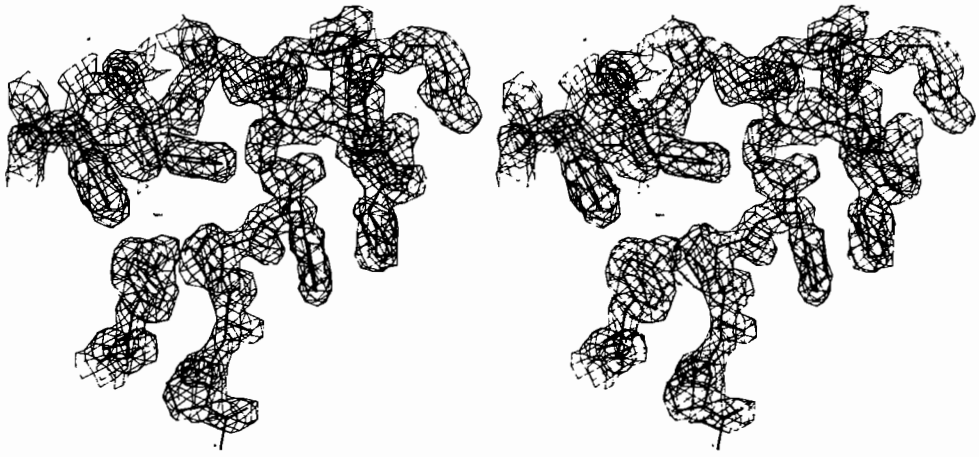


Fig. 3. $2|F_o|-|F_c|$, α_c electron density map ($|F_o|$ and $|F_c|$ are observed and calculated structure factor amplitudes from crystals of SGPA soaked in solutions of AcPAPF NH_2) at 1.8 Å resolution and R = 0.119. The electron density contours surfaces for enzyme and product molecules are +0.40 eÅ^{-3}. Distance from O^{γ} Ser195 to the carbonyl C of P_1 Phe is 2.65 Å.

region of the active site of SGPA as for Figs. 1 and 2. This is a $2|F_o|-|F_c|$, α_c map computed after 16 cycles of refinement of the data from a crystal of the AcPAPF complex with the Hendrickson-Konnert program. The current agreement factor R is 0.119 and the resolution is 1.8 Å (see Table 2 for a summary of the refinement). This map clearly shows that there is not a covalent bond from O^{γ} of Ser195 to the carboxyl carbon atom (the distance is remarkably short however, 2.65 Å) and that the carboxyl group is planar within the limits of error of this refinement (~0.1 Å).

3. REFINEMENT RESULTS

The refinement of native SGPA at 1.8 Å resolution has been described[8] as well as the mechanistic implications based on the refined structures of an inhibitor (AcProAlaProPhe-aldehyde) and two product complexes (AcProAlaProPhe and AcProAlaProTyr) with SGPA[9]. The resolution of the native data has now been increased to 1.5 Å, and refinement continued for 23 additional cycles. It is the results of the 1.5 Å structure which we shall present here. In addition, the intensity data from crystals of the three tetrapeptide complexes with SGPA: AcPAPF, AcPAPY and AcPAPF-aldehyde have been reprocessed with an improved background smoothing function (R. Read, unpublished). Refinement for a number of additional cycles has been carried out on these data. The present results of these four independent refinements of native SGPA at 1.5 Å resolution and of SGPA complexed to three tetrapeptide molecules at 1.8 Å resolution are given in Table 2. The values of the unweighted agreement factors (R) and their variation with sin θ are depicted in Fig. 4. These low values, in addition

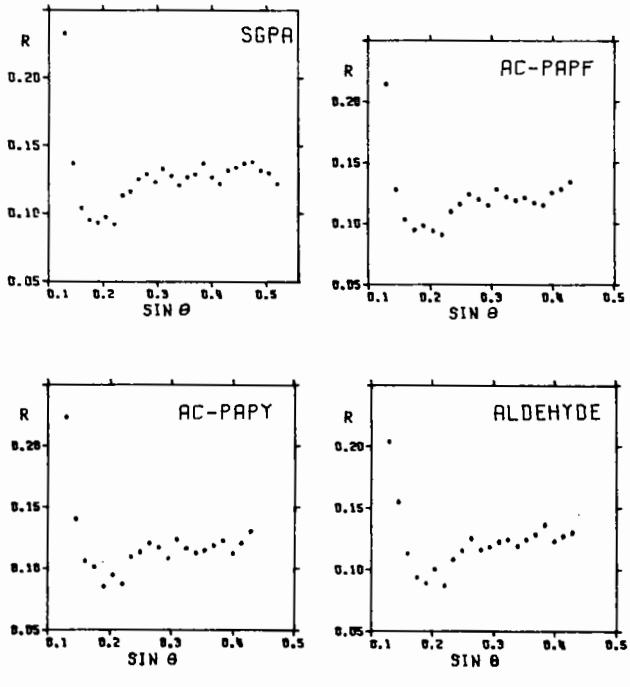


Fig. 4. Graphical representation of the variation of the crystallographic $R = \Sigma||F_o|-|F_e||/\Sigma|F_o|$ with resolution (Sin θ). All four structures have poor agreement for the low ranges of Sin θ (d > 6 Å) indicating that we have not adequately accounted for weakly occupied solvent sites or a water continuum.

Table 2

Results of least-squares refinements of SGPA and peptide ligands

	Native	AcPAPF	AcPAPY	AcPAPF-ald
$R(=\Sigma\|\|F_o\|-\|F_c\|\|/\Sigma\|F_o\|)$	0.126	0.119	0.116	0.120
No. cycles	58	16	13	21
No. reflections [$I > 3\sigma(I)$]	17194	11811	11999	10218
Resolution range (Å)	12.5-1.5	10.0-1.8	10.0-1.8	10.0-1.8
No. atoms (solvent)	1494 (235)	1498 (205)	1498 (204)	1498 (206)
$\langle B \rangle$ (Å^2)	15.7	14.4	14.9	14.4
$\langle\|F_o\|\rangle$ (e)	159	199	195	214
$\langle\|F_o\|-\|F_c\|\rangle$ (e)	19.8	23.8	22.7	27.3
r.m.s. Δ^+ (Bond length, Å)	0.017	0.020	0.019	0.017
r.m.s. Δ^+ (Angle length, Å)	0.033	0.038	0.036	0.036
r.m.s. Δ^+ (planar groups, Å)	0.015	0.014	0.015	0.012
r.m.s. Δ^+ (chiral centres, Å)	0.196	0.146	0.148	0.124
r.m.s. Δ^+ (nonbonded contacts)				
single torsion (Å)	0.254	.209	0.207	.214
multiple torsion (Å)	0.163	.180	- *	- *
possible H-bond (Å)	0.281	0.194	0.147	0.162
r.m.s. Δ^+ [planar peptide, ω (°)]	3.1	2.9	2.9	2.4

+ These Δ values are root mean square deviations from the corresponding values for ideal groups derived from small molecule structural studies[8]. The r.m.s. deviations are from the final least-squares cycle for each type of parameter restrained by the algorithm[7].

* These parameters were not restrained during the refinement of these complexes.

to the small r.m.s. deviations of the final atomic parameters from "ideal" bond lengths and angles (as determined from small molecules) indicate that these refinements have produced accurate descriptions of the SGPA molecule and its tetrapeptide complexes.

4. CONFORMATIONAL CHANGES

There are a number of conformational changes in the SGPA structure that have been induced upon substrate binding. The magnitude of these changes can be deduced by an atom by atom comparison of native SGPA coordinates with the coordinates of SGPA in each of the complexes. As each of these separate data sets has been refined independently, these comcomparisons should also provide an independent estimate of the positional accuracy that has been achieved as a result of refinement. Table 3 contains the results of these comparisons. The r.m.s. differences indicate that the resulting structures are very similar. Since there are several regions of the molecule that have significantly different conformations in the native and complexed form of the enzyme, the most appropriate measure of the coordinate precision is obtained from the comparison of the SGPA molecule in the complexes with AcPAPF and with AcPAPY. For this case, and considering main chain atoms only the r.m.s. difference is 0.05 Å (Table 3). Those regions of polypeptide

Table 3

Coordinate differences between atoms of SGPA in free and complexed forms (Å)

		SGPA + AcPAPF	SGPA + AcPAPY	SGPA + aldehyde	AcPAPF†	AcPAPY	Aldehyde
SGPA	all atoms	0.11 (0.58)*	0.10 (0.43)	0.14 (0.68)	-	-	-
	main chain	0.09 (0.44)	0.08 (0.35)	0.12 (0.49)			
SGPA + AcPAPF	all atoms	-	0.07 (0.28)	0.10 (0.43)	-	0.11 (0.21)	0.29 (0.60)
	main chain		0.05 (0.13)	0.08 (0.21)			
SGPA + AcPAPY	all atoms	-	-	0.10 (0.41)	-	-	0.33 (0.62)
	main chain			0.09 (0.20)			

* The numbers in parentheses represent the largest differences observed in each of the cases compared.

† Atoms of the tetrapeptides alone are compared in this matrix.

