DL/SCI/R27

MOLECULAR SIMULATION AND PROTEIN CRYSTALLOGRAPHY

Proceedings of the Joint CCP4/CCP5 Study Weekend,
27 - 28 January, 1989

Compiled by J. Goodfellow, K. Henrick and R. Hubbard

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

ISSN 0144-5677

MOLECULAR SIMULATION AND PROTEIN CRYSTALLOGRAPHY

Proceedings of the Joint CCP4/CCP5 Study Weekend
27 - 28 January, 1989

Compiled by
J. Goodfellow, Birkbeck College, London
K. Henrick, Daresbury Laboratory
and
R. Hubbard, University of York

SCIENCE & ENGINEERING RESEARCH COUNCIL
DARESBURY LABORATORY
1989

PREFACE

The past ten years have seen the development of a wide range of computational and theoretical approaches to the simulation of macromolecular structure. Most of these techniques are based on the empirical energy functions from which the energy (and thus the forces) of the bonded and non-bonded interactions (van der Waals, electrostatic etc) of the molecule can be evaluated. For example, molecular dynamics calculations have provided great insight into the structural dynamics of proteins and its relevance to structure and function.

Recently, these techniques have been increasingly used not to simulate some aspect of the physical behaviour or function of the molecule, but as a computational tool to satisfy experimental constraints. The important advance was the incorporation of an extra penalty term into the energy function so that, for example, movement of the distance between two atoms from a target distance is energetically unfavourable. In this way, the protein structure can be forced to satisfy the experimentally derived constraint. For the crystallographic community, the significant step was the incorporation of a penalty function representing electron density which has led to crystallographic refinement using molecular dynamics calculations.

The first CCP4 Daresbury Study Weekend on protein structure refinement in 1980 marked the maturing of protein refinement techniques and helped to catalyse the general application of these techniques by the crystallographic community. Our aim in the 1989 study weekend was to review progress in the use of molecular dynamics in structure refinement and analysis. The meeting again proved timely. The conclusion that came out of the weekend was that programs such as MDREF and XPLOR had the power to remove most of the tedious model building associated with refinement and perhaps more excitingly gave an opportunity to explore multiple conformations in refinement. Providing the community has access to enough computer time, the techniques will clearly make a contribution to the rate and hopefully quality of structure refinement.

The meeting was organised and supported by the SERC Collaborative Computational Projects in Protein Crystallography (CCP4) and in Computer Simulation of Condensed Phases (CCP5) at Daresbury Laboratory. We wish to thank the invited speakers for their considerable efforts in making the meeting a success and their cooperation in the preparation of these proceedings. Particular thanks go to Julia Goodfellow and Rod Hubbard for the considerable time and effort they invested in the planning of the meeting.

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

Kim Henrick
July 1989

CONTENTS

PROTEIN STRUCTURE REFINEMENT BY MOLECULAR DYNAMICS TECHNIQUES

P. Gros, M. Fujinaga, A. Mattevi, F.M.D. Vellieux, W.F. van Gunsteren and W.G.J. Hol

BIOSON Research Institute
Department of Chemistry
University of Groningen
Nijenborgh 16
9747 AG Groningen
The Netherlands

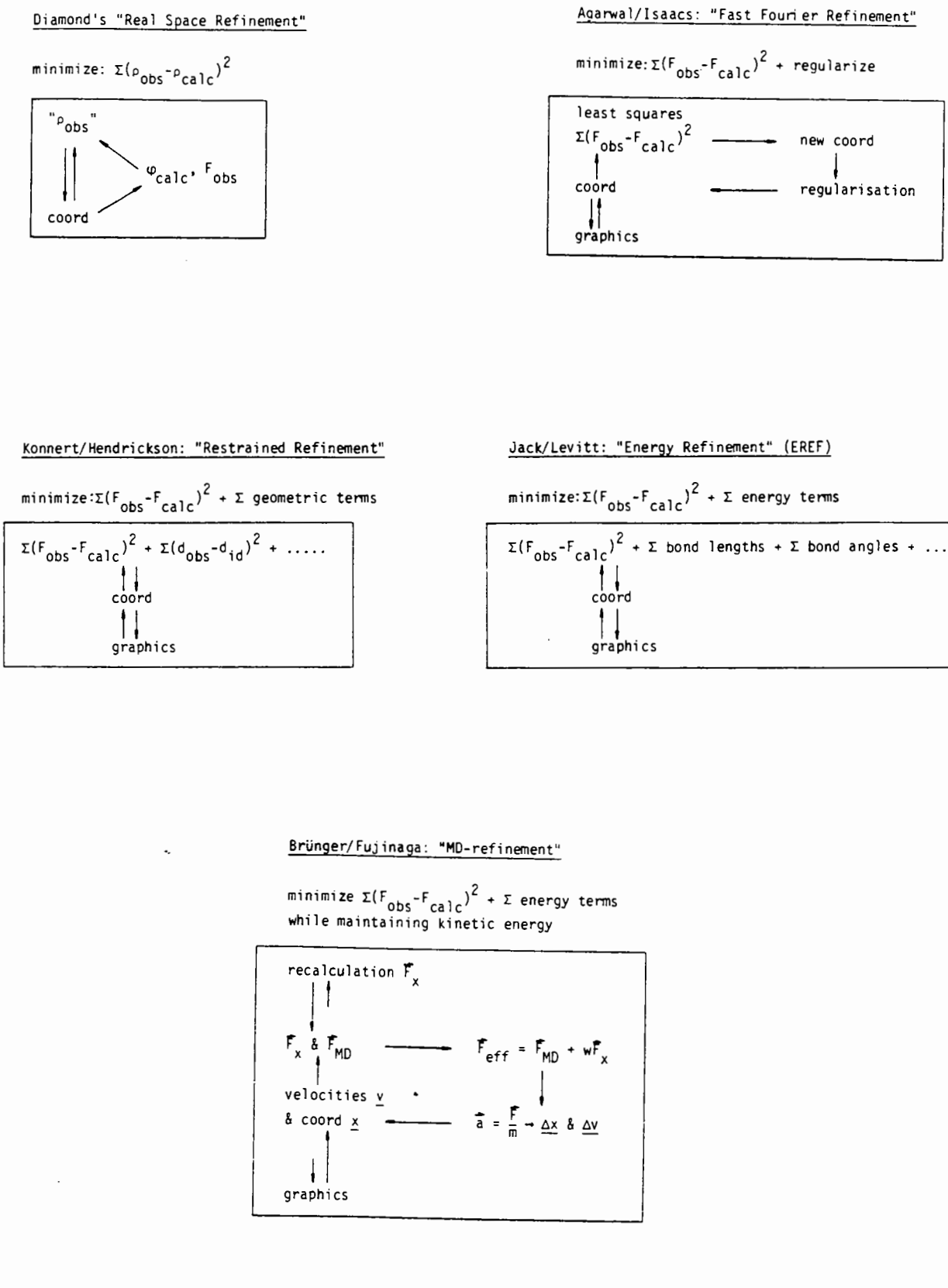
1. Introduction

Since the early seventies protein crystallographers have been trying to obtain a model of the protein under investigation agreeing as closely as possible with the observed data. In spite of the thousands to hundreds of thousands observations this is almost always an underdetermined problem because there are also thousands to hundreds of thousands parameters to be refined. Hence, from the very beginning geometric constraints and restraints have played a major role in obtaining a structure which is at the same time agreeing with stereochemical knowledge and with the observed intensities.

Early attempts included difference Fourier refinement alternated with "regularisation", that is imposing "ideal" geometry [1,2]. Gradually the real space refinement procedure of Diamond [3], applied in a cyclic manner using gradually improved phases [4], became an important tool in refinement. In addition to being computationally expensive, the number of degrees of freedom were quite limited in the sense that rigid fragments were rotated about atomic bonds and only very few bond angles were allowed to vary. A much more rapid procedure became available when Agarwal [5] developed a "Fast Fourier refinement" procedure where structure factors and gradients were calculated via maps on grids in real space and the diagonal matrix approach was used in solving the least squares problem. Shifts were applied in a careful manner and the structure was regularised by the procedure of Dodson et al. [6]. At approximately the same time Konnert & Hendrickson's restrained refinement [7] procedure became very popular. The program minimized simultaneously $\Sigma(F_o-F_c)^2$ with restrictive geometric terms. In later versions of the program more terms were added such as phase restraints, temperature factor restraints and non-crystallographic symmetry restraints. It was this tool-kit of extra facilities which contributed much to the wide spread use of the program, although the Jack-Levitt [8] program had the same philosophy with fast fourier procedures for structure factor and derivative calculations, and perhaps a physically more realistic force field.

In spite of the sophistication of these programs, and of related ones such as written by Tronrud et al. [9], they remained least squares procedures where only the nearest local minimum from the starting coordinate set will be reached. Recently, A. Brünger and co-workers have introduced a new method, where procedures from molecular dynamics are used to explore large regions of conformation space while at the same time trying to minimize the difference between observed and calculated structure factors [10,11]. A similar procedure has been incorporated by M. Fujinaga and P. Gros in the molecular dynamics package GROMOS written by W.F. van Gunsteren [12,13]. Some experience with the use of this package will be described in the present paper. At the end of this contribution we will look back to the first attempt to use molecular dynamics techniques in protein structure refinement where we aimed at a completely general description of thermal

Figure 1 Simplified flow diagram of a number of techniques used in protein structure refinement



motion of protein and solvent atoms [14].

2. Principles and important parameters of molecular dynamics refinement

The principle of molecular dynamics refinement is quite simple. It consists of adding an extra term called the "X-ray energy" to the various conventional energy terms of a molecular dynamics calculation of a protein. The total potential energy term then becomes:

$$\frac{1}{\sigma_x^2} \sum_{\underline{h}} (k|F_{calc}(\underline{h})| - |F_{obs}(\underline{h})|)^2 + \text{Energy terms.} \qquad (1)$$

This X-ray energy gives rise to an "X-ray force" on each atom which is added to the normal molecular dynamics force affecting this atom. From this force, an acceleration is obtained which in its turn gives a new position and new velocity for each atom a very small time step later. This procedure is repeated for a large number of steps while the temperature of the system is kept at a constant value, for instance by coupling to a thermal bath [15]. In this manner the system becomes quite flexible and a large number of conformations is explored, the kinetic energy of the system allowing the overcoming of barriers. The latter point is a distinct difference when compared with classic least squares procedures where atoms shift along lines of steepest descent and remain in the first local minimum found.

There are a number of important parameters which need to be considered carefully in the molecular dynamics refinement process:

2.1 The "temperature" at which the calculations are carried out. Higher temperatures allow crossing of larger barriers, searching a large area of conformation space, but one also has to take smaller time steps in the calculations in order to avoid too large a geometric distortion of the molecule which would result in too large forces and unstable behaviour.

2.2 The amount of time during which the molecular dynamics refinement is allowed to run. Obviously the longer a run lasts the more conformations can be searched, depending on the strains and errors in the molecule. When starting models contain large errors, long time periods may allow large corrections to be made automatically. If, however, the refinement has virtually been completed, longer periods may only involve fluctuations about equilibrium positions. Such calculations are interesting from a temperature factor point of view, see section 8.

2.3 The relative weights of the X-ray and energy terms, or the choice of σ_x in equation 1. Restrained least squares procedures [7] have suggested that useful results are obtained when σ_x is taken is $\frac{1}{2}\{\Sigma(F_o-F_c)^2\}^{\frac{1}{2}}$. This implies that the X-ray term in equation (1) is also independent on the scale of the structure factors. It appears that it is useful to bring σ_x in the order of this value but that it often is appropriate to lower the weight of the X-ray term i.e. using values of σ_x which are larger than $\frac{1}{2}\{\Sigma(F_o-F_c)^2\}^{\frac{1}{2}}$.

2.4 The resolution range of the X-ray data. Decreasing the resolution by, for instance, omitting terms beyond 4 Å, decreases the X-ray term in equation (1) simply because there are less terms, while the conventional MD energy terms remain essentially unaltered. In addition, omitting high resolution terms gives broader density maps and less steep gradients, allowing larger motions to occur. It is obvious that taking less high resolution terms into account means also that less computer time is needed.

2.5 The force field employed. The molecular dynamics part of the calculations is carried out in vacuum - at least bulk solvent atoms are omitted. Hence parameters which have fully charged groups can easily give erroneous results because of ion pair formation occurring in the

computer solely because of lack of dielectric screening by the solvent. Hence, the charges and dielectric constant used are of great importance. Also the parameters have to be considered carefully as artificially small Van der Waals radii, or soft Lennard Jones potentials, for instance, allow searching larger volumes of space but may also lead to atoms ending up in wrong density.

2.6 The update frequency of the X-ray forces, x_F. In our program the strategy of Brünger [10] for the update of X-ray forces has been adopted: once an atom has shifted by more than x_F Å the derivatives are recalculated explicitly, if motions remain smaller than x_F Å the derivatives are updated by a harmonic approximation using second derivatives as suggested by Jack & Levitt [8].

3. An initial test: phospholipase A_2

Bovine pancreatic phospholipase A_2 is a protein of 123 residues whose structure has been refined at high resolution by Agarwal's fast Fourier procedure [16]. The space group is $P2_12_12_1$, with one molecule per asymmetric unit. Starting from a model built in a m.i.r. electron density distribution [17], several MD refinement ("MDXREF") protocols were tested [12]. Beginning with data to 3 Å and increasing the resolution gradually to 1.7 Å it appeared possible to decrease the crystallographic R-factor from an initial 48.5% to 38.6% by energy minimization including X-ray terms ("EMX") and, subsequently, to 28.8% after 1 ps of MD refinement. More important is perhaps the fact that in the m.i.r. model 7 peptides had a wrong orientation and that the refinement procedure flipped 4 of these to the correct position rapidly and an additional peptide somewhat later, after lowering the X-ray weights and increasing them again [12]. Many atoms moved by more than 1.5 Å, while shifts of such size occur infrequently in conventional restrained refinement. In this test, where no water molecules were considered, the large convergence radius of the method was clearly illustrated. This, together with the promising results of Brünger et al. [10], encouraged us to go on with more difficult cases.

4. Thermitase:eglin-c, crystal form I

Thermitase is a heat-stable member of the subtilisin family of serine proteases. It is isolated from Thermoactinomyces vulgaris and contains 279 amino acid residues [18,19]. Eglin-c is a serine protease inhibitor from the leech Hirudo medicinalis and consists of 70 amino acid residues [20]. Crystal form I of the thermitase:eglin-c complex was obtained in the absence of calcium ions in the crystallization medium. The crystals had space group $P2_12_12_1$ with a = 63.25 Å, b = 72.10 Å and c = 89.25 Å. Data were collected up to 2.2 Å resolution from one crystal on a FAST television area detector diffractometer [21,22]. A total of 63,206 measurements yielded 16,310 unique reflections with an R-merge (= $\Sigma|F-\langle F\rangle|/\Sigma\langle F\rangle$) of 5.0%.

The structure was solved by applying the molecular replacement method using the subtilisin Carlsberg:eglin-c complex [23] as a starting model. The sequence identity of subtilisin Carlsberg and thermitase is 47%. An initial model of thermitase was obtained automatically by changing the amino acid sequence of subtilisin Carlsberg into that of thermitase, using the MUTATE option (R.J. Read, personal communication) of the molecular modelling and analysis program WHATIF (written by G. Vriend). The resulting thermitase model lacked a seven residue N-terminal extension, a single residue C-terminal extension and 3 insertions comprising a total of 4 residues. Bond breaks were created at the three insertion sites. At three sites residues had to be deleted from the Carlsberg structure and the initial thermitase model contained three extremely long bonds to close the gaps. Ambiguities in the resultant $2mF_o-DF_c,\exp i\alpha_c$ [24] electron density distribution of

thermitase:eglin-c made model building a difficult task. Hence, the molecular dynamics procedure was used to carry out the refinement.

An extensive description of refinement procedures tested and results obtained is given by Gros et al. [25]. The decrease of the R-factor during the refinement, and the resolution ranges and σ_x values used, are depicted in Figure 2.

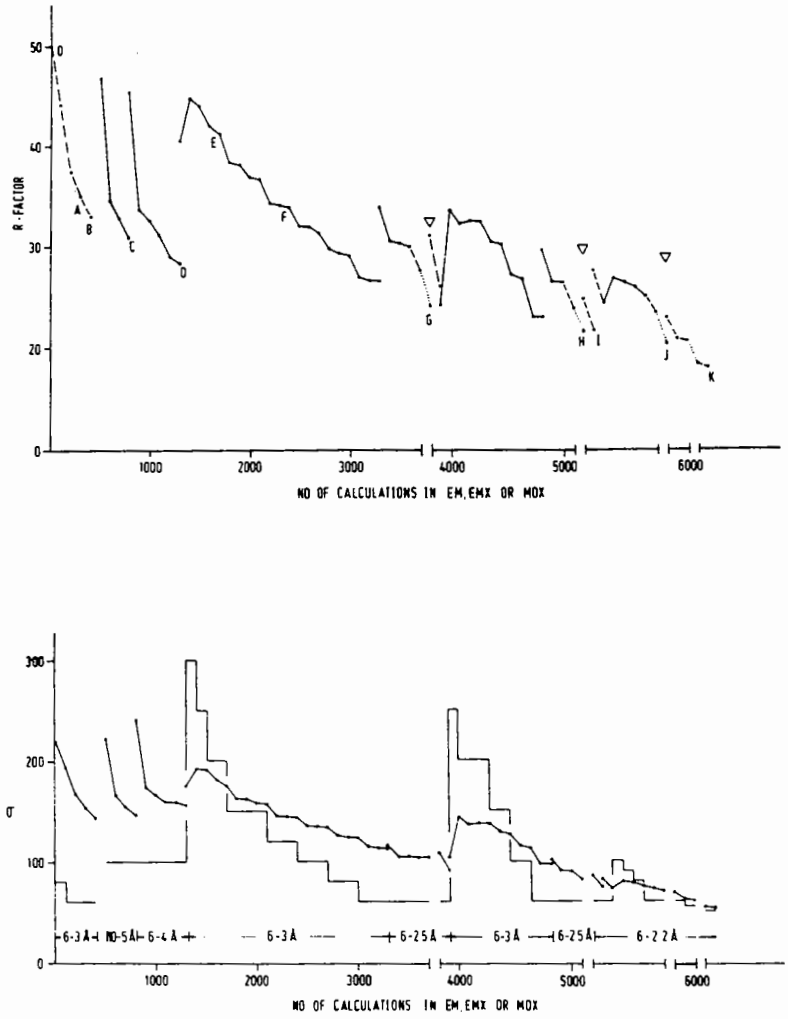


Figure 2 Course of the refinement of the complex of thermitase with eglin-c in crystal form I as indicated by the crystallographic R-factor.

a) The energy minimization with X-ray restraints (EMX) is indicated by - - -, the molecular dynamics refinement (MDXREF) by —— and the individual temperature factor refinement (BREF) by .... The triangles, ∇, indicate model building sessions using FRODO [34]. Between '400' en '500' "pure" energy minimization (no X-ray restraints) steps were carried out. MDXREF was run at 600 K; this was reduced to 300 K after adding water molecules to the model at points I and J. One step in MDXREF corresponds to a 2fs time step.

b) The root mean square difference between F_o and F_c is shown by ●——●. The weights $w_x = 1/\sigma_x^2$ were applied in a stepwise manner. σ_x is shown by thin lines.

All molecular dynamics refinement (MDXREF) steps were performed at either 300 or at 600 K using the standard GROMOS force field. Apart from a small manual correction at point A in Figure 2A which involved only residues 180 and 266, three model building sessions were carried out: at points G, H and J in Figure 2A. At point I and J solvent molecules were added which were included in the refinement without positional restraints.

Fully automatically, the molecular refinement procedure decreased the R-factor from an initial value of 50% at point O in Fig. 2A to a value of 24% at point G. An impression of the shifts occurring during these EMX and MDXREF refinement steps is given in Figure 3. Numerous atoms have moved by

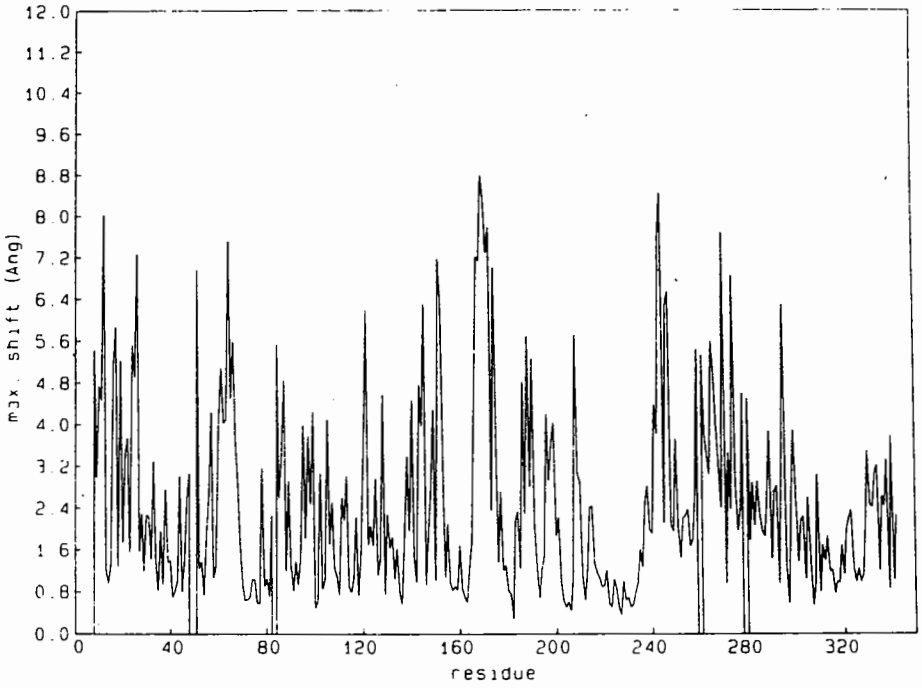


Figure 3 The maximal atomic shift per residue applied automatically to the thermitase:eglin-c model by the molecular dynamics procedure. The models at point O and point G (see Figure 2) are compared. Thermitase is shown for residue 1 to 279 and eglin-c from "280" to "342" corresponding with residues 8 to 70 of the inhibitor.

more than 4 Å, a distance which is rarely, if ever, seen in restrained refinement. The considerable improvement of the model during the EMX and MDXREF steps going from point O to point G in Fig. 2 is shown in Figure 4. Obviously, not all errors were removed by the procedure, but it should be kept in mind that the starting thermitase model contained several very large errors near insertion and deletion sites as described above.

At some points quite spectacular corrections were made automatically by the procedure used. An illustration is given in Figure 5. It is most fascinating to see how Tyr-274 first is assuming a quite distorted conformation while Gln-22 "moves out of the way". Subsequently, and within a very short time period, Tyr-274 finds its correct outward position.

The final model of crystal form I of the thermitase:eglin-c complex has a R-factor of 17.9% for 14.718 unique reflections at 2.2 Å resolution [25].

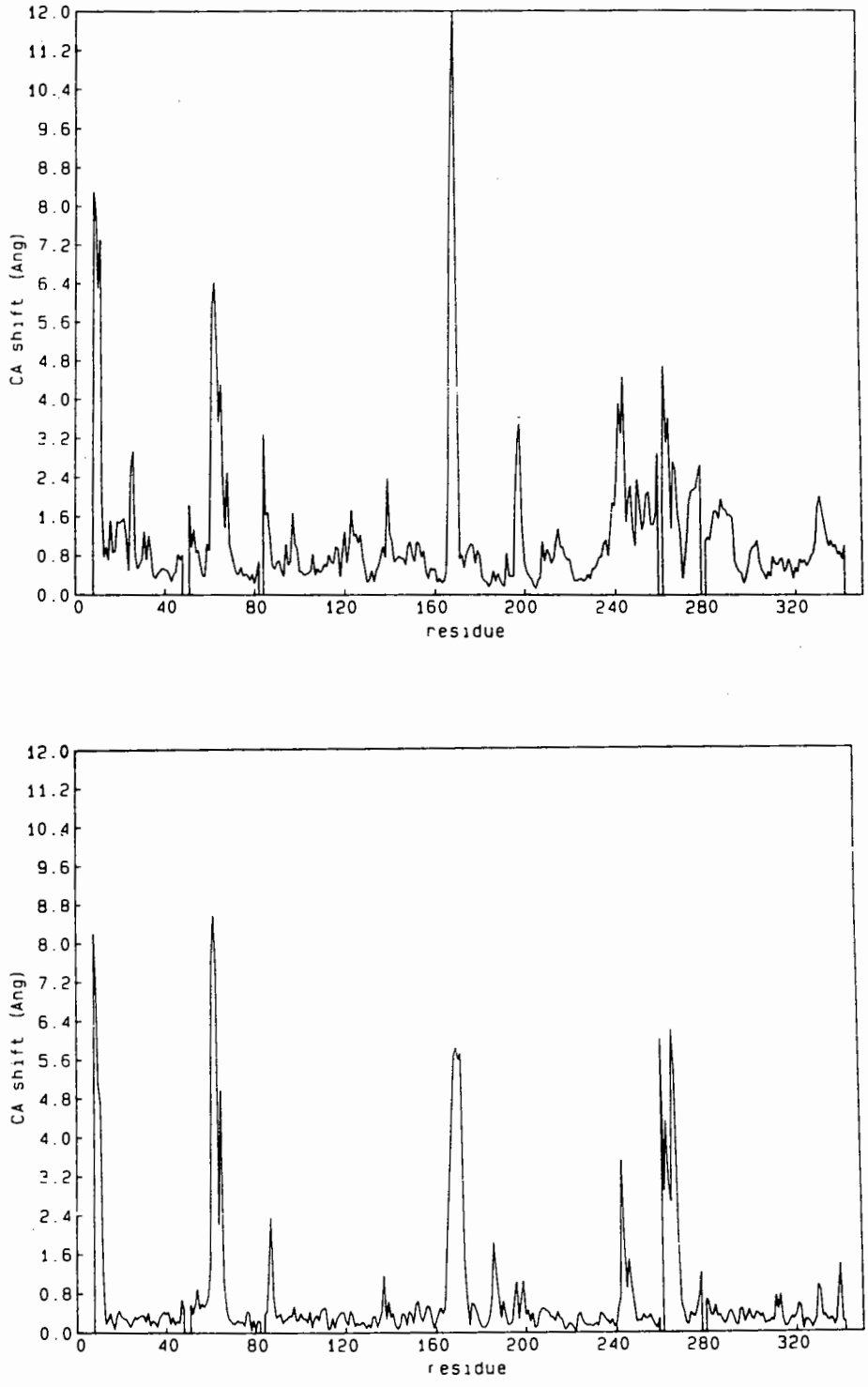


Figure 4 Shifts in C$^{\alpha}$ positions during the molecular dynamics refinement of thermitase:eglin-c in crystal form I.

a) The distance in C$^{\alpha}$ per residue between the final model, point K in figure 2, and the starting model at point O.

b) The distance per C$^{\alpha}$ between the final model, point K, and the model obtained by the molecular dynamics procedure before the first model building session, point G. Comparison with fig. 4a shows that numerous residues have been corrected by more than 1 Å.