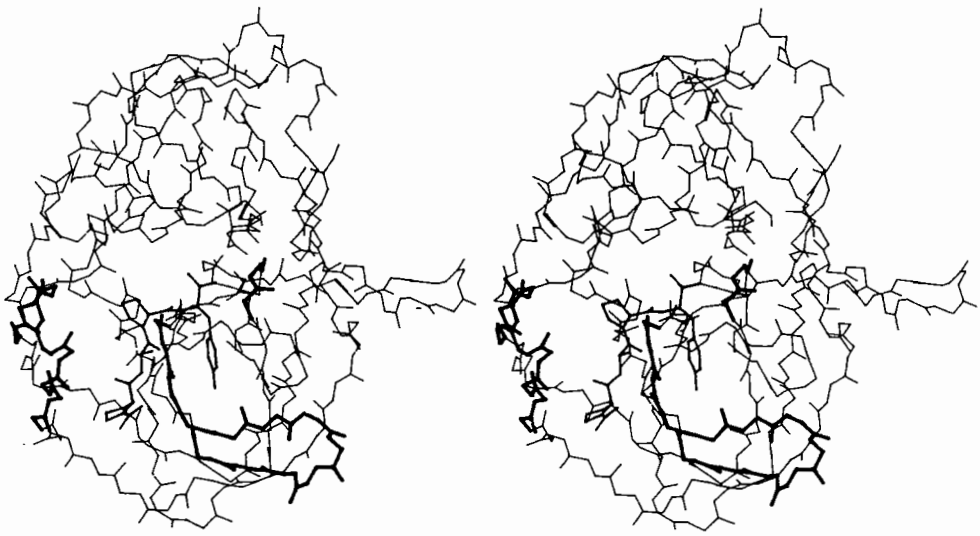


Fig. 5. The backbone of SGPA ($-N-C^{\alpha}-\overset{\overset{O}{\|}}{C}-$) with the atoms of the refined tetrapeptide AcPAPY superimposed. The regions of the backbone of the enzyme portrayed in thick lines are those segments which have moved in a concerted way by more than 0.15 Å from their positions in the native enzyme. The direction of movement is, in general, away from the product molecule.

chain which differ by more than three times this value are deemed to have significantly altered their conformations when the tetrapeptide products are bound and they can be visualized on the stereo drawing of the main chain of SGPA in Fig. 5.

Certainly these relatively small alterations in the conformation of the enzyme are residual values and do not reflect conformational changes that could have occurred in the initial encounter complex with the substrates. It also does not render any

information about movements that may occur along the catalytic pathway from substrate to product. Indeed, the observed changes are rather small (maximal changes of 0.35 - 0.49 Å are in the region of Tyr171). The atoms of the main chain and the attached side chains of the stretches highlighted in Fig. 5 have, in most cases, moved away from the atoms of the bound products. The majority of the residues do not differ significantly from their positions in the native enzyme.

In the region of the active site, the imidazole ring of His57 is slightly displaced in the complexes of SGPA with AcPAPF and AcPAPY. The movement leaves $N^{\delta 1}$ in approximately the same position, while $N^{\varepsilon 2}$ is displaced by ∿0.2 Å in a direction such that a more favorable (shorter and more linear) H-bond to the oxygen atom of the carboxylate of the products is formed. This hydrogen-bonded interaction can be seen in Fig. 3 (the $N^{\varepsilon 2}$ ... O distance in AcPAPF is 2.84 Å; in AcPAPY, 2.79 Å).

A corresponding conformational change involving His57 of the SGPA-aldehyde complex is much larger. The imidazole ring rotates from its observed position in the native structure out into the solvent region. The imidazole ring has an increased average B (25.1 A^2 from 10.6 A^2 in native SGPA). Its native position is occupied by two water molecules, hydrogen-bonded to the carboxylate of Asp102 and to $N^{\delta 1}$ of His57. Similar bridges of two water molecules between ion pairs are observed elsewhere on the surface of SGPA.

5. SOLVENT STRUCTURE

The roles played by water in enzyme mechanisms are many and varied. It is therefore important to determine the water structure as accurately as possible in order to assess how it can influence catalysis. As the interactions that water makes with the enzyme are non-covalent (electrostatic and hydrogen bonding) the positions of the oxygen atoms are less well determined than the atoms of the protein. This will vary as the number of hydrogen-bonded contacts to the protein varies[14]. In addition to the positional coordinates, the parameters refined for water molecules are the occupancy and the B factor (= $8\pi^2 \overline{u^2}$). These parameters seem to be the least well determined in the refinement of the four structures summarized in Table 2. We have selected a total of 235 solvent sites in native SGPA (total possible ∿660). These sites were chosen as those peaks with heights greater than ∿4 times the estimated standard deviation (∿0.2 $e\text{Å}^{-3}$) of difference maps ($|F_o|-|F_c|$, α_c) and making reasonable contacts to possible hydrogen-bond donor or acceptor atoms of the protein or other water molecules.

Other workers have pointed out that there is a marked correlation of the occupancy and B factors of water molecules[14]. That is, solvent atoms with high occupancy factors tend to have low B factors and vice versa. In order to examine the generality of this observation, we have plotted the occupancy factors of the solvent molecules in SGPA versus their refined B factor at the conclusion of cycle 58. This plot (Fig. 6) shows that there is little, if any, correlation (the sample correlation coefficient, r = 0.08) of these two parameters for

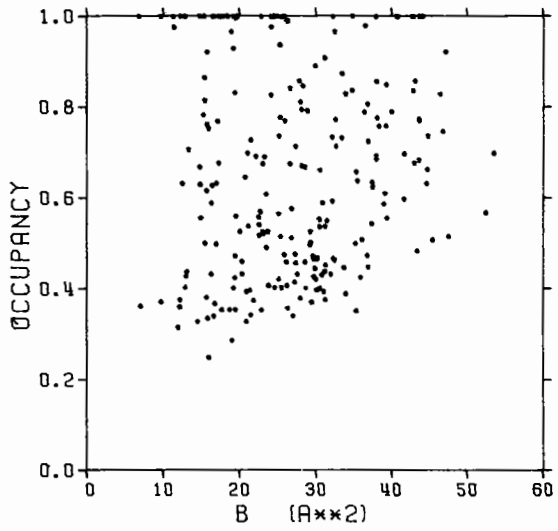
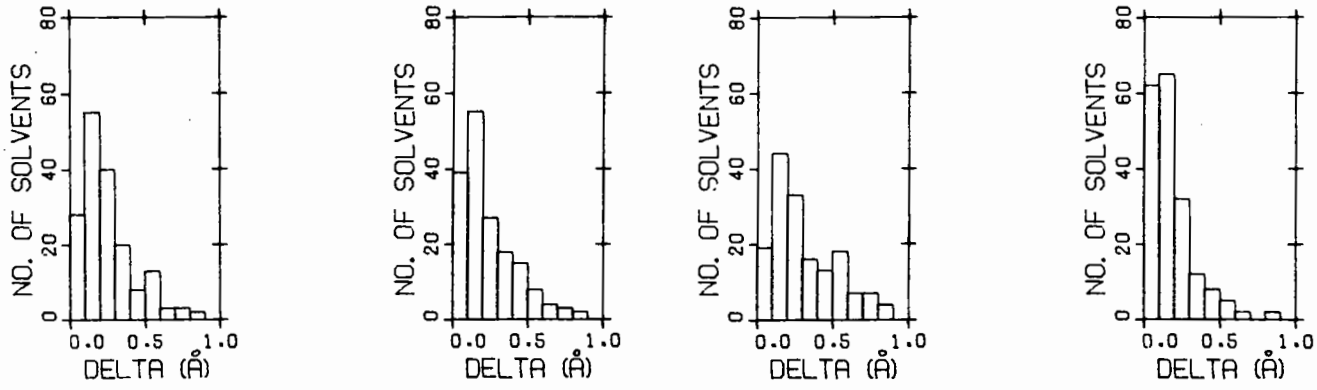


Fig. 6. Plot of the B factor (=$8\pi^2 \overline{u^2}$, Å^2) for the 235 solvent atoms in native SGPA vs the occupancy factors for those atoms. The sample correlation coefficient is 0.08.

the water molecules. This result could reflect the fact that the B factor and the occupancy are really measures of two physically different situations. While they may be strongly correlated in a least squares refinement procedure, it should not be physically unrealistic to have a highly occupied site with a large thermal motion parameter. We are comforted in that we have four independent determinations of SGPA and its water structure, so



SGPA vs. AcPAPF
175 solv. <1.0 Å
(205 Total)
r.m.s. = 0.35 Å

SGPA vs. AcPAPY
171 solv. <1.0 Å
(204 Total)
r.m.s. = 0.33 Å

SGPA vs. Aldehyde
164 solv. <1.0 Å
(206 Total)
r.m.s. = 0.41 Å

AcPAPF vs. AcPAPY
188 solv. <1.0 Å
(204 Total)
r.m.s. = 0.26 Å

Fig. 7. Results of comparing positions of 'equivalent' solvent sites in the four refined structures. Delta is the positional difference between these sites in Angstroms.

comparison of these parameters should give confidence in the results obtained.

Figure 7 shows a comparison of differences in position of the refined solvent molecule sites (one of them, a Na^+ ion, see reference 8) in the four structures. Examination of this figure indicates that not only the conformation of the enzyme is altered when product or inhibitor tetrapeptide is bound, but also the water structure is altered. Again, the comparison of solvent sites in the crystals of AcPAPF and AcPAPY shows that there are more solvent sites that are almost coincident in these two independently refined structures in which the SGPA portion undergoes similar conformational changes. Therefore one can conclude that when the segments of the SGPA molecule move, the solvent moves in concert.

It is encouraging that 1.8 Å resolution data, with careful, restrained least-squares refinement, can give meaningful data regarding solvent sites in protein structures.

6. PROTEIN FLEXIBILITY

The serine proteases have been termed mechanically rigid, electronically strained enzymes[15]. Analysis of the variation of B factor with position along the polypeptide chain shows that there are regions of the SGPA molecule that have much higher B factors than the average value[8]. Besides, if we look at the changes in B for residues of the SGPA molecule when complexed with the tetrapeptides, those regions that have some of the highest B factors in the native structure (Fig. 8) and which make contact with the tetrapeptides have reduced B values in the complexes[9]. Therefore, in view of the large values of B for some regions of the molecule and the fact that conformational movements are observed in the complex formation, we should rather consider the serine protease structure heterogeneous; like the curate's egg, "soft in parts". The fact that residues constituting the substrate binding sites have some of the highest B factors leads us to infer that residues Gly39 to Arg41 and Ile63 could possibly be involved as binding site residues on the leaving group side of the scissile bond of a good substrate (P_1' to P_3'). From these arguments and other more direct data, we have model built a P_1' Phe on to the tetrapeptide AcPAPF as shown in Fig. 8.