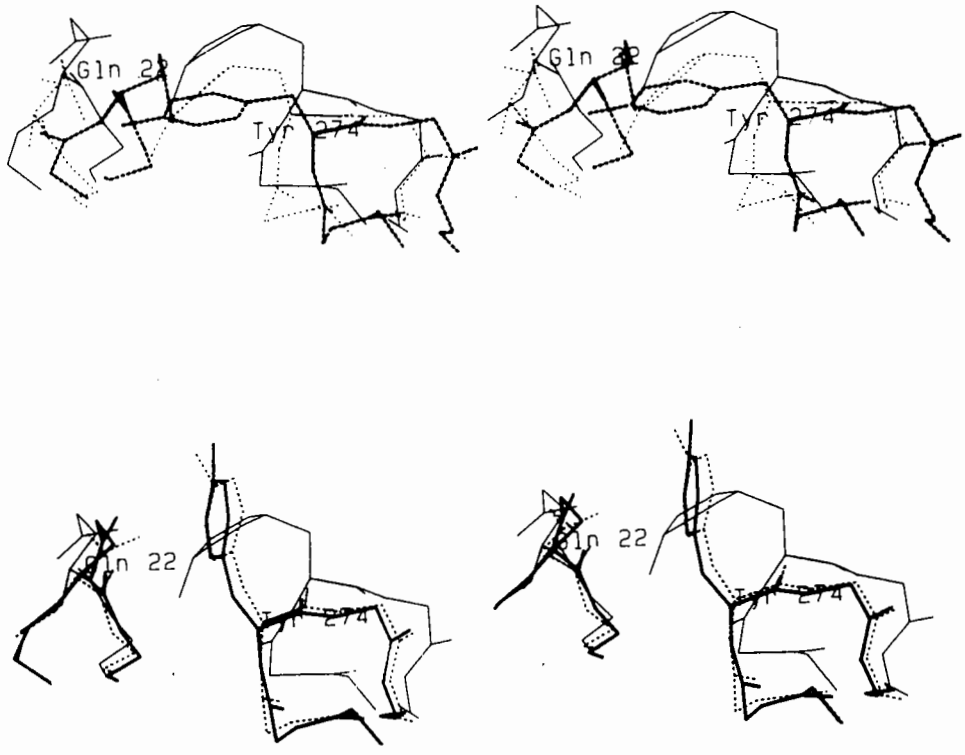


Figure 5 Subsequent stages in the refinement of Tyr-274 and Gln-22 in thermitase:eglin-c, crystal form I.
a) thick dashed lines (---) correspond to model O, thin dashed lines (---) to model C, and solid thin lines (——) to model E (see figure 2).
b) solid thin lines (——) show again model E, dashed lines (---) model F and thick solid lines the final model K.
The movements of Tyr-274 and Gln-22 shown were applied by MDXREF without any model building whatsoever.

5. Thermitase:eglin-c, crystal form II

A second crystal form of thermitase:eglin-c was obtained by Dauter et al. [26]. Data had been collected by these authors using synchrotron radiation and the oscillation film method. Two crystals yielded a total of 94298 measured intensities and 19.730 unique reflections from 8.0 to 1.98 Å used for the molecular dynamics refinement procedure. Dauter et al. [26] also solved the structure by molecular replacement using the subtilisin Carlsberg:eglin-c coordinates of McPhalen et al. [27]. Then the refined coordinates of thermitase:eglin-c in crystal form I, obtained as described in the previous section, were superimposed on the molecular replacement solution and the structure of crystal form II was subjected to molecular dynamics refinement procedures [28].

Obviously we had now a much better starting model than in the case of crystal form I and it appeared as if nothing unusual happened when the R-factor dropped smoothly from an initial 36.8% at 3 Å resolution to a final value of 22.2% at 1.98 Å resolution by the procedure depicted in Figure 6.

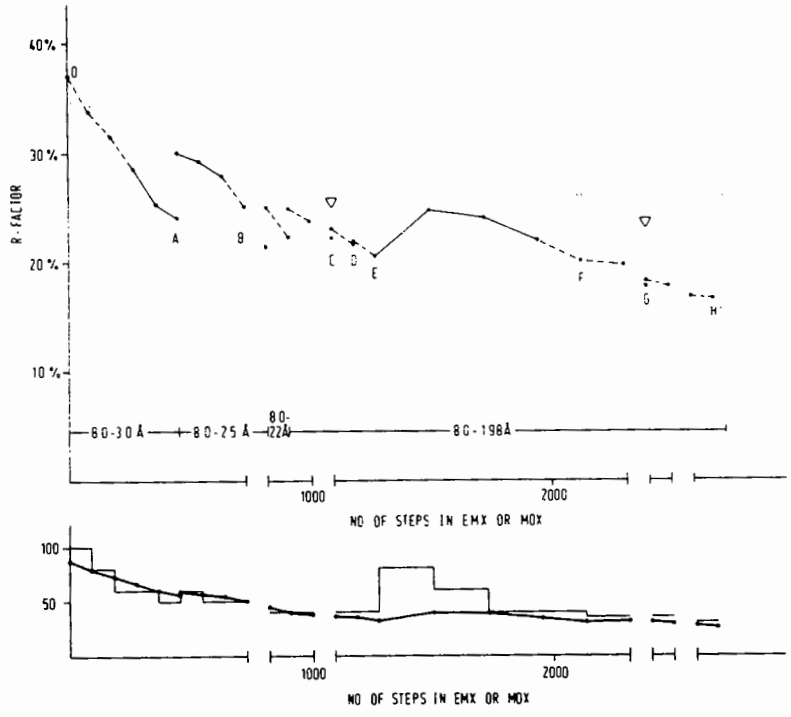


Figure 6 The course of the refinement of thermitase:eglin-c crystal form II as indicated by the crystallographic R-factor. The starting model was obtained by superposition of the model of thermitase:eglin-c crystal form I on a preliminary structure of form II, determined by Dauter et al. [26].

a) MDXREF is indicated by solid lines (——), EMX by dashed lines (– – –) and individual temperature factor refinement (BREF) by dotted lines (·····). The triangles, ▽, indicate model building sessions. Water molecules were introduced at points C ang G. Before C, MDXREF was run at 600 K and afterwards at 300 K. One step of MDXREF corresponds to a 2fs time step. The resolution ranges used for obtaining X-ray forces are also shown.

b) The root mean square difference between F_{obs} and F_{calc} is given in thick lines (•—•). The step function (thin lines) corresponds to σ_x used as a weight for the "X-ray energy" in the EMX and MDXREF runs.

However, when the final model was superimposed onto the initial model a quite remarkable result was obtained: the inhibitor eglin-c had been, fully automatically, rotated by ~ 10° in the course of the refinement. The shifts involved are shown in Figure 7 from which can also be seen that the inhibiting loop, comprising residues 41 to 47 of eglin-c, had not been rotated. So, a rigid body motion had been performed by the molecular dynamics refinement procedure of the "core" of eglin-c while automatically the hinge regions around residue 40 and 48 had been selected. In this way the core had been allowed to move while the inhibiting loop remained virtually unaltered.

The final thermitase:eglin-c model of crystal form II had a crystallographic R-factor of 16.5% including 214 solvent molecules and three cation binding sites [28].

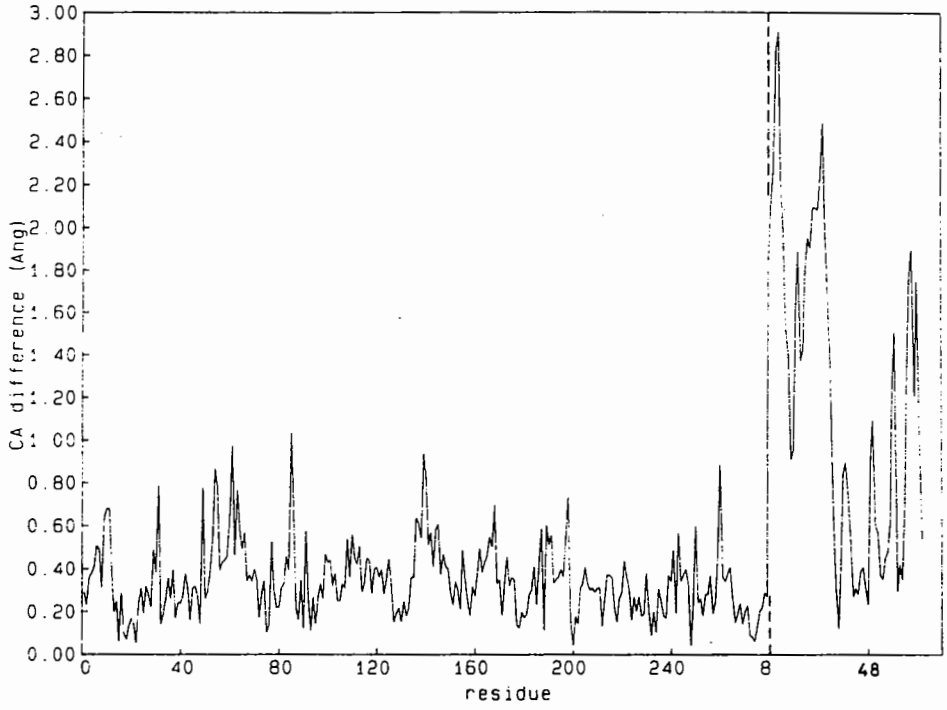


Figure 7 The distance in Cα position between the starting model, point O in Figure 6, and the final model, point H in Figure 6, of thermitase:eglin-c in crystal form II. Thermitase residues are numbered from 1 to 279. Eglin-c is numbered from 8 to 70. The first 7 residues of eglin-c are omitted from the graph, because they are absent in the electron density and hence in the model.

6. Lipoamide dehydrogenase from *Azotobacter vinelandii*

Lipoamide dehydrogenase (LipDH) is a flavoprotein which is a member of a number of multienzyme complexes, one of which is the pyruvate dehydrogenase complex. The reaction catalyzed by this enzyme is the reoxidation of lipoamide coupled to the reduction of one molecule of NAD. In the active form LipDH is a dimer with a molecular weight of 103,000 daltons.

Crystals of LipDH were grown in 22% PEG 4000 by the liquid-liquid diffusion method. Their space group was $P2_12_12_1$ and the cell dimensions were a = 61.1 Å, b = 83.8 Å, c = 192.0 Å. The crystals contained one dimer per asymmetric unit [29]. Diffraction data were collected at the DESY synchrotron in Hamburg to 2.2 Å resolution corresponding to 44,000 unique reflections. The structure of LipDH was solved by a combination of isomorphous and molecular replacement employing the "phased translation function" [30] using as starting model the related enzyme glutathione reductase [31].

The initial structure was obtained by the MUTATE option (written by R.J. Read) of the molecular analysis program WHATIF (written by G. Vriend). It gave an R-factor of 45% at 2.5 Å resolution for 30,000 reflections. The refinement by the molecular dynamics refinement GROMOS package [12] decreased the R-factor to 24.4% at 2.4 Å (35,000 reflections) without any manual intervention. During the refinement the resolution was gradually increased from 5.0 Å to 2.4 Å, performing 1100 steps (1 step = 2 fsec) of MDXREF at 600 K plus 200 steps of EMX. The course of the refinement of LipDH is summarized in Table 1.

Table 1 Course of the refinement of LipDH

STARTING MODEL: R_{fac} = 44% Resolution 8 - 2.4 Å
THE TWO CHAINS WERE IDENTICAL

REFINEMENT:

		R-factor	No steps	$\langle(F_o-F_c)^2\rangle$	WEIGHT
8 - 3.0 Å	EMX	35.2	100	140	70
8 - 5.0 Å	MDXREF	31.4	200	130	100
8 - 5.0 Å	MDXREF	25.4	200	55	100
8 - 3.0 Å	MDXREF	29.7	300	160	70
8 - 2.6 Å	MDXREF	27.2	400	110	50
8 - 2.4 Å	EMX	25.9	100	90	45
8 - 2.4 Å	BREF	24.4			

RESULTS: RMS SHIFT ALL ATOMS (START-END): 1.5 Å
RMS DIFFERENCE Cα ATOMS C1/C2 (END): 1.2 Å
R-FACTOR OF C2/C2 DIMER: 37.0%

The two chains of the dimer were always kept independent from each other and at the end of the molecular dynamics refinement the RMS difference in the positions of Cα atoms of the two chains was 1.2 Å. In most of the cases these differences are due to the packing of the molecules inside the crystal, which can alter the conformation of the loops on the surface of the protein. Their significance is confirmed by the R-factor calculated with a model consisting of a "dimer of chain 2" (see C2/C2 table 1).

Interestingly, the molecular dynamics refinement was able to correct automatically the conformation of Pro-450, which has changed the conformation of the peptide bond from "trans" to "cis". This movement took place at the beginning of the simulation when the resolution was still low (5.0 Å) and therefore the model was able to undergo large fluctuations.

7. Methylamine dehydrogenase from *Thiobacillus versutus*

Methylamine dehydrogenase is a quinoprotein which contains a PQQ-related quinone cofactor. The enzyme is a tetramer made up of two identical heavy (H) subunits and two identical light (L) subunits, the latter containing the protein bound cofactor [32]. Crystals of space group $P3_121$ (with a = b = 129.8 Å, c = 104.3 Å) were used for the structure determination. These contain half a molecule, i.e. one H and one L subunit with a total of ca. 60,000 daltons, in the asymmetric unit. Starting from a m.i.r.a.s. electron density distribution and after phase improvement and extension by solvent flattening, a model for the H subunit was built in the resulting 2.5 Å electron density map in the absence of sequence information. However, attempts to unambiguously trace the polypeptide chain for the L subunit were unsuccessful. This was a consequence of the large number of covalent cross links in that subunit [32]. Therefore, we attempted to improve the definition of the electron density, thus resolving the remaining ambiguities, with the method of partial model phasing [33]. Three protein models of increasing completeness were refined with molecular dynamics procedures using a protocol similar to that used with phospholipase A_2 [12], which is summarized in Table 2.

Table 2 The molecular dynamics refinement procedure used in the structure determination of quinoprotein methylamine dehydrogenase

Cycle #	Model refined	# atoms of model	Initial Rf	Resolution	# of E.M. steps	# of M.D. steps	Final Rf	Resolution	RMS shift (Cα)
1	H subunit	2703	44.4%	10.0 - 3.0 Å	150	700	36.0%	8.0 - 2.25 Å	0.95 Å
2	H subunit + L subunit in 4 fragments	3211	38.4%	10.0 - 2.8 Å	150	600	30.4%	8.0 - 2.25 Å	0.75 Å
3	H + L subunits	3385	29.6%	10.0 - 3.5 Å	250	1000	28.6%	6.0 - 2.25 Å	0.60 Å

In the course of the refinement of each model, the resolution limit was lowered from its starting value to 2.25 Å. Model phases were combined with solvent flattening phases to obtain three successive electron density maps, which were used for model building with the program FRODO [34]. As a result of the refinement, the R-factor dropped from 44.4% for a model of the H subunit only, to 28.6% for a "complete" model, each built with an "X-ray sequence". The electron density improved dramatically as a result of this procedure, allowing the polypeptide chain to be traced in the L subunit region. Also, the definition of side chain density had improved considerably, allowing the X-ray sequence derived from the map to be in excellent agreement with recently obtained partial sequence data (F. Huitema, J.A. Duine & J.J. Beintema, unpublished results). Examination of the map with the resulting model showed shifts of residues into electron density, while at the same time in those regions of the model where large errors had been made (such as introduction of "extra" residues) the density made obvious how the model should be modified.

8. Calculating structure factors with completely general thermal motions of atoms

Thermal motions of protein and solvent atoms in crystal structure determinations have so far virtually always been described by isotropic temperature factors. This assumes that the motion of each atom is harmonic and of equal magnitude in all directions. This is obviously wrong for nearly all atoms, but the limited number of observations even in high resolution structures up to, say, 1.5 Å prevents the application of approaches demanding more parameters because the ratio observations: parameters becomes rapidly very unfavourable. However, not only are many atoms moving anisotropically instead of isotropically, but near the surface of a protein molecule quite complicated situations can occur involving multiple conformations of residues or loops, partially occupied solvent molecule positions and weakly bound water having properties close to bulk water.

Molecular dynamics simulations of proteins and surrounding liquid provide, in principle, an elegant method to alleviate all of these shortcomings of crystallographic methods in describing the complex motions of side chains, "bound" solvent molecules, and bulk water. This is in particular true when the simulations are carried out in the crystalline state so that crystal contacts are also taken into account. The method aiming at this goal [14] was the first attempt to use molecular dynamics procedures in the crystallographic refinement process. It takes the result of a molecular dynamics simulation, i.e. a large number of configurations

describing the trajectories of all atoms, in the unit cell, and uses these configurations for generating one electron density map. This map includes then the motions of all protein and solvent atoms, without any restrictions as to their complexity. Fourier transformation of this electron density distribution yields structure factors that can be directly compared with the observed structure amplitudes. The value of the resulting R-factor will depend on the accuracy of the simulation - i.e. on details of the interaction potentials, treatment of long-range electrostatic interactions etc. - and also on the degree of static disorder occurring in the crystal. The latter point will be difficult to approach by molecular dynamics techniques because it means that different conformations with high energy barriers are occurring and these barriers cannot be passed by a conventional molecular dynamics run. Perhaps a strategy starting from two, or more, different conformations might be a way to overcome this problem.

The "general thermal motion calculations" performed [14] were carried out on BPTI with four protein molecules and 560 water molecules per unit cell. The molecular dynamics simulation was carried out for 12 ps after an equilibration period of 8 ps. It turned out that, on average, the C^{α} atoms of the four BPTI molecules, which were treated independently from each other, deviated between 1.13 and 0.94 Å from the X-ray structure. The averaged C^{α} positions of the four molecules deviated 0.82 Å from those of the X-ray results. In view of these data it is no surprise that the R-factor using Fc's obtained after Fourier inversion of all trajectories yielded a crystallographic R-factor of 52%.

Although the absolute position of the atoms in the molecular dynamics simulation have shifted by about 1 Å on average, a detailed analysis showed that the local structure in the PTI molecule has been conserved to a much greater extent. This suggests the use of "molecular dynamics fluctuations" around the X-ray positions in the structure factor calculation. The resulting R-factors, obtained by shifting the average MD positions back to the X-ray positions, are given as function of resolution in Figure 8.

In these calculations water molecules were omitted from the structure factor calculations and yielded, for data between 6.65 and 1.5 Å resolution, an R-factor of 29.0%. The X-ray coordinates, when omitting the water molecules and utilizing an overall temperature factor, yield an R-factor of 30.0%. Inclusion of the X-ray temperature factors for the protein atoms, but still omitting all water molecules, lowers the R-factor to 25.8%.

Clearly, in this first application of the "general thermal motion" approach, the individual X-ray temperature factors still give a better result than the thermal motions derived from the MD simulation. This is not very surprising as, after all, the X-ray temperature factors are the result of a best fit to the observed data, allowing some 450 temperature factors to vary, whereas the molecular dynamics fluctuations are obtained completely independent from the observed structure factors. It should also be noted that the procedure used for transferring the vibrational motions to the X-ray positions is very crude. A better procedure would be to restrain certain atoms to their X-ray positions during the simulation. Another possibility is to carry out MDXREF-type of calculations not for improving atomic coordinates, but for obtaining trajectories which simulate the complicated thermal motions in proteins.

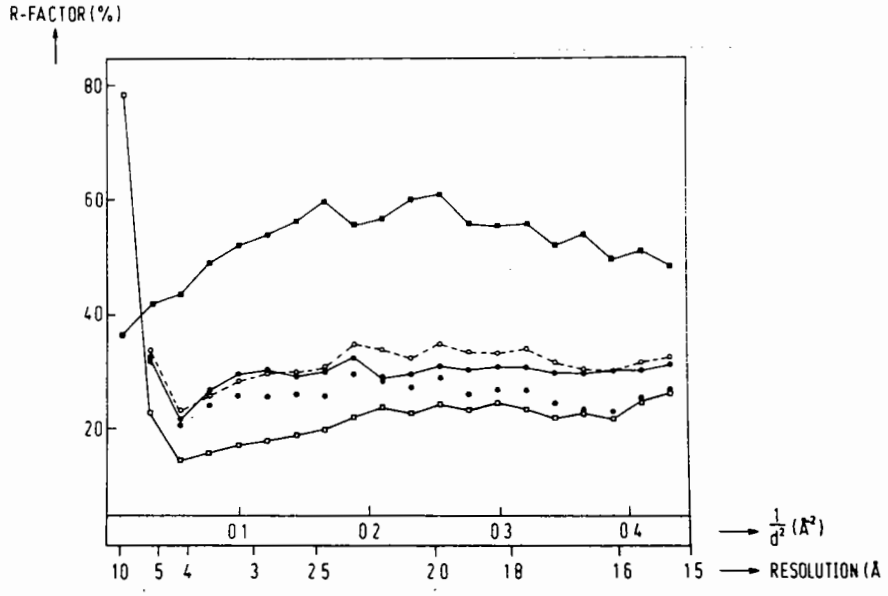


Figure 8 Reliability factors, R, as a function of resolution. The structure factors, F_c, were calculated by fast Fourier methods. Before calculating R factors [$R = (\Sigma|F_{obs} - F_{calc}| \times 100)/\Sigma|F_{obs}|$], first an overall temperature factor, B_{WILSON}, obtained from a relative Wilson plot of the data between 6.65 and 1.50 Å, was applied to the F_c values.
■—■, Structure factors obtained by summing the F_cs of 100 configurations of the 9th ps of the MD run. B_{WILSON} = 3.2 Å^2; R = 52.2% for 8,079 reflections between 50.0 and 1.50 Å.
□—□, Structure factors obtained from the X-ray coordinates, including individual temperature factors and 47 water molecules with relative occupancies. Twelve protein atoms, for which no temperature factors were known, were omitted. B_{WILSON} = 0.8 Å^2; R = 22.3% for 8,079 reflections between 50.0 and 1.50 Å.
o---o, Structure factors from X-ray coordinates without individual temperature factors or water molecules. All 454 protein atoms were included. B_{WILSON} = 11.8 Å^2; R = 30.0% for 7,963 reflections between 6.652 and 1.50 Å.
●...●, Structure factors from X-ray coordinates with individual temperature factors. For the 12 atoms for which no temperature factor was known, a value of 30 Å^2 was assumed. B_{WILSON} = 0.7 Å^2; R = 25.8% for 7,963 reflections between 6.652 and 1.50 Å.
●——●, Structure factors from 1,200 configurations of the 8- to 20-ps part of the MD simulation. B_{WILSON} = 0.5 Å^2; R = 29.0% for 7,963 reflections between 6.652 and 1.50 Å.

9. Conclusion

Incorporation of molecular dynamics procedures in protein crystal structure refinements appears to be capable of making very large corrections in molecular models. It also holds promise for obtaining a better description of thermal motion in protein molecules and the surrounding solvent.