7. CONCLUDING REMARKS

The structure of SGPA has been refined by restrained-parameter least squares to an R of 0.126 for 17194 reflections to 1.5 Å resolution. A similar level of accuracy has been attained in the refinement of SGPA when it is complexed to two tetrapeptide product molecules and to a transition state analogue inhibitor AcProAlaProPhe-aldehyde. These several refinements have allowed us to define with some confidence the enzyme-inhibitor and enzyme

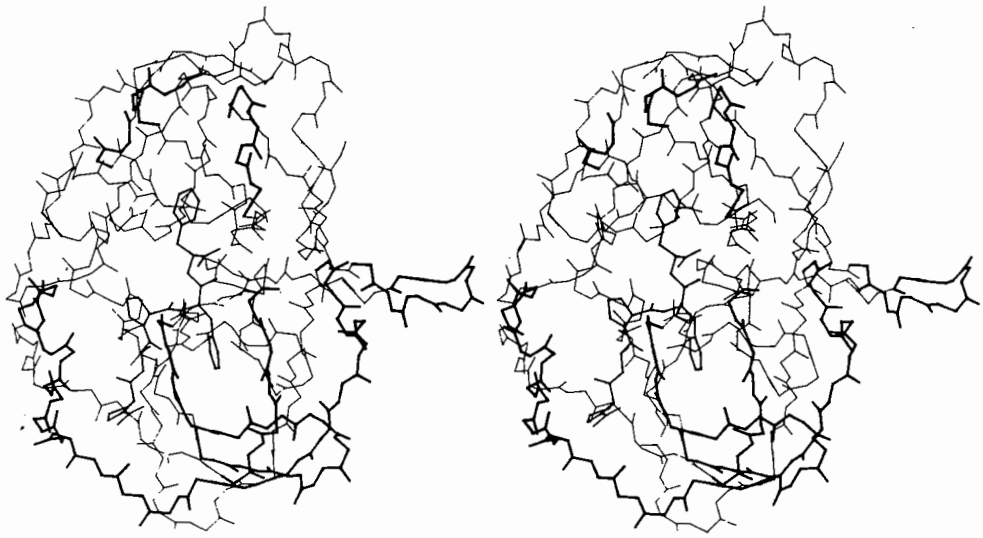


Fig. 8. The backbone atoms of SGPA with atoms of a proposed pentapeptide, AcPAPFF, bound in the active site. The regions highlighted by thick lines are those atoms which have B factors larger than 12.7 $Å^2$ (average value for atoms of main chain of the protein).

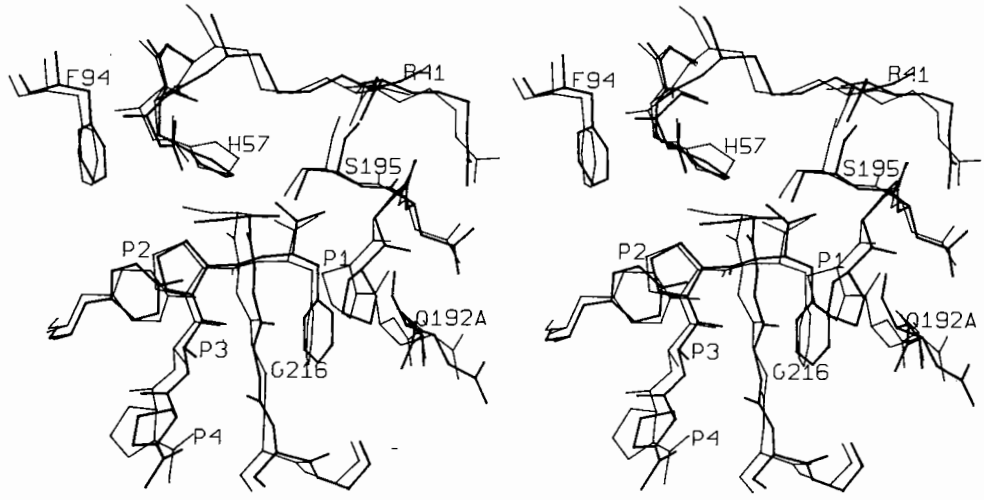


Fig. 9. Comparison of the final coordinates of AcPAPF plus the active site region of SGPA (thick lines) with those coordinates as determined from the initial interpretation of the 2.8 Å Δ F map (thin lines between atoms). The r.m.s. difference in coordinates for AcPAPF is 0.76 Å, and for the atoms of the active site of SGPA is 0.76 Å.

product molecule interactions. In addition, we have observed how SGPA reacts when it has a tetrapeptide bound in the second product binding site (S_1 to S_4). The molecule undergoes some small rearrangements that are associated with the binding. Also, these same regions have reduced thermal vibration parameters in the complexes.

In order to detect deviations from planarity in a putative tetrahedral model for the carboxyl groups of the two product complexes it is necessary to have atomic positions with an accuracy of ∿0.03 Å or better. Any minor deviations from planarity that we observe in these structures (0.09 Å and 0.03 Å for the AcPAPY and AcPAPF complexes, respectively) are less than the significance limit of our data. It is also important to have refined structural data to confirm the presence or absence of hydrogen bonding. In order for the charge relay mechanism to operate it was postulated that a hydrogen bond from O^{γ} of Ser195 to $N^{\varepsilon 2}$ of His57 was present. The refined $N^{\varepsilon 2}$ to O^{γ} distance in SGPA is 3.03 Å, on the upper limit for such a bond, and markedly non-linear. There is no alternative position for O^{γ} (by rotation about C^{α}-C^{β}) that improves the directionality of this putative bond. The

presence of water molecules bound to His57 and Ser195 in the native structure, also reduces the likelihood of the originally proposed disposition.

The correct interpretation of catalytic pathways of enzymes requires the knowledge of the precise positions of the atoms of the substrate and of the enzyme throughout the course of the reaction. The possibility of hydrogen-bonding influence, directions of dipole moments and all other positional dependent terms which are considered important to the mechanism, have to be defined precisely. The molecular drawings in Fig. 9 show that the initial interpretation of the binding of AcPAPF to SGPA is not sufficient to provide meaningful interpretations of this mechanism. Hydrogen-bonding interactions that are present in the refined structure but not in the initial interpretation (e.g. $N^{\varepsilon 2}$ His57 to OT of P_1 carboxyl group, OH of Tyr171 to OH of Ser214, NH of P_3 Ala to C = O of Gly216; NH of P_1 Phe to C = O of Ser214) emphasize the fact that we leaned heavily on our chemical intuition rather than firm experimental evidence in the early stages. Certainly, the geometry is important but it is not the only parameter required to define the pathway. The kinetics become much more interpretable in light of it though!

ACKNOWLEDGEMENTS

Koto Hayakawa has done a marvellous job in growing crystals of SGPA; Carl-Axel Bauer has been most generous in supplying samples of his valuable peptides; Gary Brayer and Louis Delbaere collected data on the FACS-1 diffractometer; Mae Wylie has been very efficient in producing this manuscript in time. We acknowledge financial support from the Medical Research Council of Canada to the MRC Group in Protein Structure and Function at the University of Alberta. Lindsay Sawyer is responsible for the curate's egg model of an enzyme.

REFERENCES

1. J. Kraut, Ann. Rev. Biochem. 46, (1977) 331.

2. P. Johnson and L.B. Smillie, Can. J. Biochem. 49, (1971) 548.

3. P. Johnson and L.B. Smillie, FEBS Letters 47, (1974) 1.

4. L. Jurášek, P. Johnson, R.W. Olafson and L.B. Smillie, Can. J. Biochem. 49, (1971) 1195.

5. G.D. Brayer, L.T.J. Delbaere and M.N.G. James, J. Mol. Biol. 124, (1978) 243.

6. G.D. Brayer, L.T.J. Delbaere and M.N.G. James, J. Mol. Biol. 124, (1978) 261.

7. W.A. Hendrickson and J.H. Konnert, in Biomolecular Structure, Function, Conformation and Evolution, Vol. 1; (Oxford: Pergamon Press, 1980).

8. A.R. Sielecki, W.A. Hendrickson, C.G. Broughton, L.T.J. Delbaere, G.D. Brayer and M.N.G. James, J. Mol. Biol. 134, (1979) 781.

9. M.N.G. James, A.R. Sielecki, G.D. Brayer, L.T.J. Delbaere and C.-A. Bauer, J. Mol. Biol. 143, (1980) in the press.

10. C.-A. Bauer, R.C. Thompson and E.R. Blout, Biochemistry 15, (1976) 1291.

11. C.-A. Bauer, R.C. Thompson and E.R. Blout, Biochemistry 15, (1976) 1296.

12. I. Schechter and A. Berger, Biochem. Biophys. Res. Commun. 27, (1967) 157.

13. G.D. Brayer, L.T.J. Delbaere, M.N.G. James, C.-A. Bauer and R.C. Thompson, Proc. Natl. Acad. Sci. (U.S.A.) 76, (1979) 96.

14. K.D. Watenpaugh, T.N. Margulis, L.C. Sieker and L.H. Jensen, J. Mol. Biol. 122, (1978) 175.

15. D.M. Blow and T.A. Steitz, Ann. Rev. Biochem. 39, (1970) 63.

MAP IMPROVEMENT BY THE COMBINATION OF PARTIAL STRUCTURE AND ISOMORPHOUS PHASE INFORMATION

by

Dr. D.W. Rice

Laboratory of Molecular Biophysics, University of Oxford, South Parks Road, Oxford OX1 3PS.

1. ABSTRACT

The structure of horse muscle phosphoglycerate kinase has been refined with the restrained least squares procedure of Hendrickson and Konnert using the X-ray data to 2.5Å resolution. During the refinement errors in the model have been detected by the use of electron density maps whose phases were produced by combining the isomorphous phase information with that obtained from the partial structure. A qualitative evaluation has been made of the level of feedback induced in the combined maps by incorrectly located atoms in the model. The Fourier syntheses computed with the combined phase angles contain much more new information concerning errors in the structure than equivalent Fouriers based on the calculated phases alone. A significant improvement in the electron density over the isomorphous map has been achieved.

2. INTRODUCTION

The aim of a protein structure determination is to be able to provide a reliable model of the atomic arrangement within the molecule. From this model deductions can then be made about the interactions that govern its stability and relate to its function. The reliability of these deductions is then directly related to the quality of the electron density map from which the model was built. Whilst several techniques of refinement or phase extension can be used to provide high quality electron density maps where the X-ray data are available at or near atomic resolution [1,2,3,4,5] these methods have not yet been successfully applied at resolutions around 2.5Å.

In these cases the interpretation of the structure usually follows from the inspection of an isomorphously phased electron density map. Phases derived by isomorphous replacement are subject to errors from a number of sources, principally non-isomorphism of the derivatives, low occupancy of the heavy atoms, inadequate heavy atom refinement and errors in the X-ray data. These errors result in a lowering of the resolution of the electron density map as well as giving rise to severe disturbances in the map near the heavy atom binding sites. Areas of diffuse density are often found in such maps and these present large problems of interpretation.

In order then to obtain a reliable set of atomic co-ordinates an improvement of the electron density is needed. During the refinement of Phosphoglycerate Kinase the use of partial structure phase information when combined with that from isomorphous replacement has been examined and will be discussed here.

3. THE METHOD OF PHASE COMBINATION

The combination of the isomorphous and partial structure phase information was achieved by multiplying the individual phase probability functions together [6]. In order to carry out the computation in a convenient manner the phase probability functions were stored as the four phase coefficients A, B, C and D suggested by Hendrickson and Lattman [7]. Thus

$$P_{ABCD}(\alpha) = \exp(N+A\cos\alpha+B\sin\alpha+C\cos 2\alpha+D\sin 2\alpha) \quad (1)$$

where $P_{ABCD}(\alpha)$ is a generalized probability function.

The isomorphous phase probability computed by the method of Blow and Crick [8] was recast in terms of the probability function P_{ABCD} by the least squares fitting procedure derived by Hendrickson [9]. The phase information from the partial structure was determined using the probability analysis of Sim [10]. Sim showed that

$$P_{par}(\alpha) = K \exp(X\cos(\alpha-\alpha_{calc})) \quad (2)$$

where $X = 2F_{obs} F_{calc}/\beta$

and $K = 1/2\pi I_o(X)$

where I_o is the modified zero order Bessel function. Sim[11] then suggested a suitable weighting scheme that would produce the map with the least mean square error, the weights being defined by

$$w = \int_{-\pi}^{\pi} \cos\alpha \; P_{par}(\alpha)\, d\alpha \qquad (3)$$

The value of β from equation (2) was in Sim's original formulation estimated from the proportion of atoms missing in the partial structure. An empirical estimate of β obtained from the lack of closure between I_{obs} and I_{calc} in ranges of $\sin^2\theta/\lambda^2$ has been used by Bricogne[12] and this approach has been taken in the modified version of Dr. Bricogne's program "Combine" which was used to compute the combined phases described here.

4. THE APPLICATION OF PHASE COMBINATION TO PHOSPHOGLYCERATE KINASE

The enzyme phosphoglycerate kinase, extracted from horse muscle, is a monomer of m.wt. 45,000 Daltons which crystallizes in the space group $P2_1$ with one molecule in the asymmetric unit and cell dimensions

$a = 50.8\text{Å}$

$b = 106.9\text{Å}$

$c = 36.3\text{Å}$

$\beta = 98.6^\circ$

The polypeptide chain contains 416 amino acids and the chain trace was obtained by inspection of a 3.0Å and later a 2.5Å M.I.R. phased electron density map[13,14]. Unfortunately no sequence information was available when these maps were calculated and the quality of the electron density was such that the chain trace was ambiguous. Furthermore the side chain definition in the maps was poor and there was little evidence of the characteristic bulge in the electron density at the positions of the peptide carbonyl groups. In an attempt to improve the map, the procedure of phase combination was used to combine the phase information available from a progressively improving interpretation of the molecular structure with the isomorphous phase probabilities[15]. The strategy of this combined refinement and map improvement was centred around the gradual incorporation of the amino acid sequence information into the molecular model[14]. The flow diagram in fig.1 illustrates the approach used.

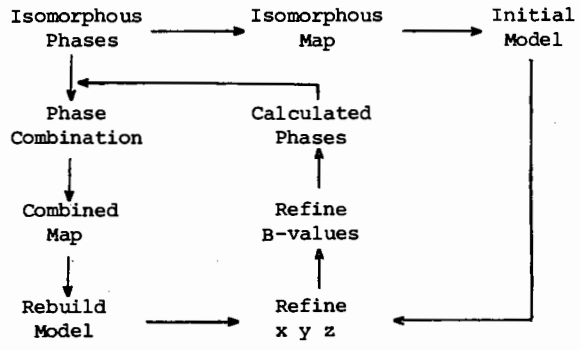


Fig.1 Flow diagram of refinement procedure incorporating phase combination.

During the refinement two techniques have been used to calculate atom shifts. Initially the method of automated difference Fourier shifts[16] followed by a model regularization step[17] was employed. More recently the restrained least squares procedure of Hendrickson and Konnert was used. In the early stages of the refinement group temperature factors (one for each side chain and main chain residue) were allowed, in the last round of refinement individual temperature factors have been used. The production of a new set of combined phases has followed each convergence of the refinement process, the details of which are given in Table 1.

Table 1

Progress of the PGK Refinement

Phase Set	Coord Set	No. of Atoms	R*	F.O.M.
ISOMORPHOUS 1	-	-	-	0.46
COMB1	PGK1	1969	0.56	0.54
COMB2	PGK2	2472	0.45	0.62
COMB3	PGK3	3016	0.32	0.73
COMB4	PGK4	3047	0.30	0.74
ISOMORPHOUS 2	Isomorphous phases after phase refinement of the derivatives using phase set COMB4			0.53
COMB5†	PGK4	3047	0.30	0.76
COMB6†	PGK6	3067	0.21	0.84

$$R^* = \sum_{hkl}^{all} |F_{obs} - F_{calc}| \Big/ \sum |F_{obs}|$$

† indicates combination of calculated phase with phase set ISOMORPHOUS 2.

5. DETECTION OF FEEDBACK IN THE COMBINED MAPS

The first type of Fourier synthesis used with the combined phases was based on coefficients $|F_{obs}|m \exp(i\alpha_{COMBINED})$, where m is a figure of merit weight. Subjectively these maps appear have greater clarity than the equivalent isomorphous map and two questions then have to be answered. First to what extent does the calculated structure feedback into the electron density in the places where the model is incorrect? Secondly can any improvement in the electron density be found in the areas where the interpretations are ambiguous?

To date two criteria have been found to be useful in assessing the level of bias in the electron density maps due to incorrect structure. The first of these involves making deliberate mistakes in the interpretation of the isomorphous map. The partially incorrect structure is then left to refine alongside the remainder of the interpretation. Then the electron density in a subsequent combined map is compared with that in the isomorphous map in order to determine whether or not the known electron density features reappear and whether there is any sign of spurious density surrounding the incorrect atomic positions. An example of this type of check is shown in fig.2, which shows the electron density in COMB2MAP (see Table 1) in the region of Tyr 323, whose position was known from the isomorphous map. The atoms shown are those used in the phase combination and it is clear that the phenol ring which was deliberately built in the wrong place is not surrounded by electron density. The true position for the phenol ring is quite clear and this density correlates well with the isomorphous map.

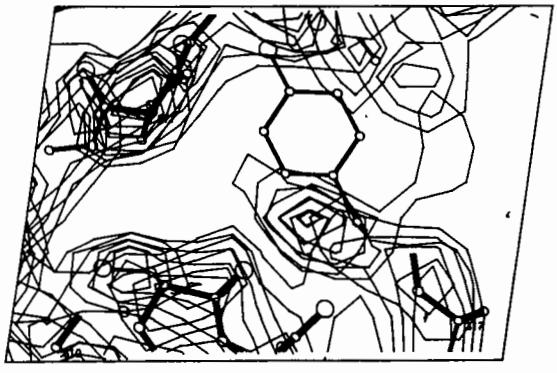


Fig.2

The second type of bias check that has been extensively used involves the examination of the refined model for atoms involved in high energy conformations or in excessively short van der Waals contacts. Clearly these atoms are likely to be incorrectly placed and therefore it only remains to observe the character of the electron density around these features and to see if this mirrors the unreasonable conformation or diverges from it in such a way that a reinterpretation is physically more reasonable. Figure 3 shows such an example taken from COMB3MAP. Here the amino group of the lysine side chain is in very bad van der Waals contact with the carboxyl group of a nearby Glutamic acid residue. Although the map does show a connection joining the lysine side chain to the glutamate residue there is a much more obvious feature indicating the true position for the lysine side chain.

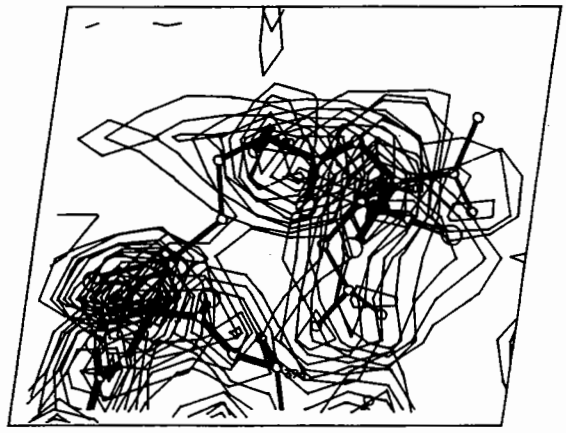


Fig.3

6. IMPROVEMENTS IN THE ELECTRON DENSITY

Improvements in the areas of weak electron density provide the best evidence for the power of the map enhancement technique and as an illustration of this it is instructive to follow the progress of the density improvement around Leu 313. Figure 4a shows an interpretation of the isomorphous map in this region. At the time the amino acid sequence for the chain 341-345 was not known and the isolated blob of density visible in this figure was assigned to the leucine side chain. Figure 4b shows the same area with the refinement at the stage COMB3. As can be seen from this figure the initially isolated density is now connected to two regions of main chain protein density. By this stage the complete amino acid sequence was available

and residue 347 was known to be a phenylalanine. Inspection of the map indicated that a reinterpretation of the electron density could be made in which the chain was grossly adjusted in the region 342-349, this modification resulting in the Phe residue occupying the position allocated to the leucine with the leucine side chain lying above its own main chain. This rearrangement would imply that the density at the β-carbon of the leucine was feedback from the incorrect structure. The map based on COMB3 phases in fact revealed many other less ambiguous errors where feedback was less noticeable and it was decided to leave the above region alone in order to see what happened to it as the refinement progressed.