10. References

1. Freer, S.T., Alders, R.A., Carter, C.W. and Kraut, J. J. Biol. Chem. 250 (1975) 46-54.
2. Watenpaugh, K.D., Sieker, L.C., Herriott, J.R. and Jensen, L.H. Acta Cryst. B29 (1973) 943-956.
3. Diamond, R. Acta Cryst. A27 (1971) 436-452.
4. Huber, R., Kukla, D., Bode, W., Schwager, P., Bartels, K., Deisenhofer, J. and Steigemann, W. J. Mol. Biol. 89 (1974) 73-101.
5. Agarwal, R.C. Acta Cryst. A34 (1978) 791-809.
6. Dodson, E.J., Isaacs, N.W. and Rollett, J.S. Acta Cryst. A32 (1976) 311-315.
7. Hendrickson, W.A. Methods in Enzymology 115 (1985) 252-270.
8. Jack, A. and Levitt, M. Acta Cryst. A34 (1978) 931-935.
9. Tronrud, D.E., Ten Eyck, L. and Matthews, B.W. Acta Cryst. A43 (1987) 489-501.
10. Brünger, A., Kuriyan, J. and Karplus, M. Science 235 (1987) 458-460.
11. Brünger, A. J. Mol. Biol. 203 (1988) 803-816.
12. Fujinaga, M., Gros, P. and Van Gunsteren, W.F. J. Appl. Cryst. 22 (1989) 1-8.
13. Van Gunsteren, W.F. and Berendsen, H.J.C. (1987) BIOMOS, Biomolecular Software, Laboratory of Physical Chemistry, University of Groningen, Groningen, The Netherlands.
14. Van Gunsteren, W.F., Berendsen, H.J.C. Hermans, J., Hol, W.G.J. and Postma, J.P.M. Proc. Natl. Acad. Sci. USA 80 (1983) 4315-4319.
15. Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., Rinola, A. and Haak, J.R., J. Chem. Phys. 81 (1984) 3684-3690.
16. Dijkstra, B.W., Kalk, K.H., Hol, W.G.J. and Drenth, J. J. Mol. Biol. 147 (1981) 97-123.
17. Dijkstra, B.W., Drenth, J., Kalk, K.H. and Vandermaelen, Ph.J., J. Mol. Biol. 124 (1978) 53-60.
18. Meloun, B., Baudys, M., Kostka, K., Jausdorf, G., Frömmel, C. and Höhne, W.E. FEBS Lett. 183 (1985) 195-199.
19. Frömmel, C. and Höhne, W.E. Biochim. Biophys. Acta 670 (1981) 25-31.
20. Seemüller, U., Fritz, H. and Eulitz, M. Methods in Enzym. 80 (1981) 804-816.
21. Arndt, U.W. Nucl. Instrum. Meth. 201 (1982) 21-25.
22. Renetseder, R., Dijkstra, B.W., Kalk, K.H., Verpoorte, J. and Drenth, J. Acta Cryst. B42 (1986) 602-605.
23. Bode, W., Papamokos, E. and Musil, D. Eur. J. Biochem. 166 (1987) 673-692.
24. Read, R.J. Acta Cryst. A42 (1986) 140-149.
25. Gros, P., Fujinaga, M., Dijkstra, B.W., Kalk, K.H. and Hol, W.G.J. Acta Cryst. (1989) in press.
26. Dauter, Z., Betzel, C., Höhne, W.E., Ingelman, M. and Wilson, K.S. FEBS Lett. 236 (1988) 171-178.
27. McPhalen, C.A., Schnebli, H.P. and James, M.N.G. FEBS Lett. 188 (1985) 55-58.
28. Gros, P., Betzel, Ch., Dauter, Z., Wilson, K.S. and Hol, W.G.J. J. Mol. Biol., submitted.
29. Schierbeek, A.J., Van der Laan, J.M., Groendijk, H., Wierenga, R.K. and Drenth, J. J. Mol. Biol. 165 (1983) 563-564.
30. Read, R.J. and Schierbeek, A.J. J. Appl. Cryst. 21 (1988) 490-495.
31. Schierbeek, A.J., Swarte, M.B.A., Dijkstra, B.W., Vriend, G., Read, R.J., Hol, W.G.J. and Drenth, J. J. Mol. Biol. 206 (1989) 365- .
32. Ishii, Y., Hase, T., Fukumori, Y., Matsubara, H. and Tobari, J. J. Biochem. (Tokyo) 93 (1983) 107-119.
33. Rice, D.W. Acta Cryst. A37 (1981) 491-500.
34. Jones, T.A. Methods Enzymol. 115 (1985) 157-171.

Crystallographic Refinement by Simulated Annealing

William I. Weis and Axel T. Brünger
The Howard Hughes Medical Institute and
Department of Molecular Biophysics and Biochemistry,
Yale University,
New Haven, CT 06511

Macromolecular crystallographic refinement aims to improve the agreement between experimentally observed structure factor amplitudes (F_o) and those calculated from an atomic model (F_c), while also improving the geometry or empirical energy of the model in accordance with prior knowledge of these quantities obtained from small molecule crystallography and spectroscopy. Mathematically, this corresponds to minimizing a target function $E_{total} = E_{empirical} + E_{X-ray}$: $E_{empirical}$ is the empirical energy of the model, and E_{X-ray} is the squared difference between F_o and F_c summed over all reflections. The atomic positions and temperature factors are the adjustable parameters of E_{total}. Unfortunately, least squares and other minimization methods are easily trapped in local minima, and manual refitting of the model to electron density maps is required to cross barriers between minima to allow the process to continue. With the computing power currently available to most crystallographers, manual model adjustment is the rate limiting step in the refinement process. Clearly, an automated method of overcoming barriers between minima would greatly speed refinement; specifically, we would like to search the configuration or phase space of the target function to find a minimum closer to the global minimum instead of simply falling into the local nearest minimum.

Simulated Annealing

The method of simulated annealing (SA), described by Kirkpatrick *et al.* (1983), allows exploration of the phase space of a complicated multi-parameter target function such as E_{total}. In contrast to conventional minimization, which allows only energetically 'downhill' steps, SA can cross barriers between minima by taking 'uphill' steps with probability e^{-E_{total}/k_bT}, using an effective temperature T as a control parameter (k_b is the Boltzmann constant). Typically, the system is first subjected to high temperatures so that high barriers separating large regions of phase space can be crossed; then the system is slowly cooled to restrict access to successively smaller regions of phase space, until a configuration closer to the global minimum than that at the start is obtained. The name of the method derives from the physical analogy of changing an amorphous glass (high energy, only local order) to a crystalline solid (low energy, long range order): the glass must first be liquified, then slowly cooled or 'annealed' into the crystalline state. Slow cooling is required to prevent the system from being trapped or 'frozen' in another metastable solid state with only local order.

Brünger, Kuriyan and Karplus (1987) and Brünger (1988a) have demonstrated the utility of simulated annealing in crystallographic refinement, and Brünger has developed a refinement package, X-PLOR, which incorporates this technique (Brünger, 1988b). Molecular dynamics simulations are employed to heat the system. An effective energy term, E_X, consisting of the squared structure factor amplitude residual, as well as non-bonded energy terms describing interactions between crystallographic symmetry-related molecules, is added to the empirical energy function describing covalent bonds, bond angles, torsion angles, and non-bonded interactions used in conventional molecular dynamics simulations. E_X restrains the dynamics trajectory to the experimental observations. It is important to note that the restrained molecular dynamics simulation does not correspond to any physical reality; it is simply a tool for carrying out a search of configuration space.

Annealing Schedules

Experience with simulated annealing in a variety of applications has shown that the 'annealing schedule', the sequence of temperatures and number of configurational rearrangements permitted at each temperature, is of crucial importance to the success of the method. At each temperature, the simulation must proceed long enough to reach steady state; if the system is cooled too quickly, it can get out of equilibrium and be trapped in a metastable energy state.

Table 1 compares several annealing schedules for the SA-refinement of the enzyme mitochondrial aspartate aminotransferase (Brünger, Krukowski and Erickson, in preparation). In all cases, the temperature is reduced in increments of 25K, with the restrained molecular dynamics carried out for a specified time. The desired temperature is maintained by coupling the dynamics to an external heat bath via frictional dampening, as described by Berendson *et al.* (1984); we have found this to be more stable than *ad hoc* velocity rescaling in several systems. A short time step of 0.5 fs is used to keep the integration of the equations of motion stable at high temperatures. Several issues concerning the choice of annealing schedule are illustrated by the overall statistics, as well as by inspection of electron density maps made from the models obtained from the different schedules. Equilibration at high temperatures prior to the onset of cooling is not necessary. Apparently, the system stays at high temperatures for a sufficiently long time to adequately sample phase space. However, spending more time at each temperature during cooling is advantageous. In particular, 6.25 fs of dynamics at each temperature does not appear to be sufficient to maintain steady state during cooling, and produces inferior results. Increasing the time spent at each temperature from 12.5 to 25 fs gives some improvement, but not as much as the difference between 6.25 and 12.5 fs per interval.

The temperature range used in the simulation is another important parameter. The data shown in Table 1 suggest that going to higher temperatures improves the result. The last three entries in the table show that simulations at 300 or 600K produce inferior results, even if one attempts to increase the radius of convergence by first starting at low resolution and then moving to higher resolution. While significant improvements are seen if the simulation is run from 4000K instead of 2000K, going to 6000K gives only marginally better results in this system. In some cases, going to higher temperatures may deteriorate the final model obtained (Brünger, Karplus and Petsko, 1989; Kuriyan *et al.*, 1989). The optimal highest temperature for annealing may therefore be a somewhat idiosyncratic choice for each SA-refinement problem. No advantage is seen in cooling all the way to 0K, as opposed to room temperature followed by minimization. The data also indicate that some improvement may be obtained by successive heating/cooling runs, even with no intervening manual model adjustment.

Non-crystallographic symmetry

The non-crystallographic symmetry (NCS) present in many systems can be utilized in SA-refinement. We have incorporated two different methods into X-PLOR to treat NCS, which we term 'strict' and 'restrained'. Strict NCS assumes that the monomers (or, more generally, protomers) are strictly identical, thereby allowing SA-refinement of the protomer only. The NCS operators (rotation matrices and translation vectors) that generate the crystallographic asymmetric unit from the protomer coordinates are applied to the protomer coordinates to generate F_{calc}'s and their derivatives with respect to the atomic parameters for the entire crystallographic asymmetric unit. Using the chain rule, the derivatives are transformed back to the non-crystallographic asymmetric unit and averaged. The strict NCS method thus improves the signal-to-noise ratio by averaging the noise in the data. For non-bonded interactions, the

NCS operators (and additionally, in the general case, a subset of the crystallographic symmetry operators needed to completely define the internal symmetry of the oligomer) are used to generate a list of NCS-related atoms within the non-bonded cutoff distance from the protomer atoms. The van der Waals and electrostatic interactions between the protomer and the atoms in this list are computed during the dynamics. Inter-subunit interactions are thus correctly treated while refining only a monomer. (Crystal packing interactions cannot be treated by this method, as atoms in lattice contacts are in non-equivalent chemical environments with respect to their NCS mates, while strict NCS assumes that such atoms are in equivalent environments.) Strict NCS reduces the computational time required to evaluate the empirical energy function by roughly a factor of n for n-fold NCS, and, more importantly, reduces the number of model atoms that must be inspected and potentially manually rebuilt by the same amount. The strict NCS method does not reduce the time needed for structure factor calculations, however.

Restrained NCS employs the formulation described by Hendrickson (1985). The entire crystallographic asymmetric unit is refined, with NCS related atoms restrained to their average positions $\overline{X}$ after least squares superposition of the protomers by adding an effective energy term $E_{NCS} = k_{NCS}(X - \overline{X})^2$ to the empirical energy function. Isotropic temperature factors can be similarly treated with the restraint term $(B - \overline{B})^2/{\sigma_{NCS}}^2$. Restrained NCS is useful when one wishes to impose NCS on only part of the structure, or to restrain some parts of the structure more tightly than others. Details for both the strict and the restrained NCS-method will be published elsewhere.

Some practical considerations in using X-PLOR

Parameters

X-PLOR uses the empirical energy parameters developed for the molecular dynamics program CHARMM (Brooks *et al.*, 1983). However, as discussed by Brünger *et al.* (1989), several modifications of the parameters are required to avoid unphysical changes due to close contacts in the initial structure, or due to the effects of high temperatures. The force constants of the improper torsion angles that maintain the handedness of chiral centers and the planarity of aromatic rings are increased to prevent distortions of these angles seen when the CHARMM parameters are used. Similarly, the force constant specifying the dihedral torsion (ω) angle for the peptide bonds of all amino acids except proline is increased to prevent formation of *cis* peptide bonds. The force constant for the proline ω angle is kept at a small value to allow *cis–trans* transitions.

It can be argued that the use of stiffer parameters has a detrimental effect on the optimization by restricting the search to a smaller region of phase space. However, we have found that use of the standard, rather than the modified, CHARMM parameters produces models with significantly poorer geometry. This occurs not only when high temperatures are used, but also in room temperature dynamics and in conventional minimization. Moreover, incorrect regions that are not improved using the modified parameters are also not corrected or improved if the more flexible parameters are used. These observations suggest that the tight force constants have the beneficial effect of by preventing access to undesirable regions of phase space, thereby raising the probability of obtaining a model with better geometry. Moreover, given the stochastic nature of the search, there is no guarantee that an incorrect region of the model will be corrected even if more flexible parameters are used; conversely, bad regions are often fixed using stiffer parameters, especially at high temperatures. As a practical matter, since most atoms in the initial model are typically 1-2 Å from their refined positions (i.e., the

model is far from a random set of atoms), one wants to search configuration space relatively close to that of the starting model, given that an exhaustive search of the space is not possible. We therefore believe that the best SA-refinement strategy is to have the procedure improve the model in most regions, and then to correct any remaining problem areas manually. While SA-refinement may or may not correct a very bad portion of the model, experience with the method has shown that the improvements which occur in most parts of the model lead to more interpretable difference Fourier maps.

Weighting of the X-ray terms

X-PLOR computes a recommended overall weight to apply to the X-ray structure factor residual to make the gradient of the structure factor residual comparable to that of the empirical energy (Brünger, Karplus and Petsko, 1989). These quantities are calculated from the model obtained from a short dynamics run without the structure factor residual, so that the empirical energy gradient is not artificially high due to bad contacts in the model. We have found that this is an adequate method for determining the relative contributions of the X-ray and empirical energy terms early in the SA-refinement process, as the success of the SA-refinement does not appear to be critically dependent on the choice of the relative weighting. However, just as in conventional refinement, the final structure is sensitive to the relative weighting of the X-ray and energy terms. We found that for the refinement discussed below, once the structure factor residual (i.e., the R-factor) had been reduced, the recommended weight tended to overweight the X-ray terms, and it was necessary to reduce it in the later stages to prevent degradation of the model geometry. We found this new weight by running short (e.g. 40 step) minimizations at a series of weights 10-40% smaller than the original, and selecting a value that did not degrade the empirical energy, nor made the R-factor significantly increase.

Treatment of charges

Experience with SA-refinement has shown that fully charged Asp, Glu, Arg and Lys residues behave abnormally during high temperature dynamics. In particular, when these residues are located at the surface of the molecule, where they are less constrained both by packing forces and by the X-ray term (since surface residues are generally more mobile), they tend to form salt links or hydrogen bonds with backbone and other side chain atoms that are clearly incorrect by inspection of difference Fourier maps. This problem is unique to SA-refinement, since residues can move much farther during dynamics than they can in conventional minimization. However, keeping the charges turned on helps to maintain chemically reasonable hydrogen bonds and salt links in the model, and to eliminate unreasonable ones. For example, if an asparagine or glutamine side chain is near a lysine, then the electrostatic term will favor formation of a hydrogen bond between the lysine and the carbonyl oxygen of the side chain rather than the amide, which would be otherwise indistinguishable at typical resolution limits. We find that by leaving the charges turned on during the minimization prior to the high temperature dynamics, then turning them off during the dynamics and final minimization, we can prevent formation of incorrect hydrogen bonds while maintaining favorable electrostatic energies. For refinement rounds involving only conventional minimization, we leave the full charges turned on to optimize the empirical energy.

A related problem occurs in histidines. In X-PLOR, polar hydrogen atoms are used for the electrostatic part of the empirical energy potential, and are placed by an automated procedure (Brünger and Karplus, 1988). By default, histidines are doubly protonated. We have found

that the double protonation can cause van der Waals and/or electrostatic repulsion of groups if one of the histidine nitrogen atoms is not actually protonated. In these instances, difference Fourier maps indicate that either the histidine or one of its bonding partners is incorrectly placed. The choice of which nitrogen must be deprotonated is usually obvious based on chemical reasonableness of the local hydrogen bonding pattern. A facility in X-PLOR can be used to remove the appropriate hydrogen. Fig. 1 gives an example of this problem.

Application to a large system: The influenza virus haemagglutinin

We have applied the methods discussed in the previous sections to the refinement of the influenza virus haemagglutinin (HA) (Wilson, Skehel and Wiley, 1981), a trimeric glycoprotein. The relevant parameters are listed in Table 2. In this section, we give an overview of the HA refinement to illustrate some strategies and problems unique to SA-refinement. A complete report of this refinement will be given elsewhere (Weis, Brünger, Skehel and Wiley, in preparation).

The HA crystallizes with a trimer in the asymmetric unit. Because of the limited resolution and the falloff of data quality past 3.2 Å , we chose to exploit the 3-fold NCS of this system. Our strategy was first to refine a monomer to acceptable R-factor and geometry using the strict NCS algorithm described above. This produced errors, as assessed by unaveraged difference Fourier maps and poor model geometry, in some of the lattice-contact regions of the molecule in which one or more of the monomers are in non-equivalent chemical environments. Thus, once the monomer SA-refinement had converged, another round of annealing was performed using restrained NCS; regions that were clearly not treated properly by the imposition of strict NCS were left unrestrained by NCS (25/503 residues per monomer). We left as few regions as possible unrestrained, since small deviations from NCS cannot be ascribed significance at 3 Å resolution. Moreover, since strict NCS was not imposed, deviations could occur in restrained regions. We note that distortions caused by the strict NCS algorithm may be somewhat specific to systems with a low degree of NCS: each monomer contributes heavily to the averaged X-ray 'force', and if one is deviant, it can drastically change this term. For systems with a higher degree of NCS at comparable resolution, it is not clear what advantage would be gained by dropping the strict NCS, given the large increase in computational time needed for the empircal energy portion of the SA-refinement.

Table 3 summarizes the progress of the SA-refinement of HA. A total of 5 rounds, each consisting of either SA or conventional minimization, followed by inspection and manual adjustment of the model, were carried out. The SA-protocol consisted of 3 parts: 1) An initial minimization of 80-120 conjugate gradient steps to relieve any bad contacts that might cause instabilities during the dynamics. We harmonically restrained C_α positions to their starting values in the first round to prevent any bad van der Waals contacts from drastically changing the structure. 2) Slow cooling from 4000K to 300K, temperature increments of 25K, with (25 steps) $\times$ (0.5 fs/step) = 12.5 fs of dynamics carried out at each temperature; the temperature bath coupling scheme was used to maintain the temperature. While Table 1 suggests that we could have gained from running longer at each temperature, the costs of the calculation did not seem to justify such gains, as we intended to do several rounds of annealing. 3) A final minimization of 80-120 steps. This protocol required 5.6 and 8.5 Cray X-MP central processor hours for the monomer and trimer, respectively; the trimer used 5.1 million words of memory. The conventional minimization used in the final two rounds consisted of alternating 20 steps of positional minimization with 20 steps of isotropic temperature factor refinement until convergence was achieved.

The ability of SA-refinement to move atoms far from their starting positions is illustrated with two examples, shown in Figures 2 and 3. In Fig. 2, a peptide bond has flipped over during the dynamics. An earlier refinement by conventional methods (Knossow *et al.*, 1986; Weis *et al.*, 1988) did not flip this bond, although the final difference map showed that the carbonyl oxygen was in negative $F_o - F_c$ density, while a corresponding postive peak was on the other side of the plane. The carbonyl oxygen forms a hydrogen bond with an arginine side chain. The flip occurred with the arginine charges turned off, confirming that this change was due to the X-ray data rather than favorable electrostatic interactions. Fig. 3 illustrates another unique aspect of SA-refinement. A poorly defined tryptophan side chain moved into strong density belonging to N-linked carbohydrate that was not included in the first round of annealing. This resulted in severe model geometry distortion at this tryptophan and four surrounding residues. The model was manually rebuilt in this region, and the missing carbohydrate added. In subsequent rounds of annealing, proper model geometry was maintained in this region. This example illustrates a caveat concerning SA-refinement: one must look very carefully at the model, both by difference Fourier maps and geometry as a function of residue number, to detect errors. The latter is particularly important, because the dynamics can move atoms far enough to at least partially compensate for missing model atoms, producing somewhat ambiguous difference Fourier maps. This is noteable for tightly bound solvent molecules, which are generally not put into the model until late in the refinement process.

As shown in Table 3, the first round of annealing produced a model with very good overall geometry. After manual adjustment of the model in regions of poor geometry, the monomer was subjected to another round of annealing. We tried cooling from both 2000K and 4000K, as we thought that because the model geometry had improved so much, it might not be worthwhile going to the higher temperature. The overall statistics (not shown) were slightly better for the higher temperature run, and $2F_o - F_c$ maps computed from the two structures were almost identical. However, a systematic comparison of the two structures revealed that in those regions where the models differed significantly, the model obtained from the 4000K annealing run fit the density better. In the 4000K structure, branched chain hydrophobic residues often were rotated by 180 ° or otherwise significantly different in side chain torsion (χ) angles from those in the 2000K and the input structures. Apparently, at 4000K the kinetic energy is sufficient to overcome some of the strong van der Waals repulsive barriers in the interior of the protein. This may in part explain the large improvement in mAATase obtained by cooling from 4000K rather than 2000K discussed earlier (Table 1).

After these runs, we noticed a systematic discrepancy between the root mean squared (rms) deviations from equilibrium of main chain *vs.* side chain angles: the overall rms deviation for bond angles was 3.6 °, but was 4.0 ° for main chain and 3.2 ° for side chain angles. We felt that, even accounting for the known flexiblility of τ angles, the discrepancy was suspicious, and was likely due to the very strong force constant required to maintain the chirality of the main chain C_α atoms (see above). We therefore increased the force constants on some of the main chain bond angles (Table 4). A similar discrepancy was noted for the rms deviations from planarity of proline peptide bonds *vs.* the other amino acids. This presumably arose from the weak force constant left on proline peptide torsion angles to allow *cis-trans* transitions during dynamics; hence, we raised the force constant for this angle to that used for the other amino acids. We then repeated the 4000-300K annealing, and found none of these systematic discrepancies. The overall statistics for this model are shown in the "round 2" entry of Table 3.

After the second round of annealing, the lattice contacts were rebuilt into unaveraged $2F_o - F_c$ maps, and a final 4000-300K annealing run was carried out on the trimer, as discussed above. A strong force constant of 300 Kcal/mole-Å^2 was used for the NCS restraint energy.

This gave rms deviations of about 0.03 Å between NCS superimposed monomers, well below the expected coordinate error at 3 Å resolution. (A trial run using a force constant of 500 Kcal/mole-Å^2 produced instabilities in the dynamics.) Comparision of the model with those obtained from the previous rounds indicated that most parts of the model returned to the same positions after annealing, except for a few disordered regions and some surface side chains. Such comparisions are useful in deciding when to stop annealing, i.e., stopping when well ordered parts of the structure are not changing significantly. Table 3 shows that after 3 rounds of annealing on the HA, the rms difference between the models of the last two rounds approach the expected coordinate error at 3 Å resolution, at least for the main chain atoms. Disordered regions or incorrect portions of the model often do not return to the same positions, and often have poor geometry, either in successive annealing runs, or in two runs using the same starting model but with the initial velocity assignments for dynamics taken from different random number seeds (Brünger, 1988a). We finished the SA-refinement of HA with two rounds of conventional refinement. Only minor adjustments in the disordered regions were required. Table 3 gives the final statistics. The model geometry is extremely good; to attain such geometry at 3 Å resolution with conventional refinement would take a great deal of effort in manual model building.

Conclusions

SA-Refinement can greatly speed the refinement process by eliminating much of the manual building effort required by conventional refinement. It is, however, not fully automatic; careful inspection and some manual model adjustment must be carried out. It is important to inspect the geometry as a function of residue number to find 'hot spots' in the structure that have been distorted, either because they are disordered or because atoms are missing from the model. Possible artifacts due to the electrostatic energy term must be considered. Finally, use of non-crystallographic symmetry for some large systems can yield substantial savings in the time required for SA-refinement.

References

Berendson *et al.* (1984) *J. Chem. Phys.* **81**, 3684–3690.

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

Brünger, A.T. (1988a). *J. Mol. Biol.* **203**, 803–816.

Brünger, A.T. (1988b). *X-PLOR Manual, Version 1.5, Yale University.*

Brünger, A.T. and Karplus, M. (1988). *Proteins* **4**, 148–156.

Brünger, A.T., Karplus, M. and Petsko, G.A. (1989). *Acta Cryst.* **A45**, 50–61.

Hendrickson, W.A. (1985) *Meth. Enzymol.*, **115**, 252–270.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983). *Science* **220**, 671–680.

Kuriyan, J., Brünger, A.T., Karplus, M. and Hendrickson, W.A. (1989). *Acta Cryst. A*, in the press.