Figure 4c shows the density at the stage of COMB5. Here the R value for the structure is 0.30 compared with 0.32 for COMB3. However a marked improvement in the electron density can be seen with the spurious connection in the region of the β-carbon of Leu 313 much reduced. A further improvement in the electron density was achieved by the use of a synthesis based on coefficients $2|F_{obs}|-|F_{calc}|\exp(i\alpha_{COMBINED})$ (see next section). The effect of using this synthesis as opposed to that using F_{obs} alone can be judged by a comparison of figs.4c and 4d; fig.4c shows the F_{obs} synthesis with COMB5 phases, fig.4d the $2|F_{obs}|-F_{calc}$ synthesis. The false connectivity has been completely broken and the carbonyl bulge associated with the position of the carbonyl oxygen of the Phe residue is more enhanced. Finally fig.4e shows the $2F_{obs}-F_{calc}$ map obtained with COMB6 phases following the reinterpretation and refinement of the structure.

Fig.4a

Fig.4b

Fig.4c

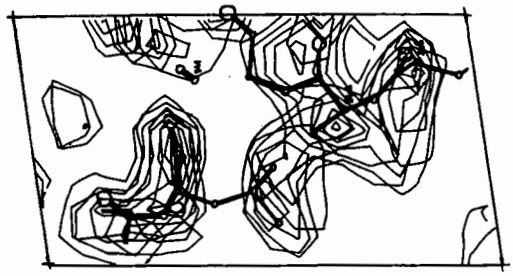


Fig.4d

Fig.4e

The conclusions then from this iterative use of the techniques of refinement and phase combination are that large initial errors in the model can be tolerated. Also, while the process is a gradual one of map enhancement, as the model is adjusted some spectacular improvements in electron density can be achieved. Carbonyl bulges in the electron density maps are now evident for almost all of the residues so that any rebuilding is greatly facilitated. In addition the peak height at the sulphur atoms in the structure are much higher than for the other atoms indicating that full use is being made of all the data available. The aromatic rings in the structure can be seen as flat discs rather than the spheroidal appearance they had in the isomorphous map, and the disturbances in the electron density near the heavy atom binding sites are no longer evident. Furthermore only one break (of 5 residues) in the course of the polypeptide chain is found and this is thought to be due to extensive disorder rather than any deficiency in the map. With this exception the molecular description is complete for the entire molecule.

7. A COMPARISON OF THE USE OF CALCULATED AND COMBINED PHASES

The use of a synthesis based on terms $(|F_{obs}|+n|F_{obs}-F_{calc}|)\exp i\alpha_{calc}$ has been described as an aid to refinement[18,19,20]. The basis of these maps is that the unknown part of a structure will be revealed in an $|F_{obs}|\alpha_{calc}$ or $|F_{obs}-F_{calc}|\alpha_{calc}$ map with peak heights $\leq 1/2$ the true value, depending on the error in the calculated phases. This follows from Luzzati[21]. Thus when virtually all of the structure is known a synthesis based on terms $(2|F_{obs}|-|F_{calc}|)$ ought to reveal the unknown structure with peak heights close to their true value. However the difficulty in a protein refinement at resolutions <2.5Å is that the initial interpretation of the structure, based on an isomorphous map, contains errors from several sources:

(i) grossly wrong structure.

(ii) random positional errors due to the limited resolution of the X-ray data.

(iii) little description of the ordered water molecules.

(iv) inadequate or no account taken of the solvent continuum.

Since these errors are all appreciable the unknown structure will appear in a $(2|F_{obs}|-|F_{calc}|)\alpha_{calc}$ map with much reduced peak heights. It would therefore be expected that such a synthesis would be unable to suppress the spurious density at incorrect atomic positions whilst also failing to reveal the as yet uninterpreted structure. An example of this may be seen in fig.5a which illustrates the region around Phe 291 in a map calculated with coefficients $(2|F_{obs}|-|F_{calc}|)\exp i\alpha_{calc}$ when the R value for the structure was 0.30. This map is to be compared with that shown in fig.5b which is an equivalent synthesis employing the combined phases. In the calculated phase map there is some electron density around the Phe ring position whereas there is none in the combined map. On the other hand, to the right of the current position of the Phe ring there is a very large feature in the combined map but only a small feature in the equivalent position of the calculated map. A reinterpretation of the model can then be made which moves the Phe ring from its current position into the electron density to its right. This change also alters the Ramachandran angles such that they move from a disallowed to an allowed region. Hence it would appear that, for refinements at 2.5Å where R is greater than 0.3, Fourier maps based on calculated phases alone contain a very significant level of bias such that large errors in interpretation will be missed. In fact in order to suppress the spurious electron density feature around the Phe ring a map (shown in Fig.5c) with coefficients $(5|F_{obs}|-4|F_{calc}|)\exp(i\alpha_{calc})$ had to be calculated. Although this map gave strong density at the true position of the aromatic ring many distortions of the electron density features and false connectivities could be identified due to its inherently higher noise level. Thus any interpretation of such a map required a degree of hindsight. The current interpretation of the structure is shown in fig.5d which is the $(2|F_{obs}|-|F_{calc}|)\exp(i\alpha_{COMBINED})$ map using the most recent set of phases.

Fig.5a Fig.5b

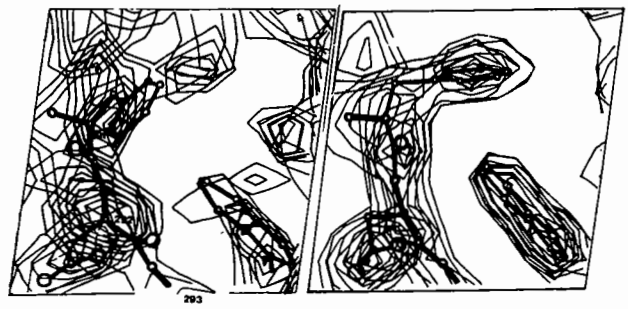


Fig.5c Fig.5d

The conclusions from this comparison, then, are that at this resolution it would have been very difficult, if not impossible, to reach the stage of interpretation now achieved with this molecule, using the calculated phases alone. The effect of combining the calculated and isomorphous phase information was to achieve a significant decrease in the phase angle error and to remove the bias associated with the calculated phases.

8. USE OF B VALUES AS A WEIGHTING FUNCTION IN PHASE COMBINATION

In a refinement at high resolution the temperature factor of an atom is often a guide to whether that atom is misbuilt or not. Thus whilst some side chains may be truly disordered and hence have high values, others may have high B values as a result of being in the wrong position. Clearly then the calculated phases would be improved if the parameters to observation ratio permitted individual temperature factors to be refined as the atoms not on true atomic positions would be automatically removed from F_{calc} and α_{calc}. However at 2.5Å a consideration of the parameters to observation ratio might suggest that such a process would not be feasible. In the most recent refinement of PGK (during the transition from COMB5 to COMB6) the atoms were allowed to have free temperature factors and following the convergence of the least squares the temperature factors were analysed in such a way as to test their physical significance. Surprisingly it was found that the B values were in most cases well behaved.

As an example of this the B values for the side chains of Arg 21, a well ordered semi-buried arginine, and Lys 85, an external lysine involved in a hydrogen bonding network, are shown below in fig.6. The good agreement of the B values for the arginine and the general increase along the length of the lysine make good physical sense.

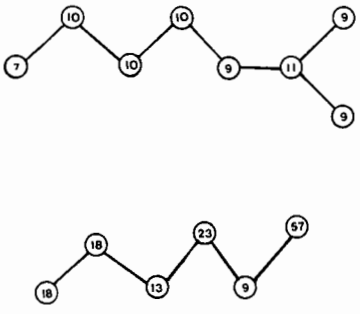


Fig.6. Temperature factors for Arg 21 and Lys 85.

One particularly interesting result was observed for the side chain of Glu 204. The temperature factors of the two carboxyl oxygens were very different, one being 20 the other 106. Inspection of the electron density map based on COMB6 phases revealed that the glutamate side chain was misplaced but that the oxygen atom with the low B value was sitting on an isolated peak. Since the level of bias in this map is very low, an explanation of this peak could be that it is an ordered water molecule and indeed its position is consistent with that finding, there being two carbonyl oxygen atoms 3Å away from this peak. The electron density in the above map and the temperature factors for the atoms are shown in fig.7.

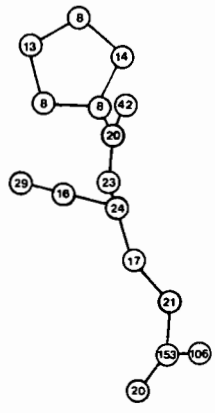


Fig.7

The temperature factors are clearly sensible parameters even at this resolution and although in the initial stages of a refinement it may be wise to use group thermal parameters, the use of individual B values later on appears to be both justifiable and useful at this limited resolution.

9. CONCLUSION

In the study of an enzyme structure the use of a refinement technique is a great aid in obtaining a more accurate set of atomic parameters. However at limited resolution the problems caused by feedback from incorrectly located atoms can give rise to a situation where a false minimum is reached and cannot be easily by-passed. By the combination of the isomorphous and calculated phase information it has been possible to overcome this situation and achieve a successful refinement for PGK which has proved both satisfying in terms of revealing many errors of interpretation and rewarding in providing much valuable information on the functional aspects of this molecule[14]. By the coupling of the refinement of temperature factors to weight out incorrectly placed atoms, and the use of mixed syntheses such as $(2|F_{obs}|-|F_{calc}|)$ with the combined phases, the technique described here has been convincingly demonstrated to have great potential in the field of protein structure determination.

REFERENCES

1. W. Hoppe and J. Gassmann, Acta Cryst. B24, (1968) 97.
2. D. Sayre, Acta Cryst. A30, (1974) 180.
3. W.A. Hendrickson, J. Mol. Biol. 91, (1975) 226.
4. C. de Rango, Y. Mauger and G. Tsoucaris, Acta Cryst. A31, (1975) 227.
5. R.C. Agarwal and N.W. Isaacs, Proc. Natl. Acad. Sci. U.S.A. 74, (1977) 2835.
6. M.G. Rossmann and D.M. Blow, Acta Cryst. 14, (1961) 631.
7. W.A. Hendrickson and E.E. Lattman, Acta Cryst. B26, (1970) 136.
8. D.M. Blow and F.H.C. Crick, Acta Cryst. 12, (1959) 794.
9. W.A. Hendrickson, Acta Cryst. B27, (1971) 1472.
10. G.A. Sim, Acta Cryst. 12, (1959) 813.
11. G.A. Sim, Acta Cryst. 13, (1960) 511.
12. G. Bricogne, Acta Cryst. A32, (1976) 832.
13. C.C.F. Blake and P.R. Evans, J. Mol. Biol. 84, (1974) 585.
14. R.D. Banks, C.C.F. Blake, P.R. Evans, R. Haser, D.W. Rice, C.W. Hardy, M. Merrett and A.W. Phillips, Nature 279, (1979) 773.
15. D.W. Rice, Acta Cryst. (1980) in press.
16. W. Cochran, Acta Cryst. 4, (1951) 408.
17. E.J. Dodson, N.W. Isaacs and J.S. Rollett, Acta Cryst. A32, (1976) 311.
18. R. Huber, D. Kukla, W. Bode, P. Schwager, K. Harteb, J. Deisenhofer and W. Steigemann, W. J. Mol. Biol. 89, (1974) 73.
19. P. Main, Acta Cryst. A35, (1979) 779.
20. M. Vijayan, Acta Cryst. A36, (1980) 295.
21. V. Luzzati, Acta Cryst. 6, (1953) 142.

SOME REFINEMENT EXPERIENCES WITH 2Zn INSULIN

by

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

One of our motives for extending the resolution of the 2Zn insulin crystal structure was to get a useful description of the electron density of the water-containing regions. A number of water molecules had been placed in the isomorphously phased 1.9A spacing map but difference Fourier calculations at this resolution indicated that some of these were incorrect. Refinement at 1.5A spacing has enabled us to identify a large proportion of the water molecules in the crystal pretty satisfactorily, though there is uncertainty in some cases as to the occupancy and thermal parameters.

Our procedure has been to place water molecules where the electron density, in different maps, was persistent and led to sensible contacts. For the well-defined molecules, these features were typically spherical and examples of these attached directly to the protein can be seen in Figure 1. Oxygen atoms refined at these positions had thermal parameters not much different from those of the adjacent protein atoms. Away from the protein, the electron density peaks are lower and sometimes extend into channels of density with no obvious maximum, as also can be seen in Figure 1. Defining positions in this density is arbitrary since first the water molecules are probably distributed in a statistical population along the density. And, secondly, although the overall appearance of the electron density is largely preserved, it is subject to the background fluctuations which alter the position of the maxima in the different maps employed to position the water molecules (Figure 2). A set of positions has nonetheless been derived which match the electron density

(continued on page 98)

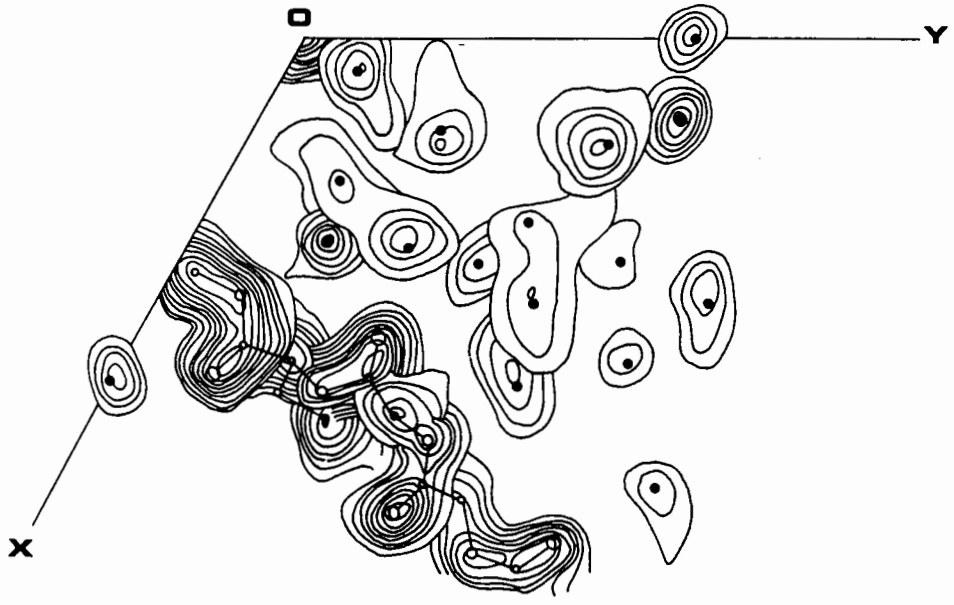


FIGURE 1: The electron density from the final systematic difference Fourier density for the 2Zn insulin crystal between sections 26 - 18/72 along the C axis. Water molecules are represented as filled-in circles. The peptide chain extends from B2 to B6.

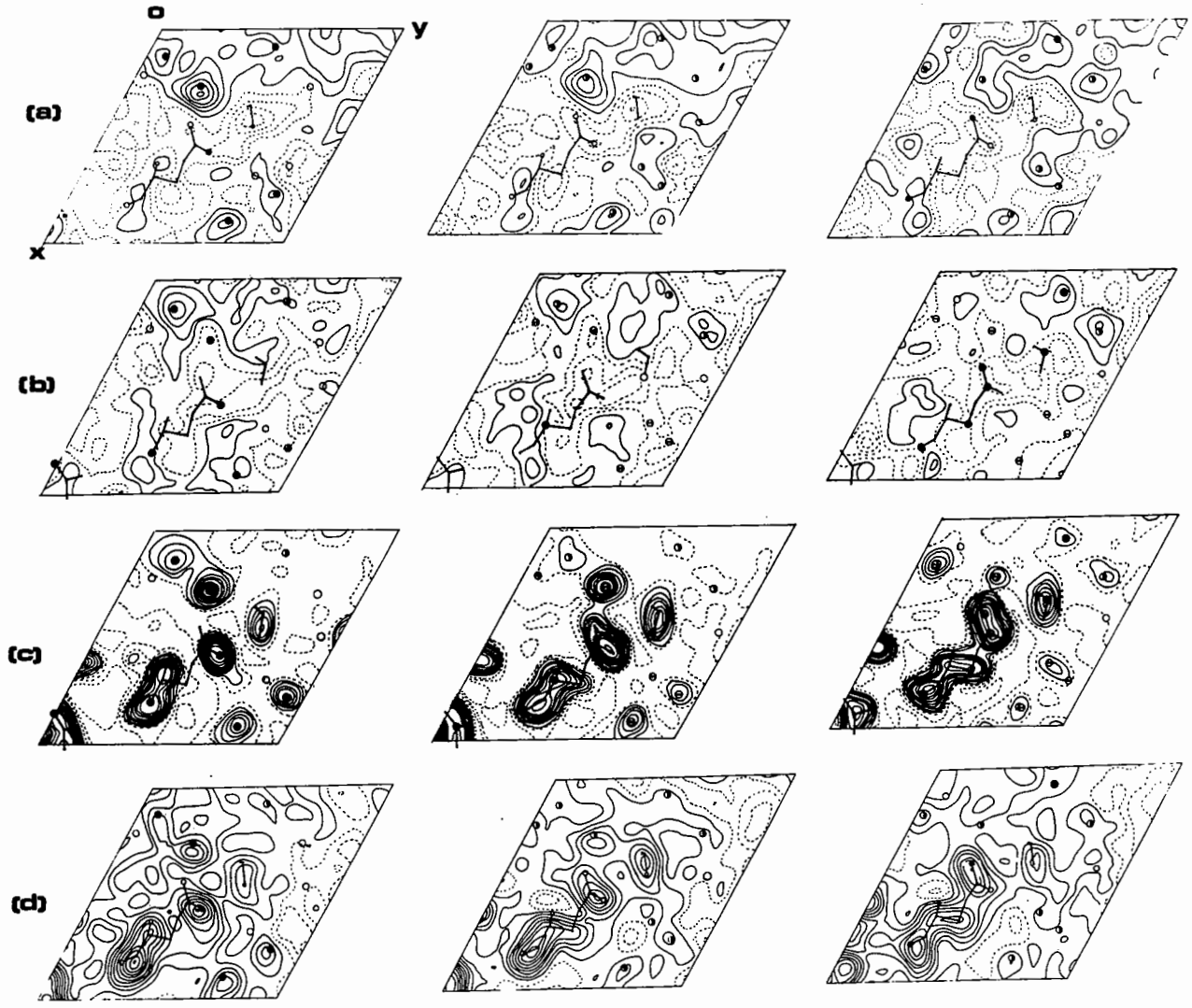


FIGURE 2: Sections 38, 39 and 40/72, showing the electron density in

(a) the difference Fourier with all water molecules and doubtful atoms excluded from the phasing calculated at the beginning of the fast Fourier least squares refinement.

(b) The subsequent difference Fourier in which 110 water molecules were included in the phasing. One cycle of refinement (3(x,y and z) + 2B) has been carried out before the map was calculated.

(c) A Fourier map calculated at the same stage as (b).

(d) A systematic difference Fourier calculated near the completion of the refinement. In this procedure the electron density in each 1/8 of the unit cell is phased by the other 7/8.

The contour levels in each map are at .1eÅ^{-3}; the protein atom and water positions are those at the completion of the refinement.

Water molecules' positions are indicated by circles.

● represents the position of an unphased water <.5Å from the section;

⊕ represents the position of a phased water <.5Å from the section;

◑ represents the position of an unphased <.1Å + >.5Å from the section;

⊖ represents the position of a phased <.1Å + >.5Å from the section;

O represents the position of a water <1Å from the section.

FIGURE 2 (continued)

Protein atoms are joined by bonds.

In (a) and (d)

● represents oxygen atoms' positions <.5Å from the section;

● " nitrogen " "

O " oxygen " >.5Å "

o " nitrogen " "

No representation is given for carbon atoms.

In (b) and (c)

⊕ represents atomic positions (oxygen, nitrogen and carbon) <.5A from the section.