Knossow, M., Lewis, M., Rees, D., Wilson, I.A., Skehel, J.J. and Wiley, D.C. (1986) *Acta Cryst. A*, **B42**, 627–632.

Weis, W., Brown, J.H., Cusack, S., Paulson, J.C., Skehel, J.J. and Wiley, D.C. (1988) *Nature*, **333**, 426–431.

Wilson, I.A., Skehel, J.J. and Wiley, D.C. (1981) *Nature*, **289**, 366–373.

Table 1: **Comparison of Slow-Cooling T-coupling Protocols for mAATase**

ID	*equilibration*	*cooling range* [K]	*cooling rate* [K] / [fs]	*R-factor*	$\Delta\Phi$ (deg)	Δ_{bonds} (Å)	Δ_{angles} (deg)
TS	5ps @ T=4000K	4000–300	25 / 12.5	21.9	56.1	0.018	4.0
2	1ps @ T=4000K	4000–300	25 / 12.5	21.6	56.8	0.017	4.1
4	none	4000–300	25 / 12.5	21.8	56.7	0.018	4.0
7	none	4000–300	25 / 6.25	23.0	57.2	0.018	4.3
5	none	4000–300	25 / 25	21.5	56.9	0.017	3.9
6	none	2000–300	25 / 25	22.6	56.7	0.019	4.2
8	none	6000–300	25 / 25	21.4	55.6	0.017	3.8
5a	none	4000–0	25 /25	21.5	56.6	0.017	3.9
55	none	4000–300	25 / 25				
		4000–300	25 / 25	21.1	56.0	0.017	3.9
555	none	4000–300	25 / 25				
		4000–300	25 / 25				
		4000–300	25 / 25	21.1	55.5	0.017	3.9
9	5ps @ T=300K	none	–	22.8	57.5	0.019	4.4
10	5ps @ T=300K (3.5Å)						
	5ps @ T=300K (2.8Å)	none	–	23.0	57.3	0.019	4.2
12	5ps @ T=600K	600–300	25 / 25	22.7	57.1	0.020	4.3

† structure before cooling obtained by minimization, 40 conjugate gradient steps with soft repulsive potential followed by 120 conjugate gradient steps with CHARMM nonbonded potential, 8.0 – 2.8 Å, C^{α}-restraints at 20 Kcal/(mole$Å^2$), W_A=130,000 Kcal/mole, W_P=12,000 Kcal/(mole rad^2), B=12.0 $Å^2$, Δ_F = 0.05Å, after cooling structure subjected to 120 conjugate gradient steps as above but without C^{α}-restraints and W_P set to zero.

† timestep=0.5 fs in all cases that start at 4000K or 2000K, timestep=1 fs in all cases that start at 600K or 300K, timestep=0.25 fs in the range between 6000K and 4000K.

Table 2: **Influenza Virus Haemagglutinin**

Protein Parameters

M.W. $\approx$ 225 kD

3 x 503 = 1509 amino acids

7 N-linked glycosylation sites/monomer; a total of 3 x 7 = 21 sugars are crystallographically observable (at 5 sites)

3 x 4052 = 12,156 non-hydrogen atoms

Crystallographic data

Crystallized from 1.32 M NaCitrate/0.15 M NaCl/pH 7.5

d_{lim} : 3.0 Å

Space group: $P4_1$; there is 1 trimer/asymmetric unit

Unit cell constants: a=162.6 Å, c=177.4 Å

Data collection: 1° oscillation photographs

Number of unique reflections, 12.0 to 3.0 Å: 75,476

(83% of theoretical)

R_{merge} = .115 (all data between 12.0 and 3.0 Å)

Data for refinement

Resolution range: 7.0 - 3.0 Å

Filtering:

1. All data between 7.0 and 3.2 Å included.
2. Data between 3.2 and 3.0 Å used only if $F_{hkl} \geq 2\sigma_{F_{hkl}}$

Number of reflections used: 67,242

Table 3: Course of HA refinement

ROUND:	start	1	2	3	4	5
unit refined:	-	monomer	monomer	trimer	trimer	trimer
method	-	SA	SA	SA	min. + B	min. + B
R factor	.390	.279	.270	.258	.233	.229
r.m.s. deviations from equilibrium :						
bond (Å)	.034	.018	.018	.017	.015	.015
angle (°)	4.3	3.7	3.3	3.0	2.9	2.9
dihedral (°)	32.	27.	27.	27.	27.	27.
improper (°)	3.4	1.4	1.4	1.4	1.5	1.5
chiral volume ($Å^3$)	.25	.14	.12	.11	.10	.10
peptide ω (°)	8.4	6.1	5.4	5.0	4.3	4.3
Non-crystallographic symmetry: r.m.s deviation from superposition (Å)						
monomer 2→1:	-	-	-	.031	.030	.030
monomer 3→1:	-	-	-	.031	.030	.031
B factors: r.m.s. deviations						
bond ($Å^2$)	-	-	-	-	1.4	1.4
angle ($Å^2$)	-	-	-	-	2.4	2.3
NCS ($Å^2$)	-	-	-	-	.96	.94
r.m.s. to previous round (Å):						
main chain	-	.73	.42	.35	.069	.049
side chain	-	1.3	.97	.69	.13	.086

SA = Simulated annealing with slow cooling: a. 80-120 steps minimization. b. Temperature bath coupled dynamics from 4000K to 300K, 25K step; for each temperature, 0.0125 ps (25 steps * 0.0005 ps/step) of dynamics was run coupled to a heat bath to maintain the temperature (reference: Berendson *et al.*, 1984). c. 80-120 steps minimization.

min. + B = Conventional minimization & B factor refinement: Alternating sets of 20 steps positional minimization, 20 steps isotropic temperature factor refinement.

Table 4: Changes to force constants

angle type	k_{old} $(Kcal/mol - rad^2)$	k_{new} $(Kcal/mol - rad^2)$
C_α—C—N	20.	60.
N—C_α—C	45.	70.
C—N—H	30.	50.
H—N—C_α	35.	70.

PRO peptide dihedral angle	k_{old} $(Kcal/mol - rad^2)$	k_{new} $(Kcal/mol - rad^2)$
	5.0	100.

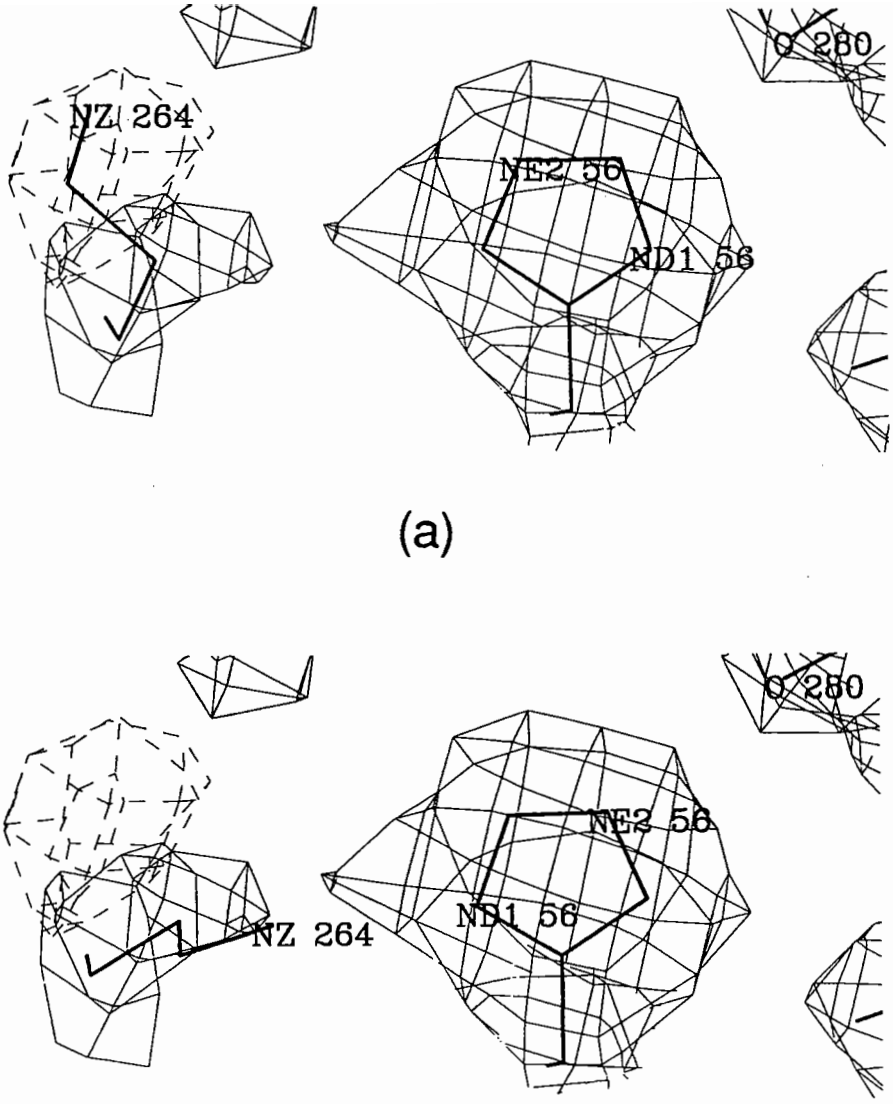


Figure 1

Double protonation of His 56 causes displacement of Lys 264. Solid lines are 1σ contour of $2F_o - F_c$ density, dashed lines are 3σ contour of $-(F_o - F_c)$ density. The maps have been averaged about the non-crystallographic 3-fold axis.

a. Result of minimization with His 56 doubly protonated. The positively charged Lys 264 side chain lies outside the $2F_o - F_c$ density, in negative $F_o - F_c$ density, to avoid unfavorable interactions with the positively charged histidine ring.

b. Result of minimization after deprotonating His 56 $N\delta 1$. The maps are the same as those in (a). The histidine ring was turned over 180 ° to optimize the hydrogen bonding between $N\epsilon 2$ and the main chain carbonyl oxygen of reside 280, and between the deprotonated $N\delta 1$ and $N\zeta$ of Lys 264.

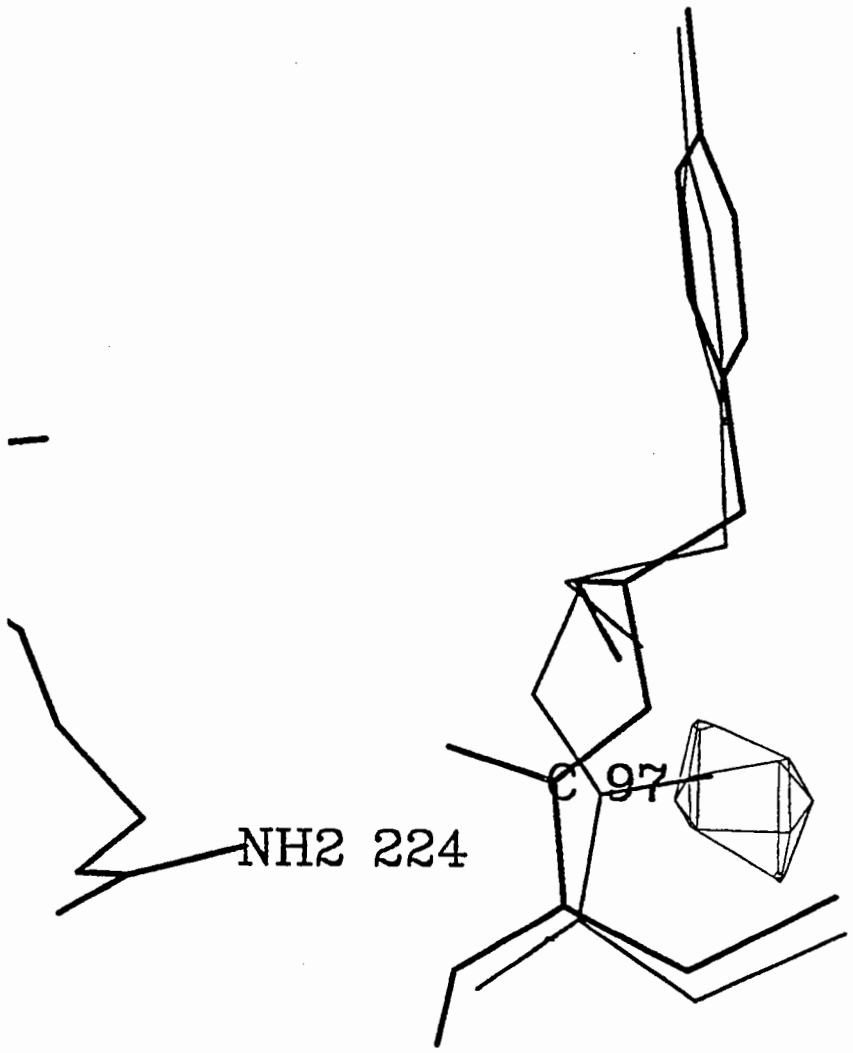


Figure 2

The peptide bond between residues 97 and 98 flips during the first annealing round of refinement. The density shown is the 3σ contour of the $-(F_o - F_c)$ density from an earlier haemagglutinin refinement performed by conventional least-squares methods (Weis *et al.*, 1988). The map has been averaged about the 3-fold non-crystallographic symmetry axis. The heavier bonds correspond to the SA-refined structure after the first round of annealing; the lighter bonds are from the least-squares refined structure. See text for further discussion.

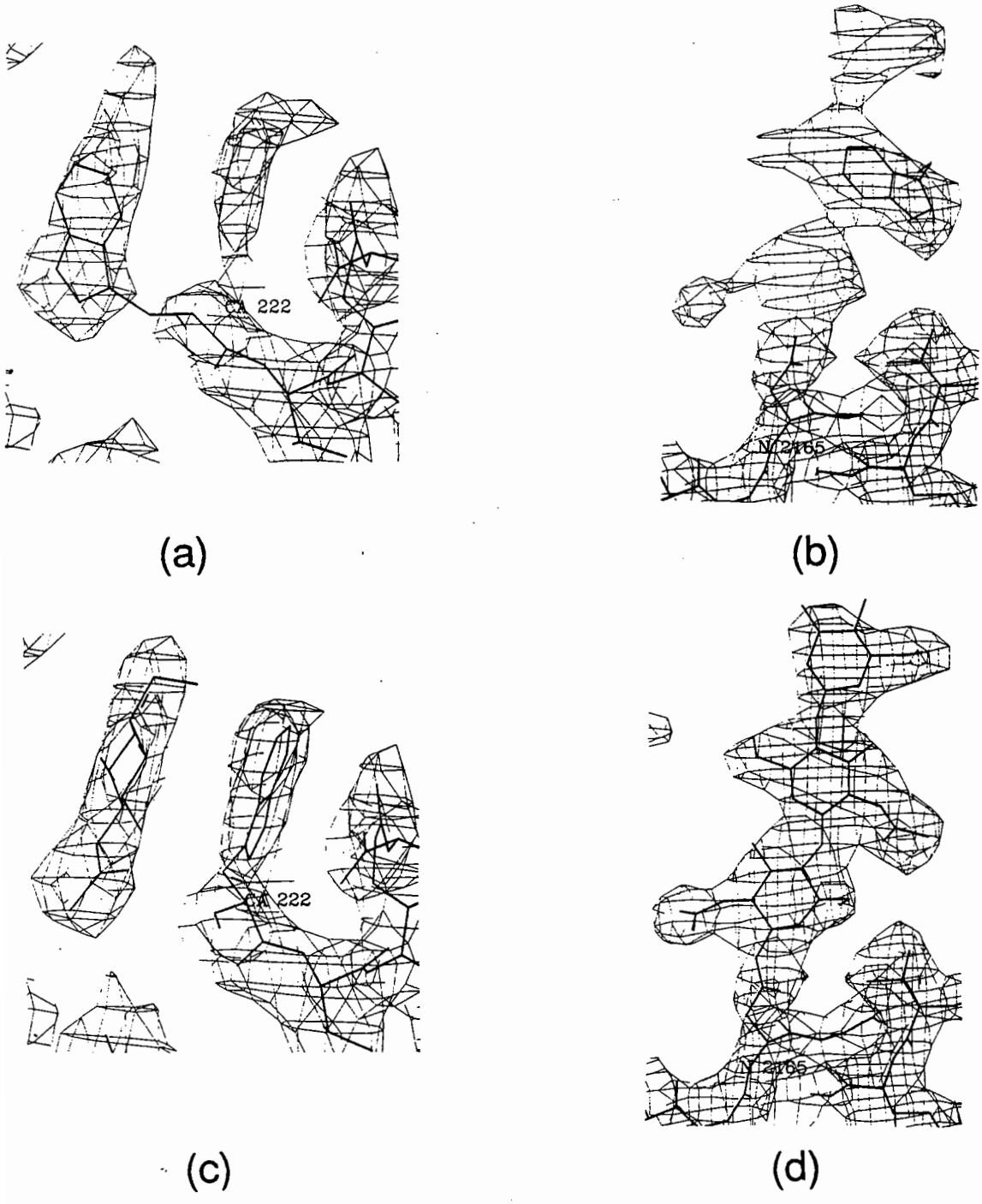


Figure 3

SA-refinement can move atoms far from their initial postions to compensate for missing model atoms. Trp 222, which has weak side chain density, moved into strong N-linked carbohydrate density in the first round of refinement, before carbohydrate was added to the model. The density shown is the 1σ contour of the $2F_o - F_c$ electron density. The maps have been averaged about the 3-fold non-crystallographic symmetry axis.

a. Electron density and coordinates after round 1, showing missing density for C_β 222. b. As (a), showing N-linked carbohydrate density protruding off Asn 165 of an adjacent monomer. c.,d. Electron density and coordinates after round 5. The views in (c) and (d) correspond to those in (a) and (b), respectively.

EXPERIENCES IN THE USE OF RESTRAINED DYNAMIC REFINEMENT

Garry Taylor

Laboratory of Molecular Biophysics, The Rex Richards Building,
South Parks Road, Oxford OX1 3QU. (Garry@UK.AC.OX.BIOP)

Many man months have been spent in the past carrying out 'manual' refinement: iterative cycles of restrained least squares followed by manual manipulation of the model at the interactive graphics. Like many other laboratories, this was carried out in Oxford using PROLSQ (Hendrickson & Konnert 1980) and FRODO. In the last four years, most refinement was carried out on a version of PROLSQ adapted to an FPS 38-bit array processor (FPS 5105) by the author, based on an earlier version for the FPS AP120B (Furey et al. 1979). This proved very successful, the machine being seldom idle, and in particular lead to the refinement of the 1.9Å structure of phosphorylase b among other proteins. The time to complete cycles ranged from minutes to several hours, but this was acceptable with a small number of proteins and the knowledge of many hours to be spent at FRODO. In the last year, however, the laboratory has solved 8 protein structures, some much larger and more complex than previously tackled. In addition, the ability to collect high quality x-ray data in house on the Xentronix detector, and in a short time, has increased the pressure on the later stages of structure solution and refinement. It was therefore very timely that the method of restrained dynamic refinement became available, with its promise of quickening the refinement process and reducing the need for excessive manual intervention.

Several new crystal structures in the laboratory have exploited the simulated annealing method of x-ray refinement in the version of Axel Brunger's excellent XPLOR package (Brunger 1988, 1989). The method appears to have been successful in all cases, leading to a more rapid convergence than the 'manual' method. Initially, we implemented XPLOR on the CRAY-X/MP48 at the Rutherford Appleton Laboratory, but more recently we have been running on our own Convex C210 processor installed as part of the SERC/MRC Oxford Centre for Molecular Sciences. The C210 runs at around 1/4 to 1/3 the speed of an X/MP processor for a typical XPLOR run.

I shall first present a test case, α-lactalbumin, whose structure is known, where the ability of the method to refine a molecular replacement solution is explored. Two new proteins whose structures were derived by molecular replacement are presented: an antigen binding fragment of an IgG antibody, and the R-state of phosphorylase b. Finally, the initial refinements of tumour necrosis factor (TNF), containing six copies of the molecule in the asymmetric unit, and of the foot and mouth disease virus (FMDV), with over 30,000 atoms in the asymmetric unit, are then discussed. The cpu times of the various stages are given for the Convex C210. As a crude comparison with other processors, if the C210 is taken as 1.0, then a microVax 3000 is 20 (times slower), a Vax 87/8800 is 10, a Cray X/MP is 0.3.

1. α-Lactalbumin, a test case

α-Lactalbumin is a calcium binding protein comprising some 15% of total protein in human milk. The structure of baboon α-lactalbumin has been solved previously to 1.7Å resolution by members of the laboratory (Stuart et al. 1986). The method of structure solution was the careful refinement of a model based on lysozyme which had been oriented in the cell through the use of low resolution isomorphous replacement phases. The careful refinement involved many cycles of CORELS (Sussman et al. 1977) and PROLSQ/FRODO. α-Lactalbumin has only a 37% sequence homology with hen egg white lysozyme (HEWL), but it had long been suggested that the two shared

structural homology (Browne et al. 1969), which was confirmed in the x-ray study. The current 'definitive' structure contains 990 non hydrogen protein atoms, 132 waters and a calcium ion with a crystallographic R-factor of 20.7% (or 31.5% for the protein atoms alone).

The strategy used to 'solve' the structure using simulated annealing was:
1. A molecular replacement solution was found using the MERLOT package (Fitzgerald 1988) with HEWL as the search model. The solution was the top solution in both the rotation function and translation functions (Ravi Acharya, unpublished results).
2. HEWL was positioned in the α-lactalbumin cell, and the residues were 'mutated' using FRODO blindly. The loops with deletions were 'annealed' by running geometric refinement (REFI) in FRODO. Only 2 sidechains were manually moved because they were clashing badly. This became the 'MERLOT' model, and the starting model for XPLOR. No rigid body refinement was carried out.

The refinement protocol used in XPLOR was as suggested in the tutorials accompanying the program. I refer to 'Jack-Levitt' refinement as performing energy minimisation with the x-ray term switched on. The heating and cooling stages represent dynamic simulations with the x-ray term switched on. Several protocols were carried out: at 2000K, 3000K with two cooling strategies, and 6000K. Details of the protocols are given below:

α-Lactalbumin 2000K MD refinement protocol
(990 non-H atoms, 10275 F's in range 8.0 - 1.7Å)

	Initial R=54.4%	(C2 cpu hours)
. Prepare stage (160 steps of 'Jack-Levitt')	R=43.7	(0.61)
. Heat stage (1000 steps of .001ps at 2000K)	R=40.7	(1.48)
. Cool stage (250 steps of .001ps at 300K)	R=37.2	(0.25)
. Final stage (40 steps of 'Jack-Levitt')	R=36.2	(0.22)
. B factor refinement	R=33.9	(0.13)
		(2.69)

Stereochemistry: rms bond deviation 0.026Å
rms angle .. 4.9°

α-Lactalbumin 3000K MD refinement protocol

	Initial R=54.4%	(C2 cpu hours)
. Heat stage (1500 steps of .001ps at 3000K)	R=40.0	(2.65)
. Cool stage (from 3000K to 300K in 25K steps)	R=34.0	(4.90)
. Final stage (40 steps of 'Jack-Levitt')	R=33.0	(0.24)
. B factor refinement	R=30.6	(0.13)
		(7.92)

Stereochemistry: rms bond deviation 0.025Å
rms angle .. 4.5°

. Fast cool stage (250 steps of .001ps at 300K)	R=34.9	(0.25)
. Final stage (as above)	R=34.0	(0.22)
. B factor refinement	R=31.5	(0.13)
		(3.25)
. Just 'Jack Levitt' refinement, 200 steps	R=43.4	(0.79)

The results show (i) that the MERLOT model does converge towards the 'definitive' structure, (ii) that 3000K is better than 2000K (figs. 1,2,3 & 7), (iii) that the slow cooling protocol produces marginally better results (fig. 4), and (iv) the B factor refinement at the end of simulated annealing correlates well with the B factors from PROLSQ (fig. 10). A run at 6000K, with a shortened time

Merlot model and definitive
Picture 1 created at 10:45:22 on 26-APR-89

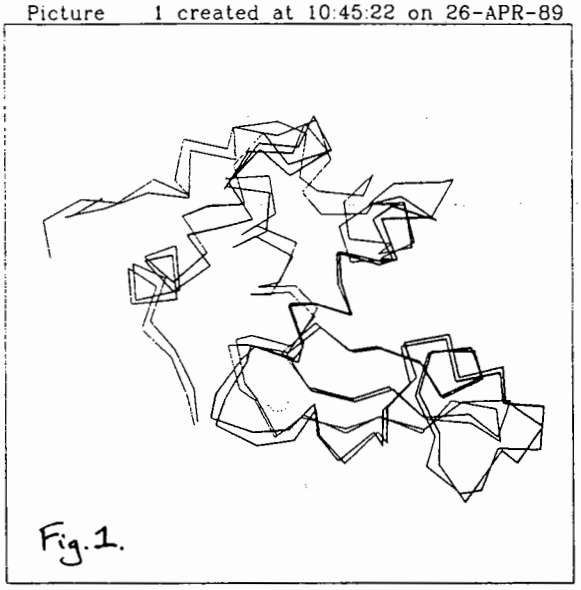


Fig.1.

2000K MD model and definitive
Picture 1 created at 10:46:24 on 26-APR-89

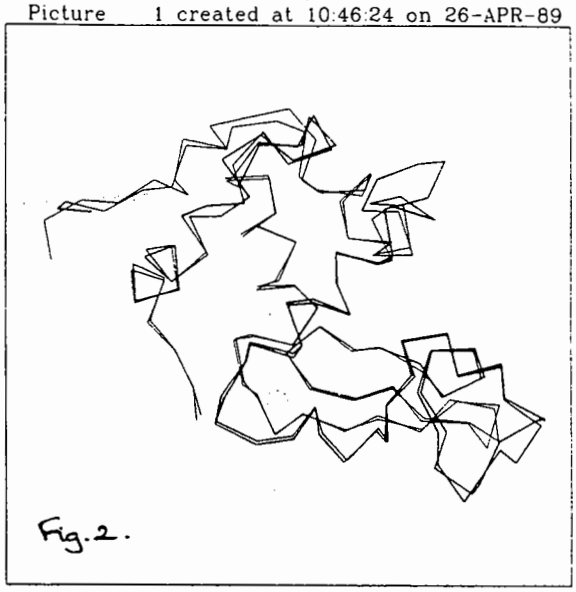


Fig.2.

3000K MD model and definitive
Picture 1 created at 10:47:10 on 26-APR-89

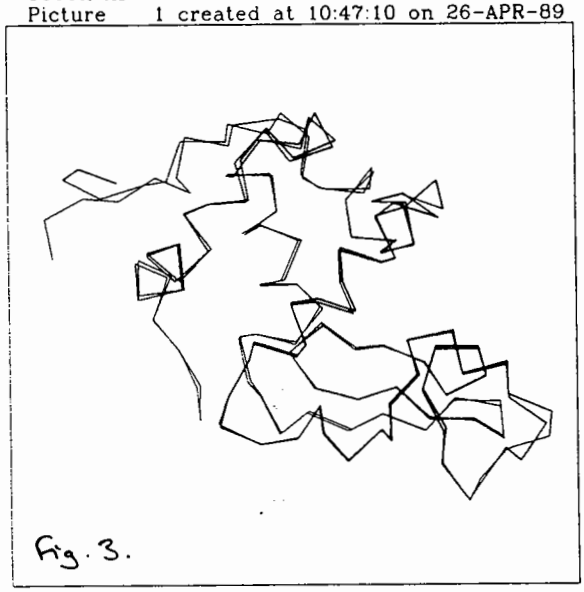


Fig.3.

2Fo-Fc map around calcium, using 3000K coords
Picture 1 created at 10:28:05 on 26-APR-89

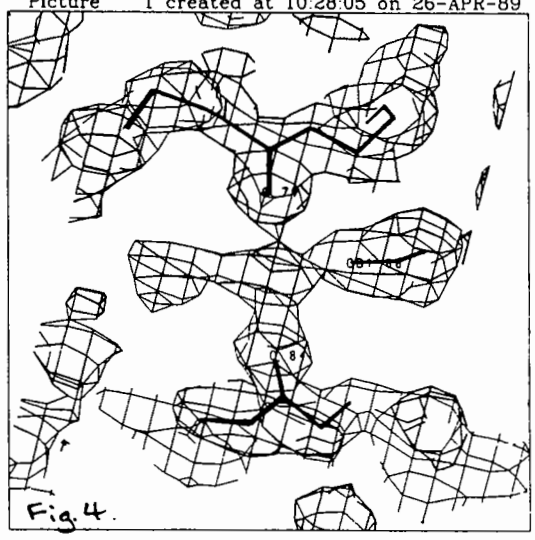


Fig.4.

3000K MD model and definitive around Pro 67
Picture 1 created at 11:07:00 on 26-APR-89

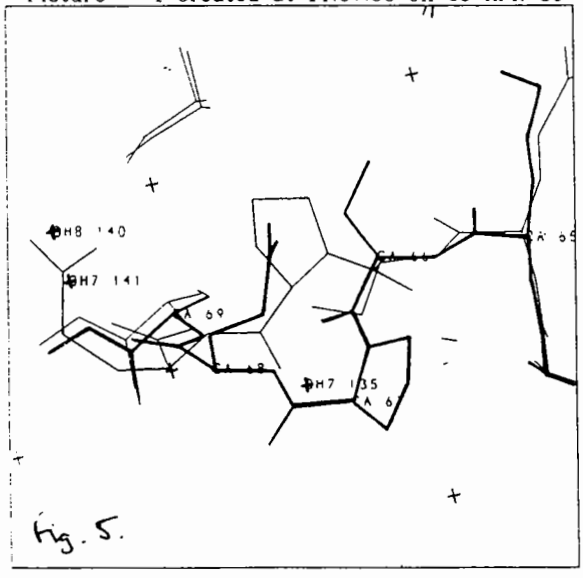


Fig.5.

2Fo-Fc map around Pro 67 using 3000K MD coords
Picture 1 created at 11:09:46 on 26-APR-89

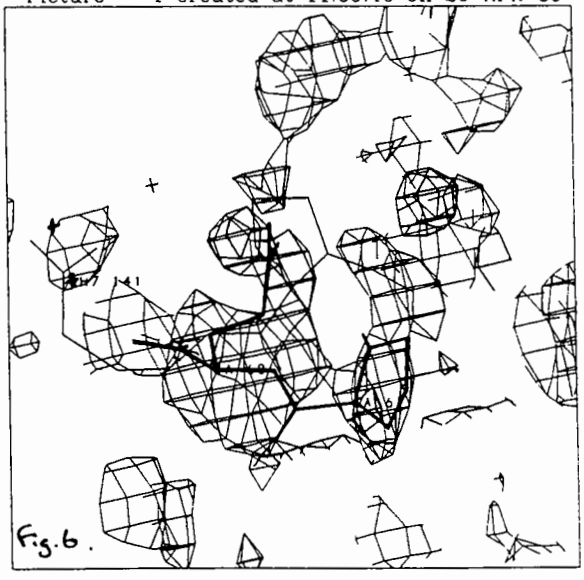


Fig.6.

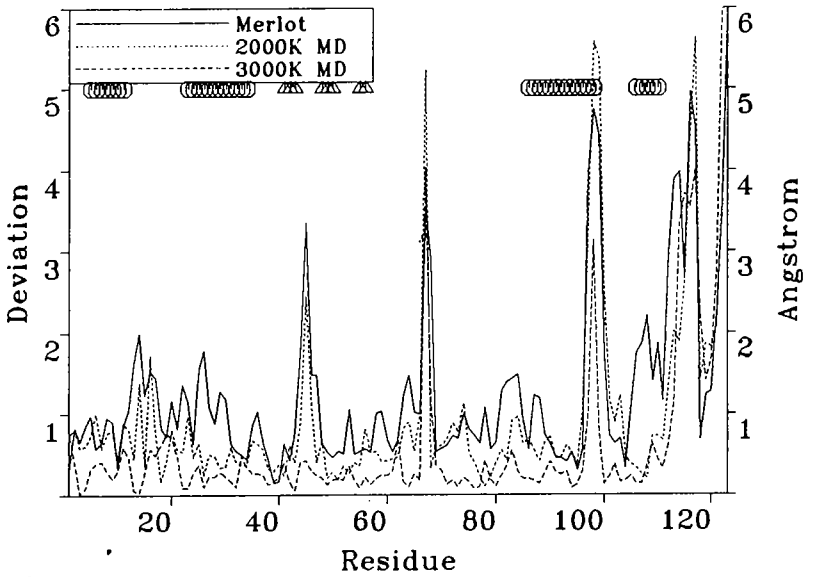


Fig.7. 1. Ca deviations from 'definitive'

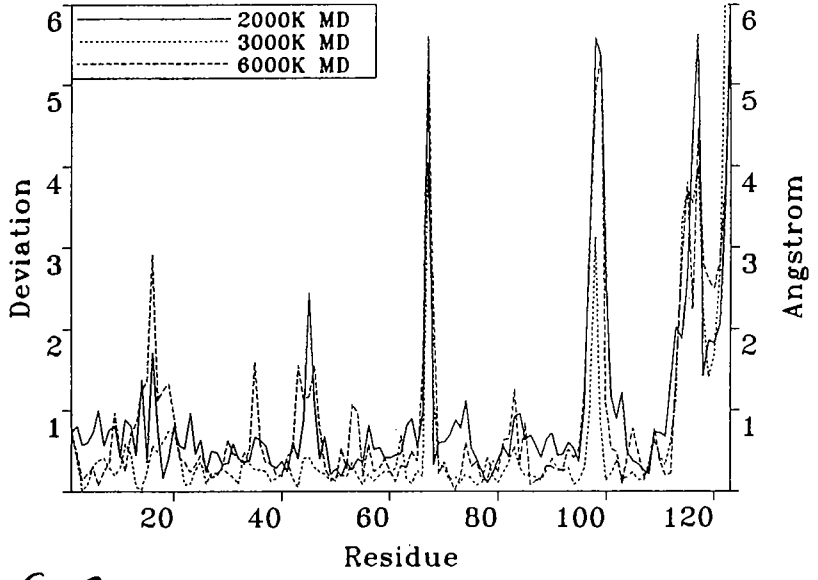


Fig.8. 2. Ca deviations from 'definitive'

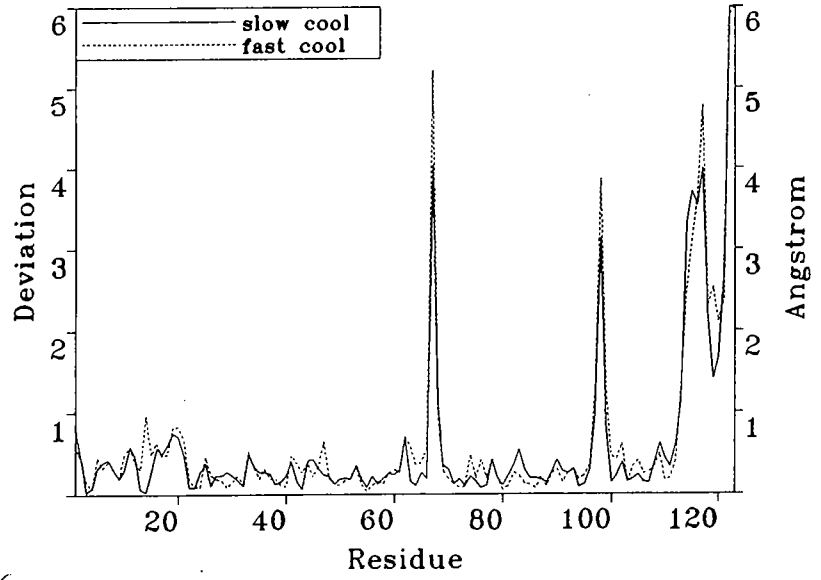


Fig.9. 3. Comparison of fast & slow cooling schemes

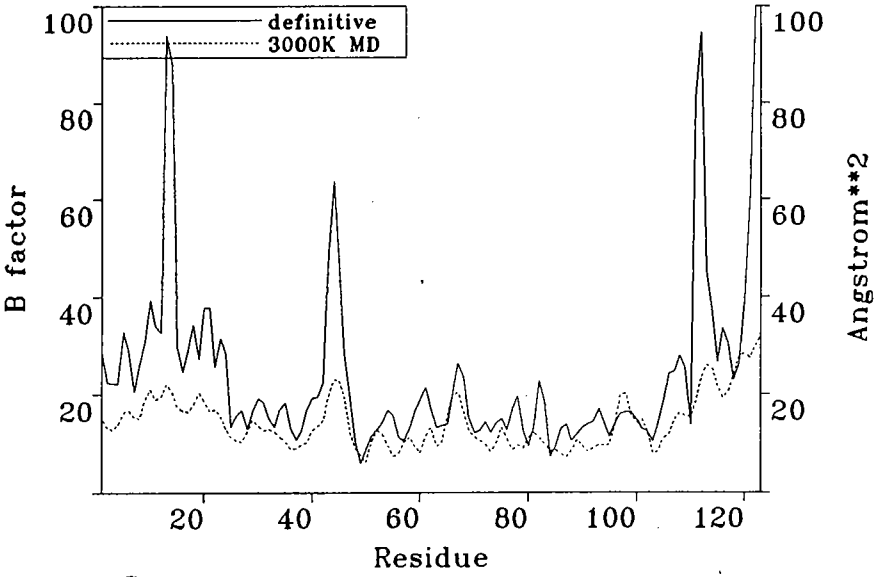


Fig 10. 4. B factor comparison

step of .5 fs during the dynamic simulation, produced worse results than at 3000K (fig.8) in contrast to the experience with crambin (Brunger et al. 1989) where best results were obtained at much higher temperatures. A $2|Fo|-|Fc|$ electron density map around the site of the calcium ion (fig. 4) which was not included in the refinement, reveals beautiful density for the calcium and both liganded waters observed in the 'definitive' structure. One region where the procedure has problems is around proline 67 which forms part of a surface loop. Figure 5 reveals that the proline has adopted a very different position to that in the 'definitive' structure, occupying the density of Gln 68, however the sidechain of Gln 68 in the 3000K structure has moved to occupy density assigned to water molecules in the 'definitive' structure. A $2|Fo|-|Fc|$ based on the 3000K coordinates (fig. 6) shows poor density for Pro 67, but strong density in the region of Pro 67's position in the 'definitive' structure; here manual intervention is required. Interestingly, a detailed examination of the potential energy of the 3000K MD structure around this region reveals the highest bond strain in the whole structure to be in the Cα-Cβ and Cβ-Cγ bonds of Gln 68. This could well be used as a way of picking up trouble areas. It should be added that this loop presented difficulties in the determination of the 'definitive' structure. Below are some statistics relating each structure to the derivative. The second column omits the C-terminal region which is particularly mobile.

rms deviations of α-lactalbumin structures from 'definitive'

Structure	Residues 1-123			Residues 1-111		
	Cα	BBone	All	Cα	BBone	All
Merlot	1.83	1.97	2.72	1.27	1.45	2.11
2000K MD	1.68	1.77	2.50	1.13	1.22	1.90
3000K MD	1.44	1.55	2.42	0.56	0.71	1.53

2. Monoclonal Antibody Antigen Binding Fragments (Gloop2)

(Phil Jeffrey, Bob Griest, Garry Taylor)

As part of a study of the nature of antibody/antigen recognition, 5 monoclonal antibodies raised against the 'loop' antigen of hen egg white lysozyme (named Gloop1 to Gloop5), are being investigated using x-ray crystallography, nmr and molecular modelling. The antigen binding fragments (Fab) of four of the monoclonals have been crystallised and the structures of two crystal forms of one of these, Gloop2, have been determined. The crystal forms are P1, where there are two Fab molecules in the unit cell, and $P2_1$ which has just one Fab. Both forms were solved using MERLOT and BRUTE (Fujinaga & Read), placing individual C_HC_L and Fv domains in the cells. In the P1 crystal form, each search model represented only 1/4 of the scattering power, nevertheless clean solutions were obtained. The orientations and positions of the models were refined using rigid body refinement in CORELS. These were then used as the starting point for XPLOR. Below are given details of the refinement protocols:

Gloop2 $P2_1$ refinement protocol

(One Fab in asymmetric unit, 3241 non-H atoms, 5423F's in range 8.0-3.3Å)

	Initial R=45.4%	(C2 cpu hours)
. Prepare stage (200 steps of J-L)	R=26.9	(0.37)
. Heat stage (1000 steps of .001ps at 2000K)	R=27.8	(0.94)
. Cool stage (220 steps at 300K)	R=23.6	(0.24)
. Final stage (200 steps of 'Jack-Levitt')	R=22.5	(0.36)
		(1.91)

Stereochemistry: rms bond deviation 0.023Å
rms angle .. 5.0°

Gloop2 P1 refinement protocol
(Two Fabs in a.u., 6174 non-H atoms, 19079 F's in range 8.0 - 2.8Å)

	Initial R=48.1%	(C2 cpu hours)
. Prepare stage (200 steps of J-L)	R=28.9	(0.89)
. Heat stage (1000 steps of .001ps at 2000K)	R=28.5	(1.95)
. Cool stage (220 steps at 300K)	R=23.7	(0.51)
. Final stage (200 steps of 'Jack-Levitt')	R=22.9	(0.22)
		(3.57)

Stereochemistry: rms bond deviation 0.026Å
rms angle .. 5.4°

It is interesting to note the large reduction in R factor during the prepare stage, which is probably due to careful choice of starting model and reflects the conservation of the immunoglobulin framework. Figure 11 shows a $2|Fo|-|Fc|$ map calculated with the L1 complimentarity determining region coordinates omitted from the calculation of phases; this shows very reasonable electron density for this loop.

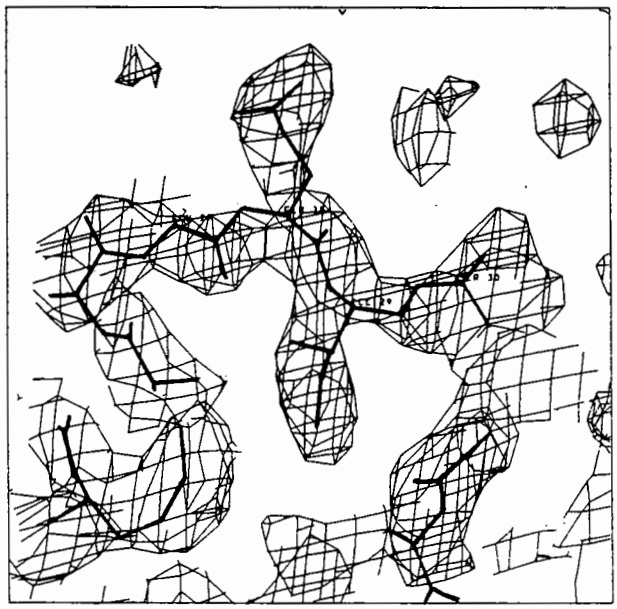


Fig. 11.

3. Phosphorylase b, R-state
(David Barford, Louise Johnson)

The R-state form of glycogen phosphorylase b crystallises in $P2_1$, with a tetramer of molecules in the asymmetric unit in 222 symmetry. The structure was solved using molecular replacement (MERLOT) with one T-state monomer as the search model. Like the P1 Fab structure above, success was achieved with only 1/4 of the scattering material as the search model. Rigid body refinement of the tetramer was carried out using CORELS prior to XPLOR refinement. Figs. 13,14 & 15 show three rthogonal views of the tetramer. The protocol used involved one complete run through XPLOR without any non-crystallographic restraints. The electron density of the4 molecules was averaged, and manual rebuilding was carried out. The tetramer was generated from this model, and Jack-Levitt refinement was carried out slowly relaxing restraints on the Cα positions. The final model showed improved electron

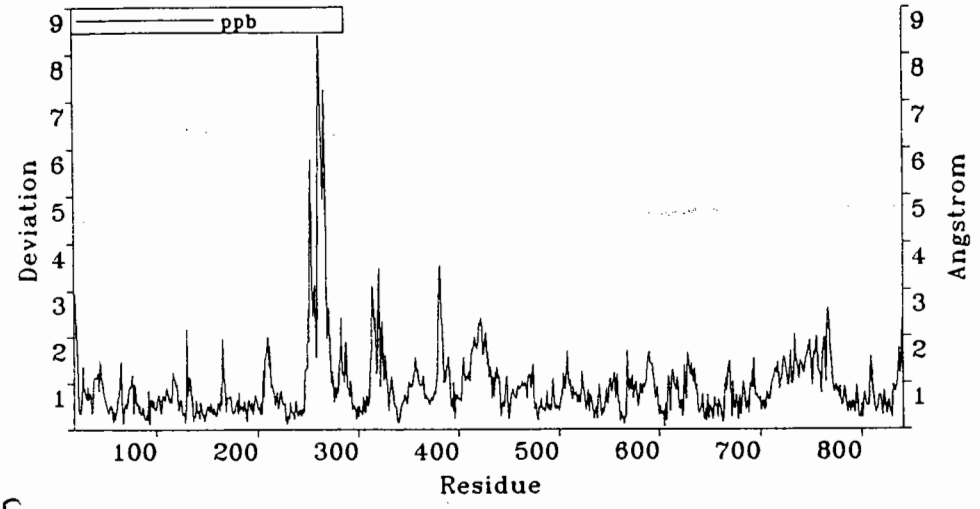


Fig. 12. Phosphorylase b: Calpha shift between T and R states

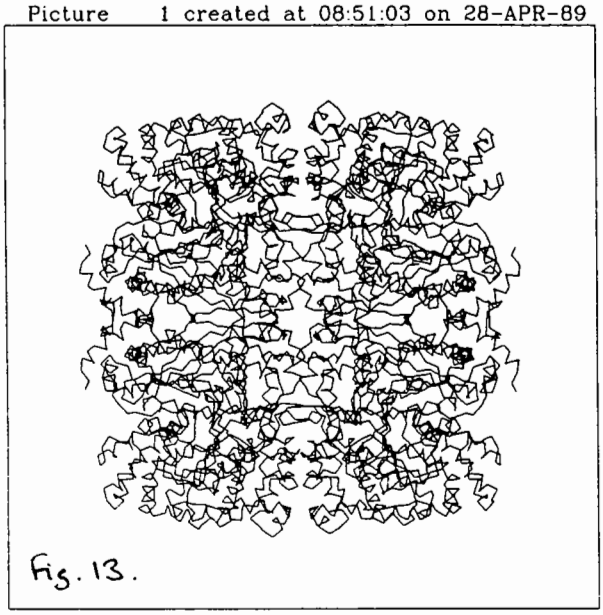


Fig. 13.

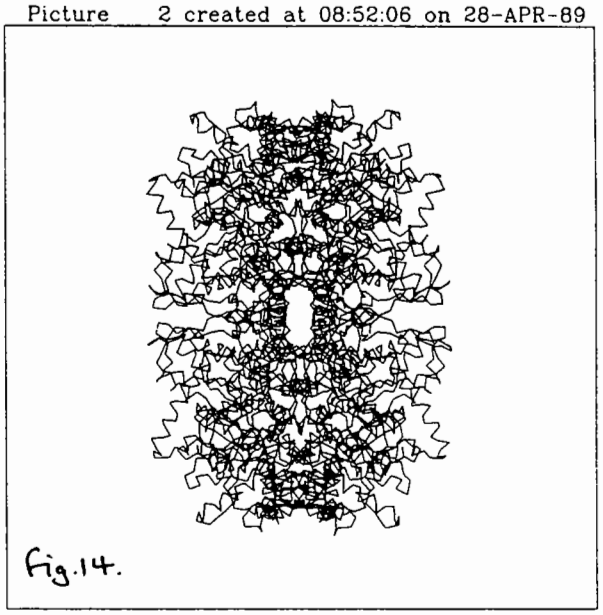


Fig. 14.

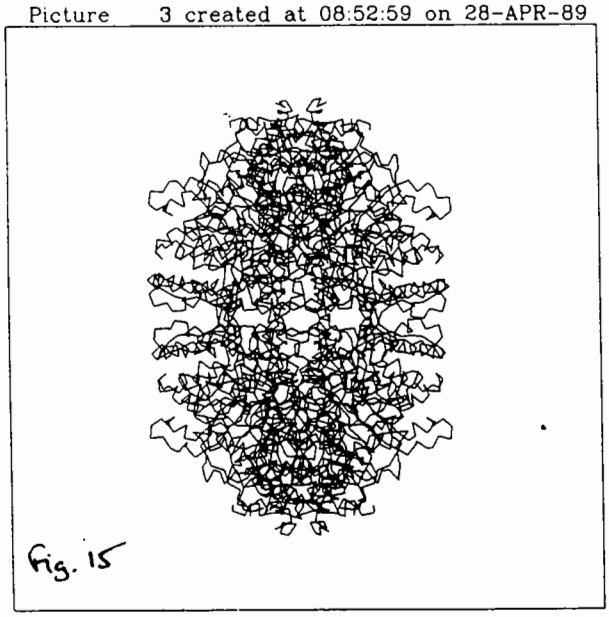


Fig. 15

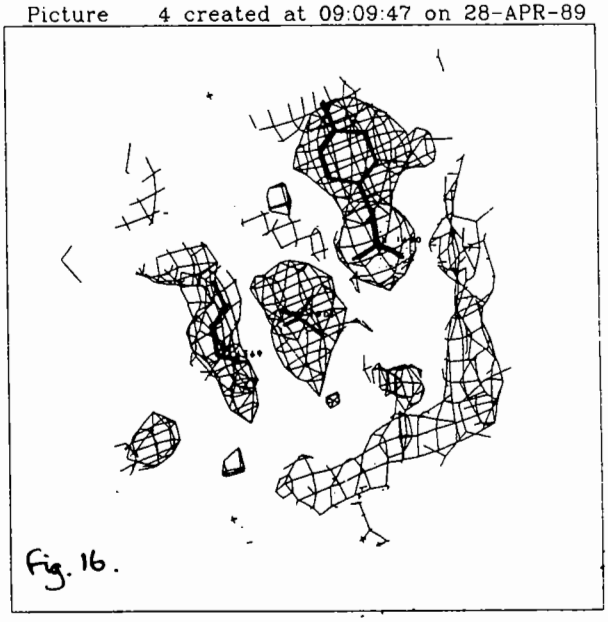


Fig. 16.

density; in particular the 'tower helix', was unclear in the initial map, but its large rotation became evident in the final map (fig.17, the T-state; fig.18, the R-state): this has proved a key feature of the proposed allosteric mechanism (Barford & Johnson, personal communication). Further evidence for the efficacy of the final model is the appearance of strong density in a difference map consistent with a sulphate ion close to the pyridoxal phosphate (fig. 16). The large shifts between the T and R state monomer structures is evident from fig. 12.