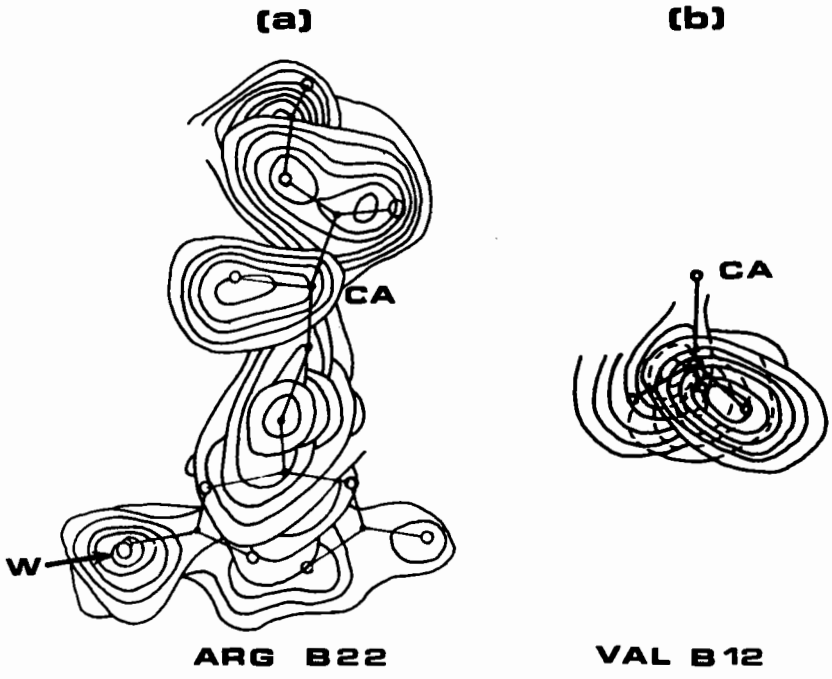


FIGURE 3: Disorder

Disorder, dynamic or statistical, can be detected in electron density maps if the conformations are few and well defined and more than 1/3 populated. In the 2Zn insulin analysis, a number of cases of disorder have been identified involving mostly charged sidechains (arginine, lysine, and glutamic acid), though a non-polar group, valine, exhibits this behaviour (b). Figure 3 (a) illustrates the electron density for one of the arginine sidechains on the systematic difference Fourier calculation near the completion of the refinement. The higher electron density peak at one of the four NH positions is attributed to the presence of a complementary water molecule (indicated by W). It makes the same H bond as the NH it replaces when the arginyl sidechain assumes the alternative conformation.

It is important that only the most definite evidence is accepted for a disordered conformation since the refinement of half weight atoms is often misleading. For this reason, review of disorder in difference and systematic difference Fouriers is essential.

acceptably, and which conform to the H-bonding and contact requirements of water molecules.

Most of the water molecules have been given full occupancy except where (for example, see Figure 3) the evidence for disorder or alternative structures was secure. Thus, the set of water positions chosen in the poorly defined electron density represents only one of the possible local structures of comparable population assumed by the solvent.

Support for the observations made directly from the maps illustrated in Figures 1 and 2 comes from an alalysis of the water molecule thermal parameters and electron density. In Figure 4, the electron density value for a selection of water molecules is plotted along the ordinate. It was obtained from the average peak value in the difference Fourier when all water was excluded and the systematic difference Fourier near the refinement's completion. The thermal parameter for each water at completion of the refinement is plotted along the abcissa.

There is a generally sensible correlation between the average value of the electron density and the thermal parameter which improves as the electron density increases. The wide spread in thermal parameters at low values of the electron density is a reflection of the errors, illustrated for example in Figure 2, present in the maps.

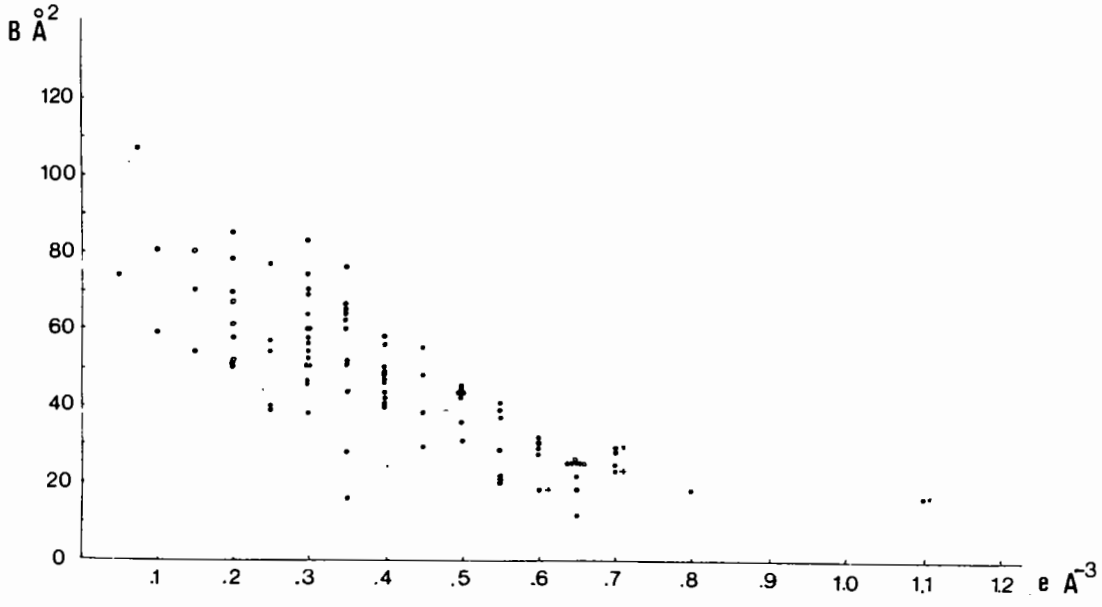


FIGURE 4:

The refined thermal parameter ($B\text{Å}^2$) and the electron density at RD17 (averaged from the difference Fourier and the systematic difference Fourier) for a selection of water molecules in the 2Zn insulin crystal.

SUMMARY OF THE GENERAL DISCUSSION PERIODS

by

John W. Campbell and Mike Elder
Daresbury Laboratory

1. INTRODUCTION

In this chapter we attempt to summarise the main points of the various discussion periods. The major Saturday afternoon discussion session is dealt with in a preceding chapter. A number of points came up in various guises throughout the small discussion periods totalling some four hours, which punctuated the meeting. We have found it easier to gather these together under a number of headings rather than to associate them with the individual papers. To some extent these headings reflect our understanding of what was said and we apologise for any omissions or misplaced emphases. At risk of offending the many contributors to the discussion periods we have decided not to attribute the various contributions. This has enabled us to summarise and abbreviate sometimes lengthy discussions and to avoid the embarrassment of wrong attributions, or even worse, crediting our misunderstanding to some otherwise innocent speaker!

2. CONSTRAINTS

(1) Whilst the power of restraints in protein structure refinement was generally accepted, it was felt that the use of constraints has certain advantages:

a) The number of parameters being refined can be reduced. This is especially valuable at low resolution where the number of X-ray observations is low.

b) Larger shifts, especially of complete sections of a protein, can be more readily accommodated, thus giving a larger radius of convergence for the refinement. Again, this feature is of particular value at low resolution.

c) It is easier to maintain planarity of groups than by using distance restraints. It would probably be realistic and desirable to maintain the geometry of aromatic rings throughout most refinements. It would therefore be desirable to have some provision for using certain constraints within restrained refinement methods.

(2) Certain disadvantages were also identified:

(a) The errors will all tend to accumulate in certain 'garbage collecting' parameters (c.f. the Tau angles in Diamond's real space refinement). In contrast, the errors tend to be spread more evenly throughout the structure when a restrained method is used.

(b) It may be very difficult to escape from false minima resulting from a constrained refinement when a restrained refinement is carried out at a higher resolution. More manual intervention may therefore be required to correct false minima before moving to higher resolution.

(3) Various experiences were reported of fitting sections of a known protein structure to a low resolution map of a related structure using computer graphics and it was felt that Sussman's constrained refinement method for nucleic acids [see his paper] could be usefully applied to fitting and refining protein structures.

3. RESTRAINT REDUNDANCY

(1) Concern was expressed that redundant information may be introduced when restraints are used. Certain regions of the molecule then become incorrectly weighted. For example, the same inter-atomic distances may be used to restrain bond lengths, bond angles and ring planarities; chiral volume restraints may be used to preserve the geometry at chiral centres for which all the parameters are also restrained in terms of bond lengths and angles. Thus the restraints, which are in effect observations in the context of the refinement, do not constitute an independent set of observations.

(2) In response to this concern, it was suggested that it is important to distinguish between the process of refinement and the analysis of the results at the end of a refinement. During the course of a refinement, the over-specification of restraints probably acts as an aid to the process. In particu-

lar, chiral volume restraints provide a strong barrier to inversion at tetrahedral centres which cannot be matched by any but the most rigid bond angle restraints.

(3) It is theoretically possible to treat the correlation between restraints properly. The relationships between the correlations in parameters and observations and the relevant matrices can be expressed as follows:

(a) Off-diagonal terms in the normal matrix represent correlations between the parameters.

(b) Off-diagonal terms in the variance-covariance matrix of the observations represent correlations between the observations.

Normally only the diagonal terms are included in the least squares observations weighting matrix on the assumption that the observations are independent but there is no reason why the off-diagonal terms expressing the correlations between the observations should not also be included. The computational problem involved would however be very large.

4. NON-CRYSTALLOGRAPHIC SYMMETRY

(1) An exact use of non-crystallographic symmetry, e.g. by averaging symmetry related regions of a difference map, is valuable when working at low resolution and during the early stages of a refinement as it reduces the number of parameters and probably increases the radius of convergence for the refinement.

(2) The restrained method also allows for the treatment of non-crystallographic symmetry but differs from other methods, such as map averaging, in that it will make allowances for some differences between the related sub-units.

(3) The use of non-crystallographic symmetry in the restrained method increases the number of observations, without any change to the number of parameters. The size of the computational problem (e.g. structure factor calculation) is therefore greater than that associated with the strict application of non-crystallographic symmetry which in general allows for a reduction in the number of parameters and therefore in the size of the problem.