Phosphorylase b R-state 2000K MD refinement protocol
(26772 non-H atoms, 57291 F's in range 8.0 - 2.8Å)

	Initial R=41.8%	(C2 cpu hours)
. Prepare stage (160 steps of 'Jack-Levitt')	**R=27.0**	(12.95)
. Heat stage (1000 steps of .001ps at 2000K)	**R=32.1**	(23.52)
. Cool stage (250 steps of .001ps at 300K)	**R=23.8**	(5.68)
. Final stage (40 steps of 'Jack-Levitt')	**R=23.2**	(0.22)
. B factor refinement	**R=20.0**	(1.20)
		(43.57)

Stereochemistry: rms bond deviation 0.022Å
rms angle .. 4.6°

. Averaged electron density of 4 molecules	
. Rebuilt one molecule, create tetramer	**R=37.5**
. 'Jack-Levitt' refinement restraining to average structure	**R=29.7**
. Reaveraged coordinates, 'Jack-Levitt'	**R=27.5**
. Reaveraged, reduce restraint, J-L	**R=25.6**
. Reaveraged, Cα only restrained, J-L	**R=23.7**
. Remove restraints, J-L	**R=22.1**
. B factor refinement	**R=19.8**

Stereochemistry: rms bond 0.024Å
rms angle 4.7°

Picture 8 created at 16:59:31 on 30-APR-89

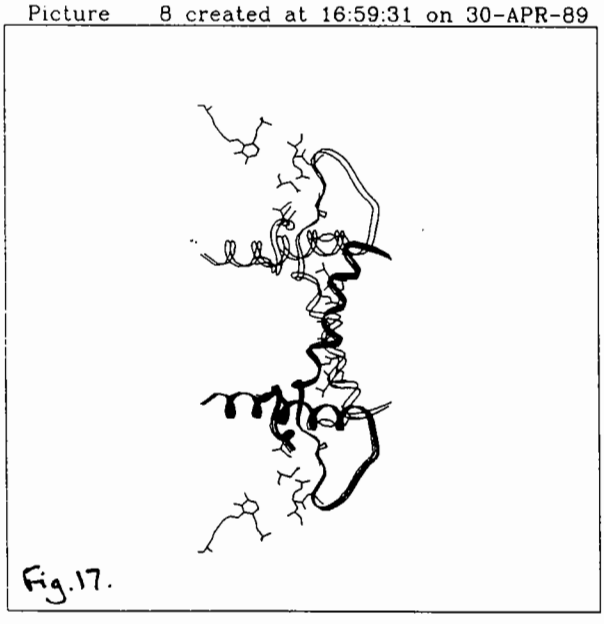


Fig.17.

Picture 1 created at 16:49:41 on 30-APR-89

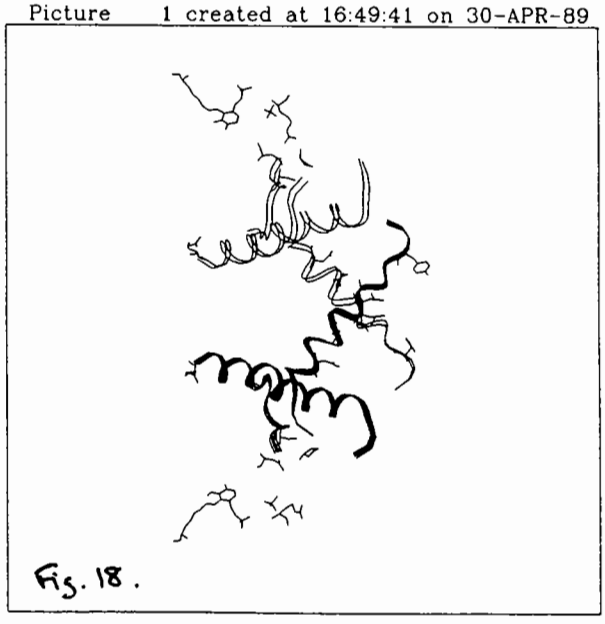


Fig. 18.

4. Tumour Necrosis Factor (TNF)

(Yvonne Jones, Dave Stuart, Nigel Walker)

TNF is a member of the family of cytokines and is of natural pharmaceutical interest. Its structure has recently been presented (Jones et al. 1989) following a major crystallographic effort because of the presence of two trimers of TNF in the asymmetric unit of a $P3_121$ crystal form. The structure was solved by novel Patterson search methods, followed by careful Bricogne averaging. Before refinement, one of the trimers had poorer electron density than the other, after XPLOR the density of the second became much clearer: figs. 19 & 20 show the electron density for part of trimer 2, before and after refinement; the corresponding density for trimer 1 was much clearer initially. The post-XPLOR map, fig. 20, is a $2|Fo|-|Fc|$ map with the atoms of the ß-strands in the picture omitted from the calculation of the phases. An important fact to note is that NO non-crystallographic restraints were imposed on the six copies during refinement, however as fig. 21 shows, the six copies showed a concerted movement from the starting model, and the refined B-factors also showed excellent correlation (fig. 23). Any deviations of the six structures from their average structure, appears mainly in the loops connecting ß-strands (fig. 22 - where the triangular symbols represent residues assigned to ß-strands).

TNF refinement protocol

(Two trimers in a.u., 7176 non-H atoms, 22845 F's in range 6.0-2.8Å)

	Initial R=48.5%	(C2 cpu hours)
. **Prepare stage (200 steps of J-L)**	**R=34.7**	(2.50)
. **Heat stage (1500 steps of .001ps at 2000K)**	**R=35.4**	(7.82)
. **Cool stage (2000K to 300K slow cooling)**	**R=29.5**	(3.90)
. **Final stage (50 steps of 'Jack-Levitt')**	**R=28.6**	(0.61)
. **B refinement (40 steps)**	**R=25.5**	(0.78)
		(15.61)

Stereochemistry: **rms bond deviation 0.017Å**
rms angle .. 3.6°

Fig. 19.

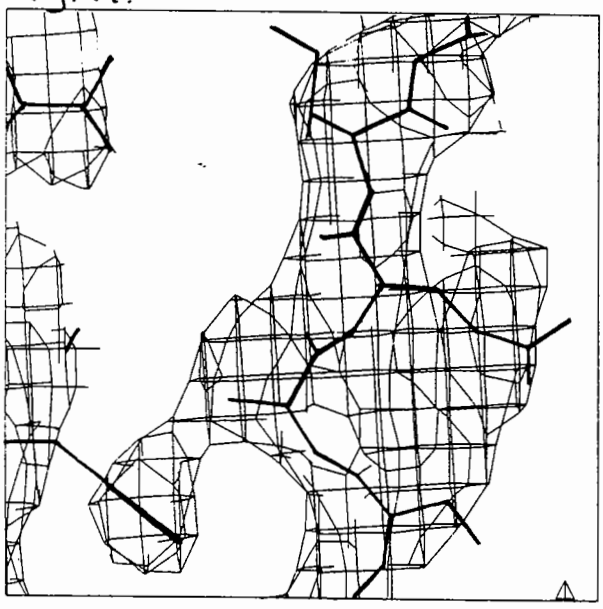
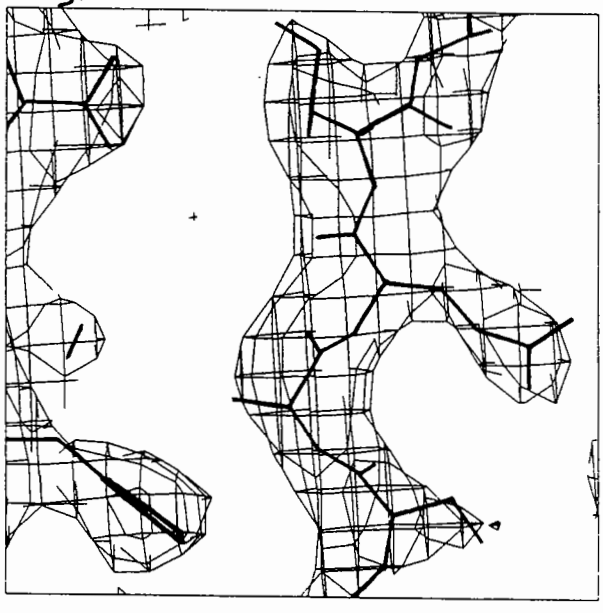


Fig. 20



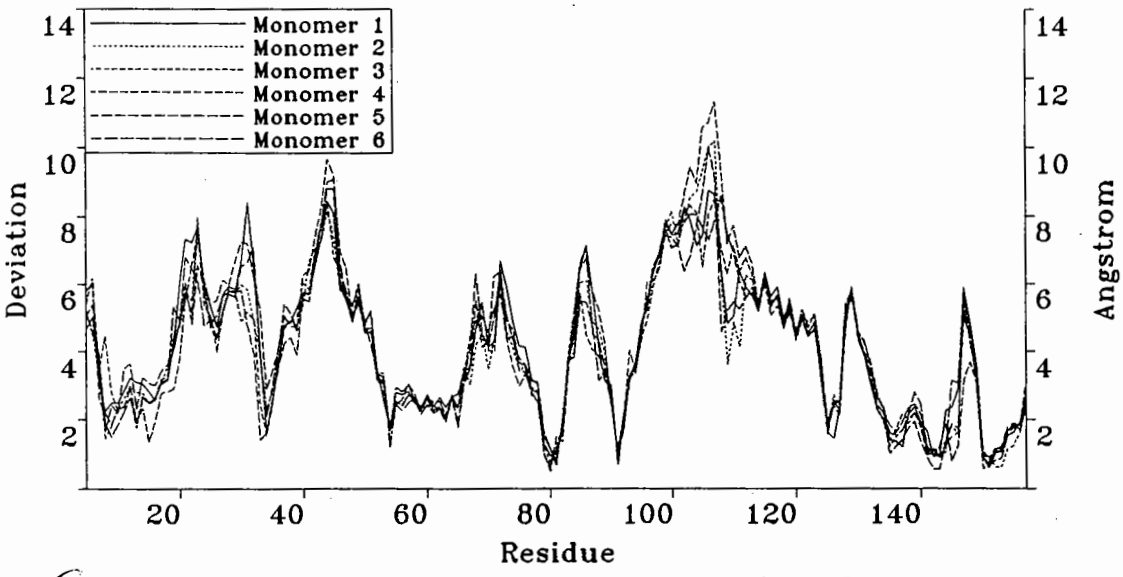


Fig.21. TNF: starting model v. refined monomers

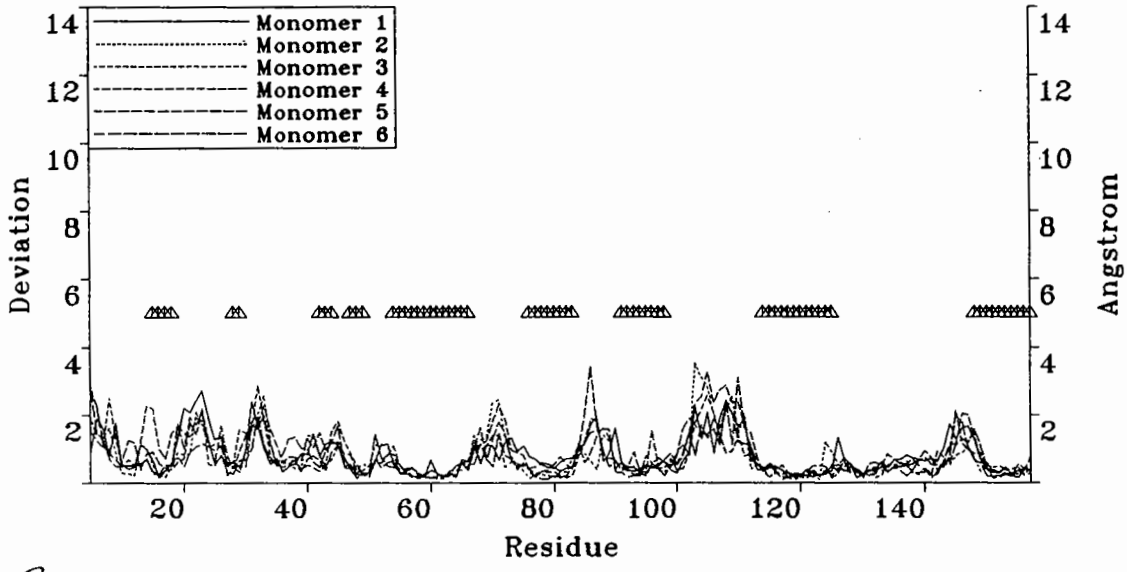


Fig 22. TNF: Deviations of the 6 monomers from the average

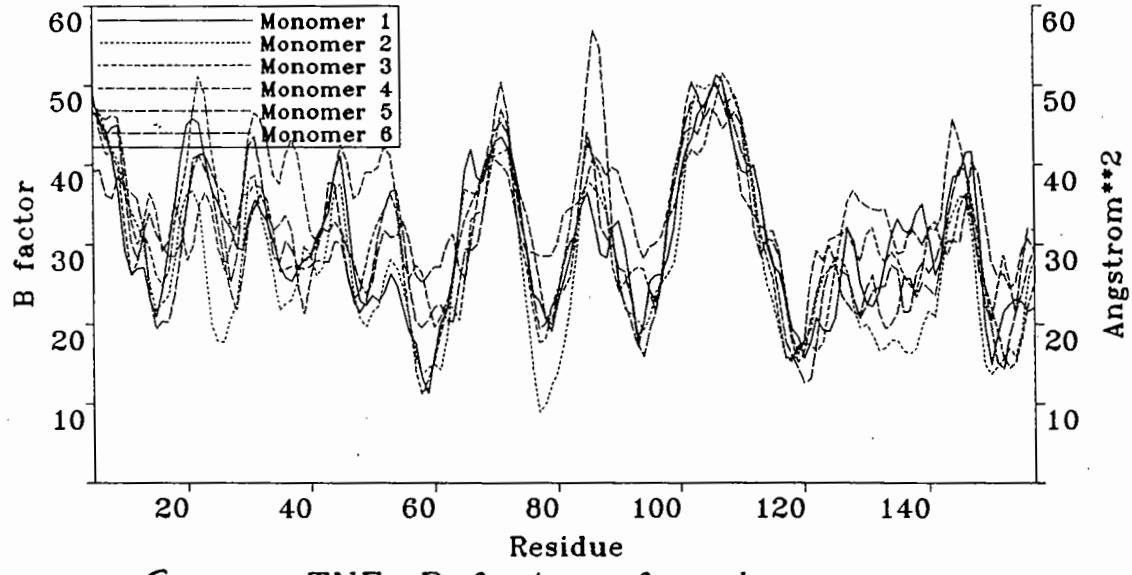


Fig.23. TNF: B factors for six monomers

5. Foot and Mouth Disease Virus (FMDV)

(Ravi Acharya, Liz Fry, Dave Stuart)

The structure of FMDV completes the structure determination of the picornavirus family of the common cold, mengo and polio viruses. Details of the structure determination have recently been published (Acharya et al. 1989. The crystal symmetry, I23, leaves a pentamer of the four coat proteins VP1,2,3,4 in the asymmetric unit. The averaged electron density map was very clear, giving an initial crystallographic R factor of 32.8% for the starting model. As for TNF, the five protomers were refined independently, with no non crystallographic restraints (XPLOR V1.3 was used for all the work presented here). Initially, we were cautious with the refinement because of the expected cpu usage. The Convex performed extremely well, however, with its virtual memory and efficient library FFT routine. The weight to be applied to the x-ray amplitude term, derived in the CHECK stage, was initially scaled up by an order of magnitude; this resulted in a rapid drop in the R-factor during Jack-Levitt refinement for 50 cycles. The weight was then adjusted to its recommended value and 10 further cycles allowed the stereochemistry to improve. Together with B-factor refinement, less than 24 hours of cpu was used. A full simulated annealing run was then carried out, after producing an average protomer from the previous refinement run. Again no restraints for non-crystallographic symmetry were imposed, and the final R dropped to 17.4% with acceptable stereochemistry. This whole run took 190 Convex cpu hours: it is humbling to think that on a microVax II, it might have taken 1.5 years!!

Figs. 24 & 26 show the correlated shifts of the five copies of VP1 relative to the starting model, and the concomitant correlation of their B-factors. The five copies also deviate very little from the final averaged structure of VP1 (fig. 25).

FMDV refinement protocol

(Pentamer of VP1,2,3,4 in a.u., 25475 non-H atoms, 102853 F's 8.0-2.9Å)

		Initial R=32.8%	(C2 cpu hours)
1. 'Jack-Levitt' refinement alone			
. Prepare stage (50 steps of J-L, 10*WA)		R=22.9	(14.24)
. Prepare stage (10 steps of J-L, WA)		R=25.4	(2.30)
. B factor refinement (20 steps)		R=22.1	(7.13)
			(23.67)
2. MD refinement			
. Averaged 5 copies of coordinates	Initial	R=24.4%	
. Prepare stage (60 steps of J-L)		R=20.4	(14.61)
. Heat stage (1500 steps of .001ps at 2000K)		R=22.8	(143.23)
. Cool stage (250 steps of .001ps at 300K)		R=18.7	(18.64)
. Final stage (40 steps of 'Jack-Levitt')		R=17.9	(10.02)
. B factor refinement (10 steps)		R=17.4	(3.64)
			(190.14)

Stereochemistry: rms bond deviation 0.017Å
rms angle .. 3.6°

Conclusion

The success obtained with the six structures described here show the power of the technique of refinement using restrained molecular dynamics. The radius of convergence is certainly much greater than that seen in least squares methods, where atom movement is limited to approximately $d_{min}/4$. Atomic shifts of up to 12Å

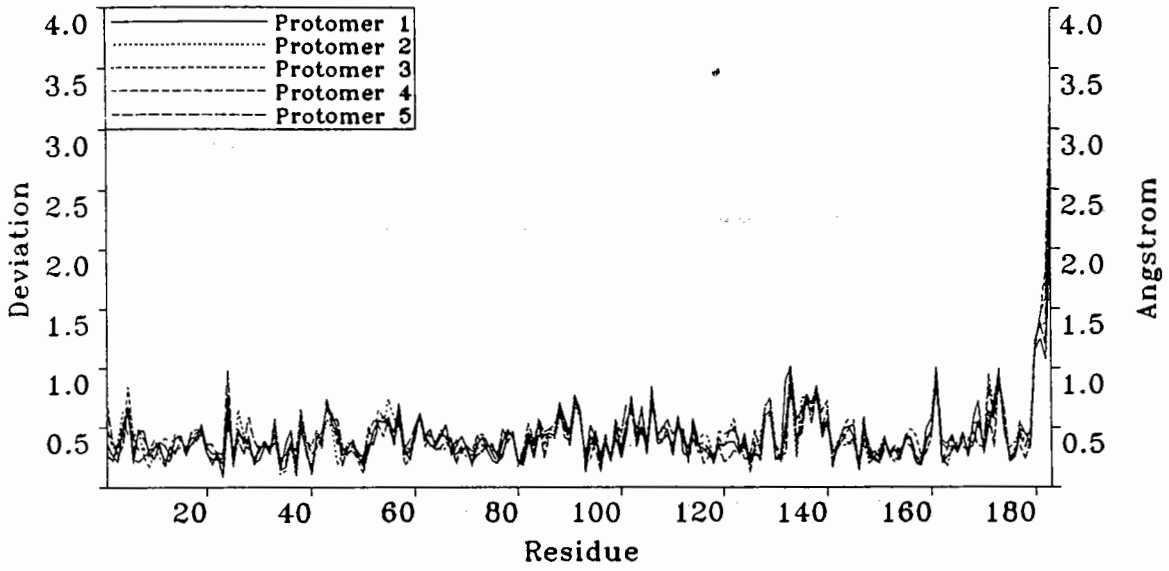


Fig. 24. FMDV: starting VP1 model v. refined protomers

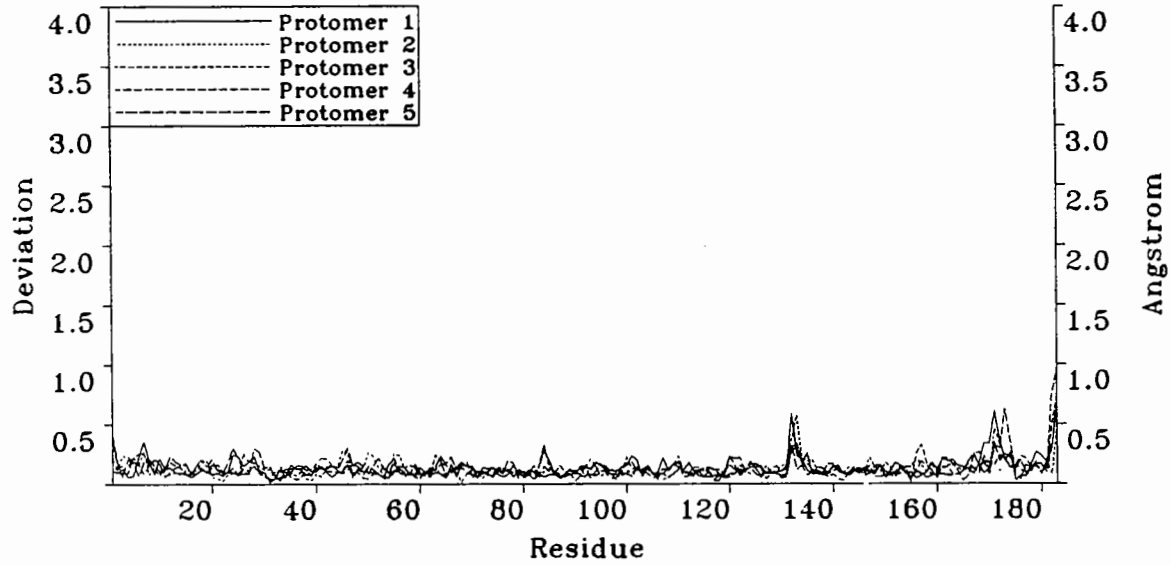


Fig. 25 FMDV: refined protomers v. average for VP1

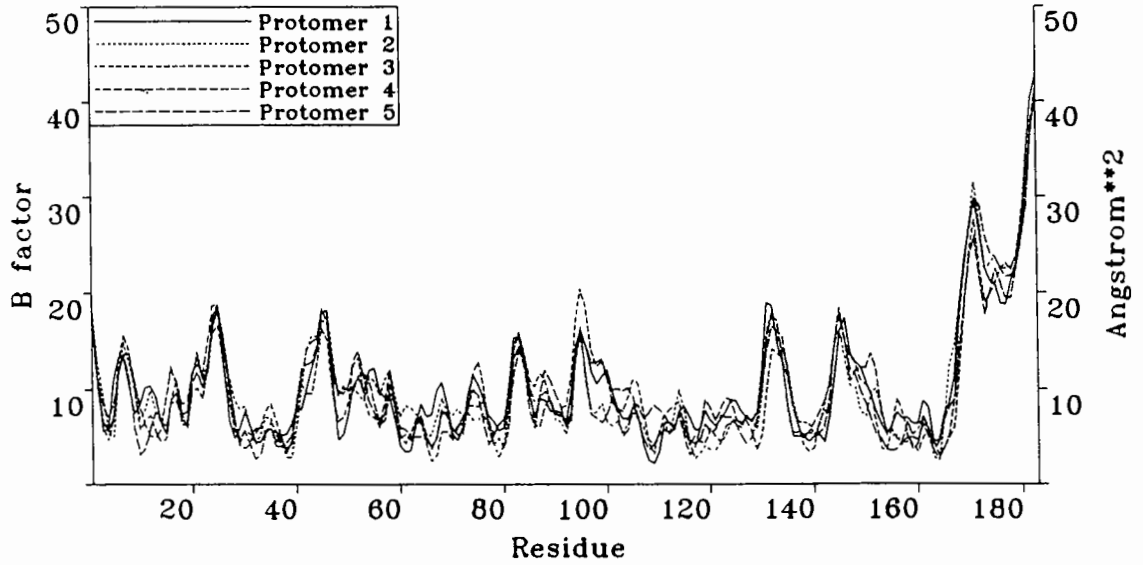


Fig. 26. FMDV: B factor comparison for VP1

from the starting model have been observed, and the concerted shifts observed in the multimeric proteins adds credence to such movements. Given a powerful enough computing resource, refinement is now a much faster procedure; the problem remains however of looking carefully at the final structure on the graphics and identifying 'problem' areas from high B-factors, poor difference maps or from energetic strain. Since presenting the talk, we have now carried out further refinements of all of the above proteins, including solvent in some cases. In addition, the structures of ß-lactamase I, 6-phosphoglycerate dehydrogenase, phosphoglycerate kinase, human α-lactalbumin, and several phosphorylase/ligand complexes have been or are in the process of being refined. The Convex C210 has proved to be an excellent machine for such work with its combined attributes of fast scalar and vector processing, high memory bandwidth and I/O. The latter stages of protein structure determination have kept pace with developments in rapid data collection: if only crystallisation methodology was as highly developed!

Acknowledgements

I would like to thank all my colleagues who generously allowed me to present the refinement aspects of their projects. I would also wish to thank the SERC/MRC for provision of the Convex.

References

Acharya, R., Fry, E., Stuart,D.I., Fox, G., Rowlands, D. & Brown, F. Nature **337** (1989) 709-716

Browne, W.J., North, A.C.T., Phillips, D.C., Brew, K., Vanamann, T.C. & Hill, R.L. J. Mol. Biol. **42** (1969) 65-86

Brunger, A.T. J. Mol. Biol. **203** (1988) 803-816

Brunger, A.T. Acta Cryst. **A45** (1989) 42-50

Brunger, A.T., Karplus,M. & Petsko,G.A. **A45**(1989) 50-61

Fitzgerald, P.M.D. J. Appl. Cryst. **21** (1988) 273-278

Fujinaga, M. & Read, R.J. J. Appl. Cryst. **20** (1987) 517-521

Furey, W., Wang, B.C., Yoo, C.S. & Sax, M. Acta Cryst. **A35** (1979) 810-817

Hendrickson, W. & Konnert, J.H. Computing in Crystallography, eds. Diamond, R., Ramaseshan, S. & Venkatesan, K. (1980) 13.01-13.23. Published by The Indian Academy of Sciences.

Jones, E.Y., Stuart,D.I. & Walker,N.P.C. Nature **338** (1989) 225-228

Stuart, D.I., Acharya, K.R., Walker, N.P.C., Smith,S.G., Lewis,M. & Phillips, D.C. Nature **324** (1986) 84-87

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

Crystallographic Refinement using Molecular Dynamics: An Application to Serum Transferrin

Harren Jhoti
Laboratory of Molecular Biology
Department of Crystallography
Birkbeck College
Malet St.
London

1 Introduction

Conventional crystallographic refinement has been used successfully in the past to refine protein and nucleic acid structures determined using X-ray diffraction techniques (Jensen,1985 see references therein). The process consists of several cycles of least-squares refinement with stereochemical restraints followed by model-building using interactive computer graphics. This process has several short-comings. The most serious of these is the problem of the model converging to a local minimum. This is due to the limited radius of convergence as restrained least-squares refinement does not usually correct atoms whose positions are more than 1Å in error. Human intervention then becomes necessary to move the model out of the local minimum and interactive computer graphics are used to rebuild the parts of the model that are in error; this stage can become very time-consuming.

Recently, a new approach to crystallographic refinement has been reported where the technique of molecular dynamics is utilised in order to search the conformational space of a molecule (Brünger *et al.*,1987; Brünger,1988a). Brünger *et al.* (1987) have shown that refinement using molecular dynamics (MD-refinement) has a larger radius of convergence than conventional restrained least-squares refinement and that the method can reduce the need for manual corrections. The 'target' function used in MD-refinement accounts for the diffraction data as well as describing the stereochemistry and non-bonding interactions of the molecule (Brünger,1988a).

The program X-PLOR has been developed to perform MD-refinement of macromolecular systems (Brünger,1988a). A description of the potential energy functions used in X-PLOR is provided by other authors. An application of MD-refinement

	R-factor after refinement	Mean FOM	Side-chains present (%)	FOURIER COEFFICIENTS		
				Fo ϕ_{comb}	2Fo-Fc ϕ_{calc}	2Fo-Fc ϕ_{comb}
MODL1	48.8%	0.69	38	COMB1A	COMB1B	—
MODL2	26.6%	0.85	50	COMB2A	COMB2B	COMB2C
MODL3	26.5%	0.79	65	COMB3A	COMB3B	COMB3C

Table 1: Progress of the refinement of serum transferrin and the Fourier synthesis calculated using RESTRAIN.

using X-PLOR was made during the crystallographic refinement of rabbit serum transferrin and is described in this paper.

2 The Problem

The molecular structure of rabbit serum transferrin has been solved to a resolution of 3.3 Å using MIR and solvent flattening techniques (Bailey *et al.*,1988). The refinement of serum transferrin was initiated using a conventional least-squares refinement program RESTRAIN (Haneef *et al.*, 1985) combined with phase combination techniques (Rice,1981). Table 1 shows the progress of the refinement using RESTRAIN.

The first model built (MODL1) contained about 90% of the backbone and was refined using rigid-body refinement only with data in the 3.3-5.0 Å range and no sigma cut-off. It should be realised that at this stage the model-building was based largely on the sequence of human serum transferrin (Yang *et al.*,1984) as only a little of the rabbit sequence was known (H. McKenzie, personal communication). The rabbit sequence, which was found to be highly homologous to human, was incorporated as it became known.

Both MODL2 and MODL3 were refined using restrained refinement with a data range of 3.3-5.0 Å with an overall temperature factor. A unit weighting scheme where all reflections, above a 3σ cut-off, have an equal weight was applied to the data. As shown in table 1 combined phase sets were used in the calculation of electron density maps at several stages. In these cases the original MIR phase set was combined with the phase set calculated from the relevant model using the program COMBINE (Bricogne,1976). Also shown in table 1 is the mean figure of merit (FOM) for the combined phase sets, with the original MIR phase set having a value of 0.52.

After the refinement of MODL2 the electron density in several loop regions remained ambiguous and so these loops were removed before the refinement of MODL3. As seen there was no further reduction in the R-factor during the refinement of MODL3 (R=26.5%) and furthermore there were no significant improvements in the electron density maps COMB3A, COMB3B or COMB3C which were

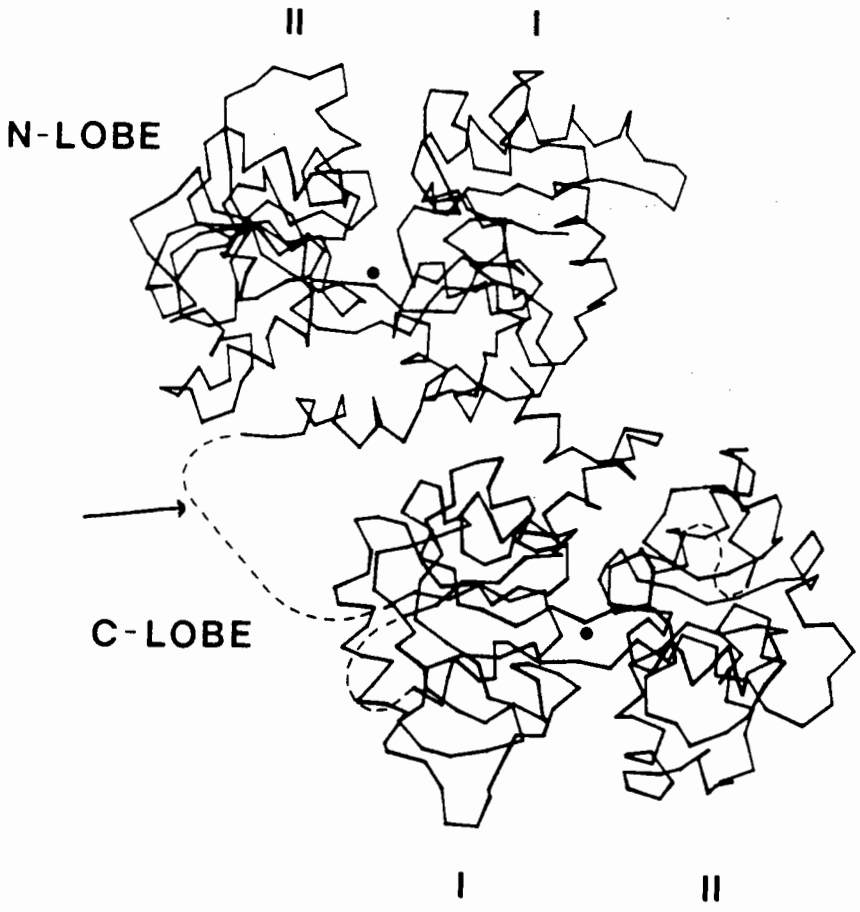


Figure 1: An C^{α} plot of MODL3 the final model refined using RESTRAIN. The structure has a bilobal shape with each lobe divided into domains I and II. Dotted lines indicate breaks in the model. The inter-connecting peptide region is arrowed and the two dots represent the iron atoms.

based on the refined MODL3.

During protein structure refinement an impasse is often encountered when the R-factor is in the mid-thirties. It may be noted that the R-factor calculated for MODL3 using the data range 8.0-3.3 Å with a 2σ cut-off was actually 34.4%. For the refinement to have proceeded further major re-building of parts of the model would have been required. Indeed the refinement had fallen into a local minimum. An application of MD-refinement to this problem seemed to be an ideal test to evaluate the potential of this novel refinement technique.

3 An Application to Transferrin

The program X-PLOR (version 1.3) was implemented on the Cray X-MP at the Rutherford Appleton Laboratory, Didcot, U.K.

3.1 Pass 1

3.1.1 Method

The starting model used was the final model refined using RESTRAIN (MODL3) after some modifications had been made to it. Firstly, the inter-connecting peptide which consists of seven residues (323-329) and connects the two lobes in the molecule (see figure 1) was inserted as a poly-alanine chain. At each end of this seven residue stretch are located disulphide bridges for which there was electron density in previous maps. Hence, residues 322 and 330 are both half disulphide bridges which 'pin down' the ends of the inter-connecting peptide to the two lobes (see figure 2). Based on the amino-acid sequence of human serum transferrin it was known that there were seven residues between these two disulphide bridges. The conformation of these residues was unknown as there was only ever putative density in this region of the previous electron density maps. Thus, it was decided to build the inter-connecting peptide into an arbituary conformation.

The topology of the two lobes in serum transferrin are very similar. Indeed, using the present model the two lobes can be superimposed to give an r.m.s deviation of 1.1 Å for 223 pairs of structurally equivalent C^{α} atom positions. A loop region consisting of 18 residues (601-619) which was only defined well in the N-lobe, but according to the human serum sequence was present in both, was built into the corresponding position in the C-lobe where the density was ill-defined using a similar conformation as observed from the N-lobe.

These insertions were made on the basis that the radius of convergence for MD-refinement is in the order of 5-7 Å (Brünger,1988b) and so the correct conformations for these segments would be found. Hence, it was to be a test to investigate the ability of the program to search conformational space.

The new starting model (X-PLOR1) contained 4284 atoms and was refined using data in the 8.0-3.3 Å range with a 2σ cut-off. The protocol used in pass 1 consisted of 6 stages as outlined in table 2. It should be remembered that during energy minimization cycles the function being minimized includes terms for the diffraction data unless stated otherwise.

Stage 1

The first stage establishes the relative weighting of the empirical potential energy E_i to the effective potential energy E_x that is to be used (Brünger,1988a). A suggested value of 255690 Kcal/mole for W_A was calculated and used throughout pass 1. As phase restraints were not used in this MD-refinement W_P was kept as zero.

Stage 2

Once the relevant weights for E_i and E_x have been determined energy minimization is used to relieve bad inter- and intramolecular contacts and also to reduce the potential energy of the molecule. To avoid large shifts of the backbone while relieving steric conflicts the C^{α} positions are fixed to their original positions using

harmonic restraints at 20 Kcal/(mole Å). A tolerance value of 0.05 Å was used such that the structure factors are updated if an atom moves greater than this distance.

Stage 3

The heating stage consisted of a molecular dynamics simulation over 3 ps at a temperature of 2000K with a 1 fs timestep. The initial velocities were assigned from a Maxwellian distribution at 2000K and rescaled every 25 steps.

Stage 4

A slow cooling option was employed in order to anneal the system. The temperature of the system is reduced from 2000K to 300K over a 2 ps time period.

Stage 5

This stage consisted of a further 120 energy minimization cycles using the full nonbonded potential.

Stage 6

For the final stage a grouped temperature factor refinement was performed where each residue is assigned two temperature factors, one for the backbone atoms and one for the side-chain atoms. In view of the limit of the resolution this type of temperature factor refinement was considered to be justified. New electron density maps on the basis of the refined model (X-PLORED1) were then calculated using Fourier coefficients $|2|F_o| - |F_c||\phi_c$ and $||F_o| - |F_c||\phi_c$ and investigated for improvements (see table 3).

3.1.2 Results

As shown in table 4 the progress of the MD-refinement of transferrin can be followed by monitoring the R-factor during the different stages. An initial R-factor of 43% (and not 34%) was due to the changes made to the model (MODL3) which had been previously refined to an R-factor of 34% using restrained least-squares refinement. After the second stage, the Jack-Levitt type refinement, the R-factor was reduced to 32.5% and a similar but slightly better situation to that produced using restrained least-squares refinement was achieved. However, simulated annealing as performed in stages 3 and 4 was successful in reducing the R-factor of the structure to 28.7% which otherwise would have only been achieved by extensive re-modelling. The stereochemistry and R-factor were further optimized by stage 5 and after the separate refinement of the temperature factors, stage 6, a final R-factor of 26% was achieved. The overall geometry of the model appears to be satisfactory (table 5) and so the reduction in the R-factor is not at the expense of the geometry.

A comparison of the final model (X-PLORED1) with the initial model (X-PLOR1) is summarised in table 6. The analysis shows that 191 backbone atoms, i.e about 10% of the total backbone, moved by more than 2.0 Å. This obviously would not have been possible using a conventional least-squares refinement approach without manual intervention. Furthermore, the maximum shift for a backbone atom was observed at the C^{α} position of residue 327. This residue is located in the interconnecting peptide and its movement is discussed later. Most of the movement

Stage	Parameters
1.	Determination of weights W_A,W_P
2.	Conjugate gradient minimization - 40 cycles using the repulsive potential followed by 120 cycles using the full nonbonded potential - resolution range; 8.0-3.3 Å, B=15.0 Å - C^{α} restraints at 20Kcal/ (mole Å),W_A=255690 - tolerance=0.05 Å
3.	Heat - 3 ps, T=2000K, timestep=1fs, velocities rescaled every 25 fs, resolution range; 8.0-3.3 Å W_A=255690,B=15.0 Å, tolerance=0.2 Å.
4.	Cool - 2ps, initial T=2000K, final T=300K, timestep=1fs temperature increment =-25K, resolution range; 8.0-3.3 Å B=15.0 Å W_A=255690, tolerance=0.2 Å
5.	Conjugate gradient minimization - 120 cycles using full nonbonded potential - resolution range; 8.0-3.3 Å B=15.0 Å - W_A=255690, tolerance=0.005 Å
6.	Temperature factor refinement - grouped temperature factor refinement, - two for each residue, one for sidechain and one for the backbone

Table 2: PASS 1: Protocol for the MD-refinement of transferrin.

	Fourier coefficients	
Model	$2F_o - F_c\phi_{calc}$	$F_o - F_c\phi_{calc}$
X-PLORED1	XPMAP1A	XPMAP1B
X-PLORED2	XPMAP2A	XPMAP2B

Table 3: The different electron density maps calculated using the two refined models during MD-refinement.

Stages	R-factor (%) PASS 1	PASS 2
1. Weight determination	43.0	37.4
2. Energy minimization	32.5	27.2
3. Heat	32.4	29.0
4. Cool	28.7	25.1
5. Energy minimization	27.9	24.4
6. B-factor refinement	26.0	21.9

Table 4: Progress of MD-refinement of transferrin: R-factor during Pass 1 and Pass 2.

Type	R.M.S. Deviation in Stereochemistry PASS 1	PASS 2
Bonds (Å)	0.021	0.021
Angles (°)	4.750	4.322
Dihedrals (°)	26.31	27.11
Impropers (°)	2.63	2.58

Table 5: R.M.S. deviations in the stereochemistry for both X-PLORED1 (Pass 1) and X-PLORED2 (Pass 2) models.

in the backbone of the structure corresponded to loop regions with the secondary structural elements showing little change.

One region of the structure that showed a large concerted movement was the inter-connecting peptide. Figure 2 shows the initial and final conformations for this region of structure after pass 1. It can be seen in figure 2 that the new conformation for the inter-connecting peptide corresponds well with electron density in the new maps XPMAP1A and XPMAP1B which were calculated using the refined model X-PLORED1 (see table 3). The electron density map XPMAP1B is an omit map calculated using the refined model but with the seven residues of the connecting peptide omitted from the structure factor calculations. Although the electron density for the inter-connecting peptide is discontinuous in the new maps it is a significant improvement compared with previous electron density maps.

As described earlier an 18 residue loop was positioned into an ill-defined region of the electron density map in the C-lobe on the basis of structural homology with the N-lobe. As expected there was little difference between the initial and final conformations of this loop. A slight improvement in the electron density of this region was observed. The maximum shifts observed were in side-chain positions where atoms moved distances of up to 7 Å into density (see figure 3). As expected MD-refinement proved rather cpu intensive; the whole of pass 1 required about 12 hours of cpu time on the Cray X-MP.

Comparison between X-PLOR1 and X-PLORED1	
Backbone atoms (C,C^{α},N) :	-191 atoms shifted by > 2.0Å -max. shift of 4.96 Å observed at C^{α} of residue 327 -R.M.S shift =1.26 Å
Side-chain atoms (the rest) :	-24 atoms shifted by > 5.0Å -max. shift of 7.1 Å observed at Nz of residue 345 -R.M.S shift =1.85 Å

Table 6: Comparison between the initial (X-PLOR1) and final (X-PLORED1) models in Pass 1 of MD-refinement.

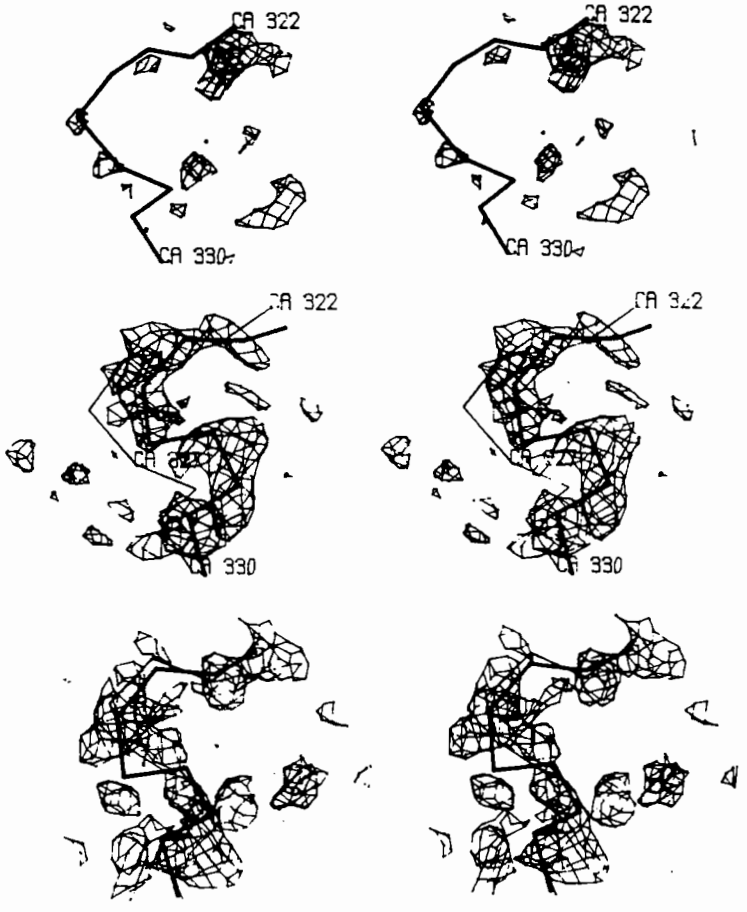


Figure 2: TOP: The inter-connecting peptide built into an arbituary conformation as there is little density (COMB3C) present. MIDDLE: A concerted movement is observed during pass 1. The new position (bold) corresponds well with density from the omit map XPMAP1B. BOTTOM: Only a slight shift from the old position (bold) is observed during pass 2 suggesting the right conformation had been found during pass 1. The density shown is from XPMAP2A.

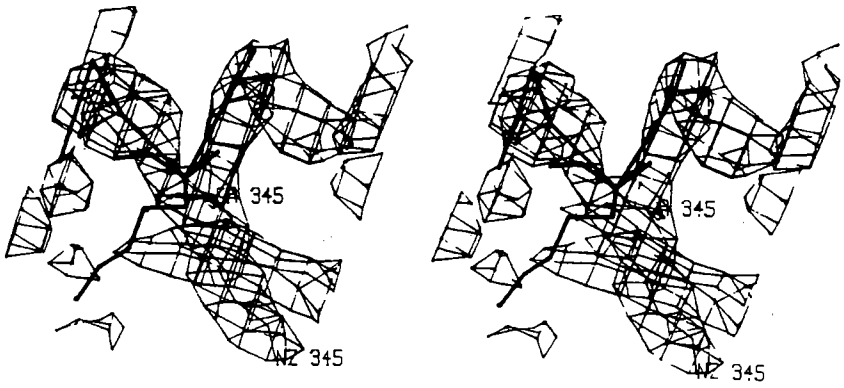


Figure 3: The maximum shift during pass 1 was observed in lysine 345 where the Nz atom moved 7.1 Å into new density (XPMAP1A). Note the movement of the backbone in this helical region. The old position of the side-chain and part of the helix is shown in bold.

3.2 Model-building strategy

The new electron density map XMAP1A (table 3), calculated using the refined model (X-PLORED1), showed significant improvements. This map in conjunction with XMAP1B allowed many more side-chains to be inserted and the model was re-built using the program FRODO (Jones,1978) implemented on the Evans and Sutherland picture system 300.

As the radius of convergence for MD-refinement was shown to be in the order of 5-7 Å, an adventurous strategy was employed when model-building. It was felt that to exploit this novel technique to the full the model needed to be as complete as possible. So, as many side-chains as possible were inserted even if there was no apparent density for them (as was the case for about 20 lysine and arginine residues located on the surface of the protein). Indeed, it would be dangerous to omit parts of the structure that were known to exist (see below).

The resulting model (X-PLOR2) contained 5230 atoms and according to the human serum transferrin sequence contained all but six of the 679 residues. At this stage about 60% of the rabbit serum transferrin sequence had been elucidated and incorporated (H.McKenzie, personal communication; R.MacGillivray, personal communication).

3.3 Pass 2

3.3.1 Method

As pass 1 had significantly improved the model it was decided to perform pass 2 using a very similar protocol to the one described above (table 2). In stage 1 W_A was calculated to be 293000 and this value was used throughout pass 2. Stage 2 was slightly modified such that 160 cycles of energy minimization using the full nonbonded potential were performed, otherwise the protocol used in pass 2 was

Comparison between X-PLOR2 and X-PLORED2	
Backbone atoms (C,C^{α},N):	-47 atoms shifted by > 2.0 Å -max. shift of 7.6 Å observed at N of residue 1 -R.M.S shift=0.81 Å
Side-chain atoms (the rest):	-38 atoms shifted by > 5.0 Å -max. shift of 8.5 Å observed at Nz of residue 148 -R.M.S shift =1.61 Å

Table 7: Comparison between the initial (X-PLOR2) and final (X-PLORED2) models in pass 2 of MD-refinement.

identical to the one shown in table 2.

The starting model for pass 2 was X-PLOR2 which contained 5230 atoms and it was refined against the data range 3.3 – 8.0 Å using a 2σ cutoff. Before pass 2 the model X-PLOR2 was re-modelled such that the alanine residues in the inter-connecting peptide were replaced by the relevant residues based on the human sequence.

3.3.2 Results

The progress of the refinement can be monitored by the R-factor (see table 4). It can be seen that the simulated annealing (stages 3 and 4) reduced the R-factor by about 2% compared with about 4% in pass 1. This suggests that fewer atoms needed to be moved large distances in pass 2. The final R-factor after grouped temperature factor refinement was 21.9% resulting in a drop of about 4% from the value obtained after pass 1 which was 26%.

Table 5 shows r.m.s deviations in the geometry for X-PLORED2 to be comparable to those calculated for X-PLORED1. A comparison of the initial (X-PLOR2) and final (X-PLORED2) models in pass 2 (table 7) show that fewer backbone atoms have moved a distance greater than 2 Å compared with pass 1.

It is interesting to note that the conformation of the inter-connecting peptide changed only slightly after pass 2. This suggests that MD-refinement was successful in finding the correct conformation for this region (figure 3). The new electron density maps (see figure 3), although much better, are still discontinuous in this region. This may be due to sequence changes in the rabbit protein which are now known to be present in this region (R. MacGillivray, personal communication).