5. HYDROGEN BONDS

(1) When using non-bonded contacts as restraints in a refinement it is necessary to allow for hydrogen bonds as these allow non-bonded atoms to come closer than the normal Van der Waals' distance. For this reason, potential hydrogen bonds are identified in Hendrickson's refinement method [see his paper] and separate distance restraints are applied to them. Such hydrogen bonds are identified only by the types of the non-bonded atoms and without reference to the geometry involving the hydrogen atom.

(2) In a more general context, it was pointed out that the calculations of theoretical quantum chemists would suggest that the protein crystallographer's traditional view of what constitutes a hydrogen bond is probably too conservative and that significant hydrogen bonding effects can take place with hydrogen bond distances as great as 4 Å or with angles as low as 150 degrees at the hydrogen as atom.

6. ENERGY REFINEMENT

(1) In energy refinements, hydrogen atoms play a more important part than in X-ray refinement. In particular, parameters depending on non-bonded contacts involving hydrogen atoms are likely to be better defined from energy calculations than from X-ray refinement.

(2) The present methods for modelling energies are in many ways inadequate. Also, important terms such as long range electrostatic interactions are normally ignored in energy calculations though they may have significant impact on the structure. It is fortunate that long range effects are well treated by X-ray refinement and it would seem to be generally true that energy refinement and X-ray refinement provide complementary information.

(3) Though there is good evidence of the value of doing energy as well as X-ray refinement, it is not necessary that they should be combined in one calculation. Provided that the matrices from the separate calculations were retained, it would be possible to calculate the most probable structure based on the results of the two independent refinements.

7. WEIGHTING SCHEMES

(1) When considering the problem of the correct relative weights to give to the X-ray data and the various restraints in a restrained refinement, it is important to distinguish between the weighting requirements in the final stages of refinement and the weighting requirements during the course of the refinement. Meaningful interpretation of the results at the end of a refinement is only possible if correct weights are used in the latter stages. During the course of the refinement, however, pragmatic considerations are the most important. In particular, it is unsatisfactory to weight reflection data using σ values based on counting statistics. Unit weights give a better refinement and perhaps the most helpful weighting is obtained using a sloping function based on the sin θ/λ values for the reflections.

(2) The question was raised as to where the bond lengths and angles to be used in restrained or energy refinement should be found and, more importantly where the values to be used as weights or force constants should come from. It was observed that average values and their distribution about this average could be obtained from a survey of relevant small molecule structures such as are available in the Cambridge Crystallographic data base.

(3) At the end of a refinement the X-ray weights need to be put on a correct scale relative to those for the restraints. The problem of doing this has not yet been adequately solved.

(4) The basic approach to the weighting of energy refinement results and restrained X-ray refinement results is through probability considerations. From the energy point of view, the probability of a given structure can be calculated using Boltzmann statistics and for the restrained X-ray refinement the probability can be calculated, at least in theory, using Gaussian statistics.

8. THE TREATMENT OF SOLVENT

(1) It was generally felt that the ability to represent solvent structure is an important factor in improving the refinement of the protein itself and that the placing of individual solvent molecules is valuable even if the complete solvent structure cannot be described. Much detailed work has been done on the solvent structure of Insulin including a careful analysis of the contents of the unit cell. All the water structure (283 molecules) has been described, mainly in terms of individually placed oxygen atoms with B values from protein oxygen values to around 80.

(2) Where it is not possible to place individual water molecules, it may be profitable to include uniform electron density for the regions of solvent for which the structure cannot be described in detail. Some experiences have shown however that only marginal improvements take place and those only in the very low resolution terms.

(3) Though the computer simulation approach to describing solvent structures using Monte Carlo techniques has both in Hermans' results [see his paper] and in Finney's coenzyme B12 results, given some correspondance between predicted and observed closely bound solvent structure, there are many problems to be solved before the method will be of real use to protein crystallographers. In particular the problem of choosing appropriate potential functions, which model accurately the relevant properties of water, is far from being solved and this is critical because small differences in the potential functions used tend to give large differences in the predicted solvent structure.

(4) The question of whether the presence of salt inhibits the formation of extensive solvent structure was raised but no evidence was presented in reply.

9. THERMAL EFFECTS

(1) The correlation of temperature factors is much stronger in the direction of bonds than perpendicular to bonds. A model calculation by Hendrickson assuming an RMS bond length restraint of say 0.05 Å gave the following results for the possible variation of temperature factor along an extended Lysine side chain:

(a) Isotropic model:

4.0	4.2	4.4	4.6	4.8	5.0

(b) Anisotropic model:

4.0	7.4	10.7	14.1	17.4	20.8

It can be seen that the anisotropic model allows for a much greater variation in temperature factor along a chain and is therefore more suitable, though still inadequate, for describing a side chain which is well anchored at one end and relatively free to move at the other.

(2) Other approaches to the temperature factor description of proteins need to be considered. For example, TLS modelling may be appropriate. Here a side chain may be treated as a rigid body with freedom to translate and librate. Looking further into the future, it may be possible to address the problem of thermal effects by considering molecular dynamics of the whole protein.

(3) When considering the effects of thermal motion, the possibility of alternative conformations being present should also be considered. A side chain which alternates between two conformations may give difficulties in refinement if an attempt is made to describe it in terms of a side chain vibrating about the average position. A particular problem will occur if such an average position is not sterically feasible because of unfavourable non-bonded contacts.

10. FFT METHODS

(1) At the present time, Agarwal's FFT refinement method [see his paper] only deals with isotropic temperature factors but there is no reason why it should not be extended to cope with anisotropic temperature factors.

(2) The approximations made in Agarwal's method are probably satisfactory provided that the matrix is well conditioned but concern was expressed over the case where this is not so. It was felt that if restraints were introduced into the method then the matrix would be likely to be less well conditioned. The regions over which the approximations are valid need to be explored in some detail.

(3) When convoluting atomic densities with difference Fourier maps, the Gaussian approximations to the atom shapes should not be modified for resolution; it is sufficient to take the Fourier terms as zero after the resolution cut-off.

11. PHASE IMPROVEMENTS

(1) Bhat's phase combination method [see his paper] is capable of making significant changes and improvements in an electron density map compared with that obtained using MIR phases alone. The phase combination approach is especially valuable where sequence information is not known or where it is not well known.

(2) In contrast, no dramatic results have been obtained using direct methods to refine or extend sets of phases where the answer was not already known. Sayre's phase extension method on Rubredoxin and 2 Zn Insulin did give fairly sensible results especially when the generated phases were held close to the isomorphous phases where available. A 1.5 Å 2 Zn insulin map calculated using phases extended from 1.9 Å had clear maxima for many of the atoms and it was easy to assign coordinates to these. However the connectivity of the protein was not always visible, i.e. some of the atoms had disappeared, and it was not straightforward to interpret such a map. The matrix methods of Tsoucaris and de Rango also generated phases to 1.5 Å on Insulin which agreed well with the calculated model phases when the isomorphous phases were held fairly closely. However they have not had much success when starting with lower resolution sets or less accurate sets of isomorphous phases.

(3) Direct methods may prove to be useful in a limited context, particularly in such areas as finding improved positions for water molecules.

(4) When attempting to use direct methods, experience has shown that it is necessary to preserve some of the MIR phases strictly throughout the phase extension or refinement process.

12. PRACTICAL ASPECTS OF REFINEMENT

(1) The need for frequent manual intervention during a protein structure refinement was stressed. This involves the inspection of Difference Fourier maps and model building which can most effectively be done using a computer graphics system. It is normal practice to inspect the complete model during manual intervention though it is probably a matter of personal preference whether minor adjustments are made

throughout the structure as well as correcting the more gross errors. A useful facility, which is not at present available in the restrained refinement method, would be the ability to specify which parts of the model are in good agreement with the electron density map and to put an extra restraint on moving such parts.

(2) Both the speed of a refinement and the volume of parameter space explored are probably increased by using some alternation between X-ray based and geometry based refinement. At one extreme, this may be done using an X-ray refinement program in alternation with a model building program. A similar, though less extreme, solution is to vary the relative weighting of the X-ray and geometry restraint terms in a restrained X-ray refinement. This latter approach enables a more even refinement than is possible with the strictly alternating approach.

(3) The concept of alternation is also important as a means of looking for problem areas within a structure. Errors in interpretation are probably indicated if the shifts produced by giving a high weight to the X-ray terms are in opposition to those produced by restraining the geometry. The use of this technique cannot however substitute for regular manual intervention as described above.

(4) One general worry about using restraints in refinement is that the emergence of new information will tend to be suppressed if each structure is restrained to results based on previously determined structures which have themselves been restrained.

(5) It was noted that the inclusion of the X-ray data off-diagonal terms between different atoms in a restrained refinement did not give any improvement in the speed or results of the refinement. Though the inclusion of off-diagonal terms in X-ray refinements is valuable in small molecule refinements, it was felt that their inclusion in Agarwal's FFT refinement method would probably be of little benefit when dealing with large and relatively poor models, except perhaps at the end of the refinement.

(6) A request was made, on behalf of those about to start on restrained refinement, for more practical advice and examples of actual numbers used in weighting schemes etc. It was hoped that groups with experience in this field would be able to make contributions for this purpose.

13. THE ACCURACY OF THE RESULTS

(1) To give a feel for the quality of the model produced at the end of a refinement, it would be useful to devise a geometry factor which could be presented in parallel with the standard crystallographic R-factor. This would give an indication of how far the model deviates from the ideal geometry used to restrain it.

(2) Even more importantly it is necessary at the end of a refinement to be able to estimate, meaningfully, the accuracy of the results obtained. Though in principle the standard deviations of the final parameters can be derived from the normal matrix of the least squares, there are still in practice problems of weighting to be solved.

(3) More qualitatively, some feel for the accuracy of the results of a refinement may be obtained by observing the effects on the model when restraints are relaxed.

(4) In some cases information on the accuracy of results has been obtained from different structure determinations of the same protein containing, for example, different inhibitors. It would be unwise, however, to assume that the X-ray data sets measured for the different structures are truly independent, especially if the protein always crystallises in the same space group.

(5) The importance of considering the precision of the observed X-ray data was also stressed and care should be taken when making corrections, such as absorption corrections, to data which are to be used for accurate refinement.