The movements in side-chains during pass 2 where of the same order as observed in pass 1. The maximum shift in a side-chain atom was observed at the Nz of lysine 148 (figure 4). This example highlights the advantage of using a complete model (and the danger of not). There was no previous density for lysine 148, however, it

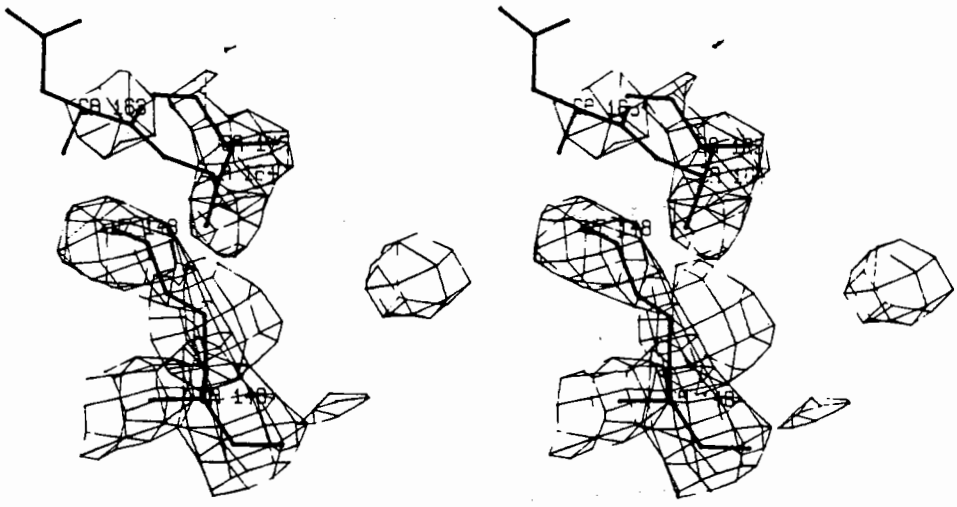


Figure 4: Lysine 148 moved into density (XPMAP2A) which was previously, and incorrectly, occupied by residue 164. The Nz atom moved 8.5 Å during pass 2.

was still inserted during the model-building after pass 1. Its new position, after pass 2, corresponds well with very good density which was previously (and incorrectly) occupied by the backbone region around residue 164. This stretch of backbone moved into a different region of density which resulted in a better connectivity for residue 164. This local reshuffling probably would not have occured had lysine 148 not been inserted.

4 Conclusions

These results show the immense potential of crystallographic refinement using molecular dynamics as applied to protein structures. It has been shown that the radius of convergence for MD-refinement is much larger than for conventional restrained least-squares refinement. This greatly reduces the level of manual intervention required during the refinement process. MD-refinement using X-PLOR has greatly accelerated the refinement of the serum transferrin structure.

The ability of MD-refinement to search conformational space was successfully exploited in the case of the inter-connecting peptide. The final conformation of the inter-connecting peptide corresponded well to the electron density in the region. This has important implications in terms of refining protein structures where there are ill-defined loop regions. Indeed, during the refinement of most protein structures it is the usually the loop regions that pose the greatest problems. As seen, there are dangers when an incomplete model is refined using molecular dynamics.

Using MD-refinement it may be possible to obtain several structures, varying slightly in conformation, that agree equally well to the experimental data. This information would be of more relevance to understanding biological function than a single averaged protein structure.

4.1 Acknowledgements

I acknowledge the support and encouragement I received from my colleagues in the Transferrin group at Birkbeck who are Dr. P.F.Lindley, Dr. B.Gorinsky, Dr. A. Walton, R.C. Garratt and B. Sarra. Also, I thank Dr. Ian Tickle for assistance during the implementation of X-PLOR. Financial support was provided by the Medical Research Council.

References

[1] Bailey,S., Evans,R.W., Garratt,R.C., Gorinsky,B., Hasnain,S.S., Horsburgh,C., Jhoti,H., Lindley,P.F., Mydin,A., Sarra,R. and Watson,J.L. Biochemistry, **27**, (1988) 5804.

[2] Bricogne,G. Acta Cryst., A**32**, (1976) 382.

[3] Brünger,A.T., (1988a) in 'Crystallographic Computing 4: Techniques and New Technologies', edited by N.W. Isaacs and M.R. Taylor. Oxford Univ. Press.

[4] Brünger,A.T. J. Mol. Biol., **203**, (1988b) 803.

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

[6] Brünger,A.T., Karplus,M. and Petsko,G.A. Acta Cryst., A**45**, (1989) 50.

[7] Haneef,I., Moss,D.S., Stanford,M.J. and Borkakoti,N. Acta Cryst., A**41**, (1985) 426.

[8] Jensen,L.H. in 'Methods in Enzymology', **115**, (1985) 227

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

[10] Rice,D.W. Acta Cryst., A**37**, (1981) 491.

[11] Yang,F., Lum,J.b., McGill,J.R., Moore,C.M., Naylor,S.L., van Bragt,P.H., Baldwin,W.D. and Bowman,B.H. PNAS, **81**, (1984) 2752.

Computer Simulations of Many Particle Systems

I Haneef

Astbury Department of Biophysics
University of Leeds
Leeds LS2 9JT, U.K.

1 Introduction

There now exist a large number of exquisite biological examples of molecular recognition processes ranging from the the highly specific binding of small ligands to proteins, to the interaction between two large macromolecules such as proteins/nucleic acids and antibodies/antigens. Site directed mutagenesis (Winter and Fersht, 1984) has provided a very powerful approach to altering the biological function and structural stability of proteins, and can be used to test structure-function hypotheses by introducing new amino acids in positions thought to be responsible for molecular recognition. Such experiments provide, in principle, a powerful approach for the design of molecules with modified or novel properties of clinical or industrial importance.

Despite the considerable body of structural and biochemical data available, the exact nature of the interactions responsible for highly specific recognition processes is not well understood. This lack of understanding currently poses a major obstacle to the design of novel molecules, proteins and drugs, with specific properties (Blundell and Sternberg, 1985). It is important, therefore, to develop quantitative methods that can be used to understand these processes in terms of basic interactions at the atomic level and enhance our understanding of biological processes. Theoretical calculations, using empirically derived potential functions, provide one means for quantitative studies of interactions in many biological processes. However, the value of such studies has often been questioned due to considerable scepticism about the accuracy of empirically derived potential functions (see, for example, Roberts *et al*, 1986). In this article we show that the errors in potential functions should not be the major source of concern in computer simulations; there exist many other problems, and these can only be solved by a thorough investigation of computer simulation techniques. We also show that simpler and computationally less expensive simulation techniques can be used to good effect, and can be used to rationalize experimental data and to make useful predictions.

2 Theoretical Calculations

Statistical mechanics is the central discipline for analyzing the aggregate properties of a many particle system subject to some interaction potential V(r). The most important statistical thermodynamic quantity is the partition function Q_N:

1.

$$Q_N = \int e^{-E_N/kT} dr$$

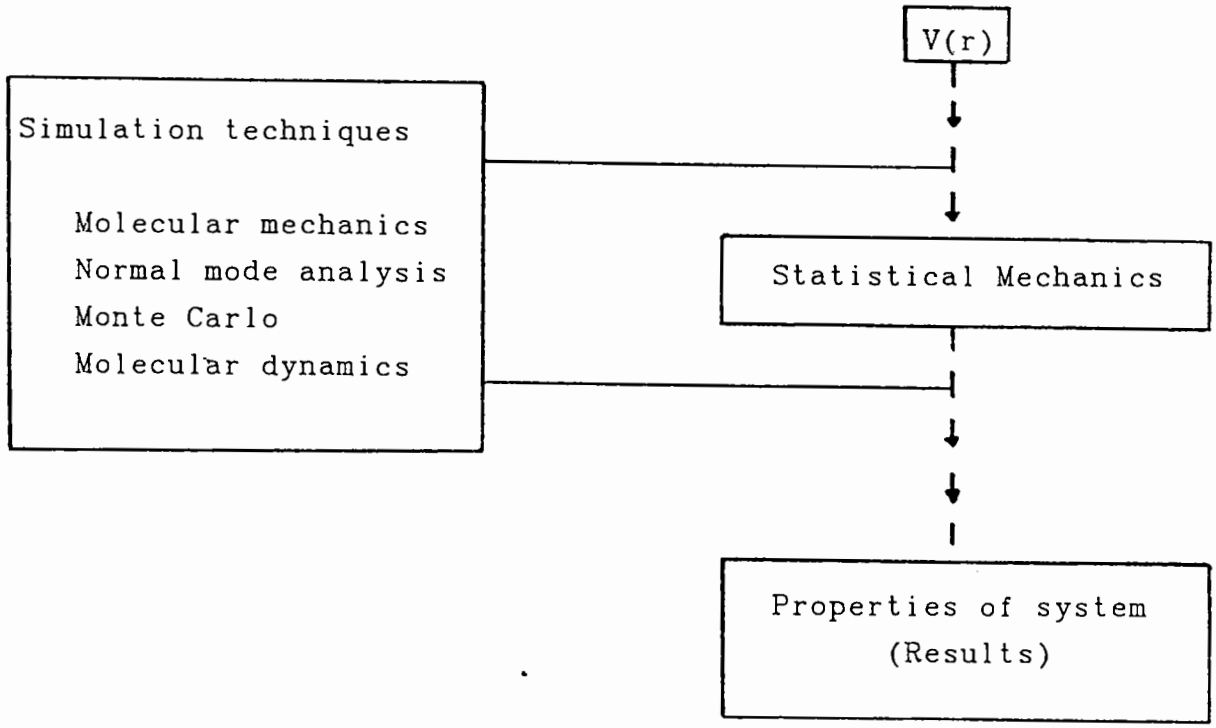
where E_N is the configurational energy, k the Boltzmann constant, and T the absolute temperature; the integration is carried out over whole of the configurational space of the N particle system. Other statistical thermodynamic quantities can be expressed in terms of the partition function using the well-known results of statistical mechanics (McQuarrie, 1976).

Central to the calculation of the properties of any given system is the knowledge of the interaction potential for that system. Thus, given a potential function V(r) for any system, statistical mechanical results provide a mechanism by which we can calculate the properties of that system. These properties can then be compared with experimental data. Any discrepency between the calculated and observed properties can be ascribed to errors in the potential function V(r).

Unfortunately, the picture of theoretical calculations presented so far is very limited. In practice, the direct calculation of the properties of a system from the potential function is not possible for realistic systems due to the very complex nature of statistical machanical expressions that relate V(r) to the properties of the system. Indeed, we find ourself in a situation similar to that exemplified by the famous statement of Dirac for quantum mechanical studies of molecules:

'The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble'.

Fig. 1



Lousy V(r) ==> Lousy results

Lousy results ==> Lousy V(r)

Statistical mechanical methods are essentially analytical and can only be applied in an exact form for extremely simple systems. In particular, for systems with very large number of degrees of freedom, the exact analytical approach is impossible and one has to resort to using numerical methods (computer simulations) to solve statistical mechanical equations (Fig. 1). The essential problem in computer simulations can be appreciated simply from the expression for the partition function. The integration over whole of the configurational space for a realistic many particle system is simply impractical. In computer simulations, therefore, one attempts to restrict the integration over those parts of configurational space which make the most significant contribution to the integrand. For most practical applications, this is done by starting the simulations from a configuration of the system that is reasonably close to the real system being studied. Even in such studies, the computational cost of such simulations is prohibitive and only a partial sampling of the configurational space is possible (typically $O(10^4)$-$O(10^6)$ configurations).

A number of computational techniques are used to derive the properties of a system from its potential function. Various simulation methods can often be grouped into one of the four techniques listed in Fig. 1. Of these, energy minimization (EM) and molecular dynamics (md) are by far the most commonly employed techniques for studying protein molecules.

1. Energy minimization (EM) In EM one searches the potential energy surface for configurations of the system at energy minima. One makes the assumption that the configuration at an energy minimum is significantly populated, and contributes significantly to the aggregate properties of the system. However, since the potential energy surface for complex systems is likely to have vary large number of minima, we are necessarily limited to sampling only a small fraction of these minima for large systems.

2. Molecular dynamics (md) Md simulations generate a trajectory (configurations as a function of time) of the system by numerically integrating Newton's equations of motion. The advantage of md over EM is that it provides a Boltzmann weighted ensemble of configurations for the system and, thus, permits the calculation of aggregate properties of the system. However, in order to obtain accurate trajectories it is necessary to employ (integration) time step of the order of femto-seconds. For typical protein molecules (*ca.* 2000 atoms), the computational cost effectively limits such md simulations to few tens of pico-seconds and at best to a few hundred pico-seconds.

The accuracy of the results obtained from statistical mechanics depends solely on the accuracy of the potential functions *i.e.* if one feeds in a lousy potential into statistical mechanics, one obtains lousy results; conversly, lousy results necessarily imply a lousy potential. This simple concept ('lousy results imply lousy potentials') plays an important role in theoretical calculations of macromolecules. For protein molecules, it provides the only means for designing accurate potential functions for these important molecules. (For small molecules, the potential functions can be obtained from detailed quantum mechanical studies; for macromolecules, however, empirical energy functions of the molecular mechanics (Boyd and Lipkowitz, 1982) type are the only possible source of such information.)

The results obtained from computer simulations are also affected by the accuracy of the potential functions; here, however, there exists another major problem - namely the inability of simulations to sample the configurational space adequately. The results from simulations will necessarily have some error due to incomplete sampling of configurational space. In this article we attempt to gauge the relative importance of the errors in computer simulations from a) the

incomplete sampling of the configurational space, and b) the errors due to inaccuracies in the potential functions.

3 Molecular Dynamics Studies

In computer simulations, structural properties of a system are likely to converge faster than the other properties (Mehrotra, *et al*, 1983) such as simple average quantities (*e.g.* mean energy) or fluctuation properties (*e.g.* heat capacity). However, there is a widely held view that simulations of *ca.* 10-100 pico-second are sufficient to give converged structural properties and that any discrepency between experiment and calculations is due largely to errors in potential functions (see, for example, van Gunsteren *et al*, 1983, and van Gunsteren, 1988). Can we set up simple calculations to test this hypothesis?

Consider an N particle system subject to some potential function V(r). If we were to perform m computer simulations of this system under identical conditions (*i.e.* at the same temperature, pressure, *etc.*) then the m converged structures should be identical (or at least the differences between them should be small). In table 1 are listed the rms differences between four molecular dynamics time averaged and experimental structures of avian pancreatic polypeptide (aPP, Blundell *et al*, 1981). The time averaged structures are from a 15 ps simulation of the unit cell of aPP with its full complement of solvent. The rms differences between the X-ray structure and the md structures range from 0.7Å-1.4Å for the main chain atoms, and 1.2Å-1.8Å for all atoms. Do these differences represent the errors in the potential functions? This would certainly be the case if the rms differences between the various md structures were to be, say, an order of magnitude smaller than the rms' between experimental and md structures. The pairwise rms differences between the four simulated structures of aPP range from 0.8Å-1.5Å for the main chain atoms, and 1.3Å-1.8Å for all atoms. These results clearly show that, contrary to popular view, *ca.* 10 ps md simulations are insufficient to give a converged time averaged structure for molecules the size of aPP.

Table 1. Rms differences (Å) between the X-ray structure (X) and the four time averaged structures from 15 ps md simulation of the full unit cell of aPP. Figures in upper triangle pertain to main chain atoms, those in lower triangle for all atoms. Taken from Haneef, 1985.

	X	1	2	3	4
X	-	1.0	1.4	0.7	0.8
1	1.4	-	1.4	1.0	1.0
2	1.8	1.7	-	1.5	1.4
3	1.4	1.4	1.8	-	0.8
4	1.2	1.3	1.8	1.4	-

We have carried out two independent 1000 ps simulations of a 19-mer RNA fragment from MS2 bacteriophage (Bernardi and Spahr, 1972). This structure has a stem-loop conformation with all but two of the bases involved in base-pairing. The structure is highly constrained due to the complementary base pairing and optimal base stacking. Such constrained structures provide ideal cases for studying the convergence properties of long md simulations. The simulations of this system were *in vacuo*, and used the potential functions of Weiner *et al* (1984). Both simulations were started from the same structure; the simulations differed only in that different initial velocities were assigned. The rms differences between the various time averaged structures from the two independent simulations are presented in table 2. The rms differences for 200 ps time averaged structures are typical of *in vacuo* simulations of nucleic acids. The interesting feature of these results is that the rms difference between 1000 ps time averaged structures from the two independent simulations is *ca.* 0.6Å, in excess of 30% smaller than corresponding rms differences between 200 ps time averaged structures; cleary the two simulations are converging towards the same average. However, even though both simulations were started from the same structure, and despite the constrained nature of the system, the simulations have not converged to an identical average structure after 1000 ps simulation. The conclusion from these results is unmistakable - md simulations in excess of 1000 ps are required to give a converged structure even for such a constrained system.

Table 2. Rms differences (Å) for 200 ps and 1000 ps time averaged structures of MS2 19-mer RNA from two independent md simulations. The two simulations were carried out *in vacuo* under identical conditions, but differed only in the initial velocities assigned at the start of the simulation. Rms differences are for all atoms.

Time span of structures	MD1	MD2	MD1/MD2
MD(1- 200)/MD(201- 400)	0.96	1.17	
MD(1- 200)/MD(401- 600)	1.02	1.32	
MD(1- 200)/MD(601- 800)	0.88	1.31	
MD(1- 200)/MD(801-1000)	1.01	1.55	
MD(1- 200)/MD(1- 200)			1.28
MD(201- 400)/MD(201- 400)			0.90
MD(401- 600)/MD(401- 600)			0.92
MD(601- 800)/MD(601- 800)			0.98
MD(801-1000)/MD(801-1000)			1.07
MD(1-1000)/MD(1-1000)			0.62

In table 3 are listed results from two independent 1000 ps simulations of the dimer of *E coli* methionine repressor (MetJ, Rafferty *et al*, 1988). These results follow essentially the same trends as those in table 2. Although the rms difference between the two 1000 ps time averaged structures is less than 20% smaller than the corresponding numbers for 200 ps time averaged structures, the difference is significant and shows that the two simulations are converging, albeit very slowly; for large systems, which inherently possess much lower frequency modes of vibrations, convergence to a true average structure is likely to require much longer simulations. Even

though both simulations were started from the experimentally determined structure, the 1000 ps time averaged structures differ by more than 1Å. Comparing the time averaged structures with the crystal structure shows that 1000 ps time averaged structures are significantly closer to the crystal structure than are the 200 ps time averaged structures. These results signal a clear warning about using rms differences between experimentally determined and simulated structures as a measure of the accuracy of the potential functions - such comparisons can only be warranted if the simulations have converged.

Table 3. Rms differences (Å) for 200 ps and 1000 ps time averaged structures of MetJ dimer from two independent md simulations, and comparison with X-ray (X) structure. The two simulations were carried out *in vacuo* under identical conditions, but differed only in the initial velocities assigned at the start of the simulation. Rms differences are for all atoms and main chain atoms (in brackets).

Time span of structure	X/MD1	X/MD2	MD1/MD2
MD(1- 200)	1.64(1.06)	1.72(1.10)	1.55(0.97)
MD(201- 400)	1.56(0.91)	1.76(1.10)	1.48(0.93)
MD(401- 600)	1.70(1.06)	1.75(1.10)	1.54(0.95)
MD(601- 800)	1.72(1.07)	1.75(1.11)	1.54(0.95)
MD(801-1000)	1.72(1.07)	1.73(1.07)	1.53(0.97)
MD(1-1000)	1.48(0.89)	1.69(1.05)	1.31(0.81)

4 Energy Minimization Studies

In studying the interactions between enzymes and substrates, the binding energy for the process

2. $E + S \longrightarrow E.S$

where E represents an enzyme, S the substrate and E.S the enzyme substrate complex is obtained from

3. $U(\text{binding}) = U(E.S) - U(E) - U(S)$

where U(X) represents the potential energy of the minimized structure X. Now consider the hypothetical case where there are no interactions between E and S; for this system, one would expect U(binding) and RMS(E,Ec) to be exactly zero (where E is the conformation of the isolated energy minimized E and Ec is conformation for E in the complex, and RMS represents the rms difference between two conformations). We have carried out a number of calculations for such a hypothetical binding of trimethoprim (TMP) to dihydrofolate reductase (DHFR); results from such calculations are presented in table 4. In all cases, the structures were restrained to their initial conformations using a harmonic restraining term. The restraining force constant was decreased from 100 kCal/mol/Å^2 to zero over 2000 steps. Further minimizations were carried out without the use of any restraining terms. The minimizations were stopped when all

components of the gradient vector had fallen below 0.01 kCal/mol/Å. The various calculations differ only in that we have used different search protocols in the conjugate gradient minimization routines.

Table 4. Total and binding energies for the hypothetical binding of trimethoprim to dihydrofolate reductase where there are no interaction between the two molecules. For each calculation, all molecules were minimized using identical protocols. The three calculation differ in that different search methods were used in conjugate gradient minimization (Fletcher and Reeves, 1966). In all calculations the minimizations were started from the same structures; the sarting structures of DHFR in isolation and in the 'complex' were also identical. Rms' are for all atoms and Cα atoms between isolated and complexed enzyme.

Calculation	U(DHFR.TMP)	U(DHFR)	U(TMP)	U(binding)	RMS
1	-5257	-5248	10	-19	0.39,0.34
2	-5245	-5268	10	13	0.33,0.26
3	-5259	-5267	10	-2	0.22,0.16

The connotations from the results in table 4 are clear. Consider the situation where we are interested in the difference in the binding energy of E to two similar substrates S and S'. If we were to use the method given above for calculating binding energy differences, we would obtain results in the range ± 40 kCal/mol - even if the two substrates did not interact with the enzyme! Comparing the result of ± 40 kCal/mol with the exact result of zero kCal/mol, the unwary would immediately jump to the conclusion that these lousy results were due to errors in the potential functions (no doubt due to leaving out an important electrostatic interaction of some sort!). The truth in this case is simple, and far more informative: the only difference in the minimizations of E on the left and right hand sides of equ. 2 is the number of degrees of freedom of the system. On the left hand side, minimization is carried out with N_E degrees of freedom; on the right hand side the total number of degrees of freedom is $N_E + N_S$, where N_E and N_S degrees of freedom are totally independent of each other. Indubitably, the total number of degrees of freedom has a profound effect on numerical minimization algorithms. In an exact analytical approach, the minimum obtained would depend solely on the starting structure of the molecule. Although the total energies of the enzyme and the complex, respectively, are very similar in magnitude in the different calculations, the rms differences in the resulting structures range from 0.5Å-0.8Å for Cα atoms and 0.7Å-1.1Å for all atoms (table 5). Results obtained from computer simulations are not identical to those expected from an exact analytical treatment - indeed, one would obtain different results on different computers due to the various precisions of these machines and the way they handle rounding errors, *etc.*

Simulations on such hypothetical systems are invaluable in providing insight into types of problems encountered in theoretical calculations. We have recently developed a minimization technique, SIMMIN, which is free from the problem described above. The essential feature of the technique is to minimize all three structures (*i.e.* E, S and E.S) in equ. 2 in the same simulation. The technique has considerable advantages over currently available methods for calculating binding energies; in particular, the technique delivers exact results for a number of hypothetical situations of the type described above. We have used this method to study the binding of avian egg-white lysozymes to the monoclonal antibody D1.3 Fab (Amit *et al*, 1986).

Table 5. Rms differences (Å) between crystal (X) and energy minimized isolated (E) and complexed (Ec) structures of DHFR from the three calculations referred to in text. Upper triangle, rms' for Cα atoms; lower triangle, rms' for all atoms.

	X	E(1)	E(2)	E(3)	Ec(1)	Ec(2)	Ec(3)
X	-	0.78	0.87	0.80	0.84	0.83	0.82
E(1)	0.99	-	0.83	0.76	0.34	0.77	0.77
E(2)	1.11	1.09	-	0.50	0.83	0.26	0.52
E(3)	1.04	0.97	0.73	-	0.80	0.49	0.16
Ec(1)	1.04	0.39	1.10	1.01	-	0.78	0.78
Ec(2)	1.06	1.04	0.33	0.72	1.06	-	0.51
Ec(3)	1.06	0.99	0.75	0.22	1.00	0.75	-

D1.3 binds hen lysozyme with an affinity constant of 5×10^7 M^{-1}, but if Gln 121 of lysozyme is replaced by His, complex formation is effectively abolished. We used the multi-start method to sample six energy minima in the vicinity of the crystal structure of the D1.3/hen lysozyme complex; pairwise rms differences between the six energy minimized structures range from 0.8Å-1.1Å for all atoms. All calculations, table 6, show that lysozymes with Gln at position 121 bind more tightly to D1.3 than lysozymes with His at position 121 - the energy difference being in the range 7-21 kCal/mol in favour of lysozymes with Gln at position 121. Since the energy minima obtained from multi-start methods represent a random sample, these calculations show that everywhere in the vicinity (*ca.* 1Å) of the crystal structure the potential energy surface of the Gln containing complexes is lower than that for His containing complexes. The binding energy differences obtained from these calculations compare favourably with experimental data which suggest that the difference in free energy of binding should be 5 kCal/mol or greater. Our results from the SIMMIN method should also be compared with those from currently available molecular mechanics methods which gave binding energy differences in the range $\pm$ 250 kCal/mol.

Table 6. Total and binding energies for Gln (GLN) and His (HIS) containing lysozyme/D1.3 Fab complexes. Figures in brackets are differences in energies. Energies in kCal/mol.

Model	GLN/HIS Total potential energy	GLN/HIS Binding energy
1	-11421/-11415 (5)	-139/-127 (12)
2	-11412/-11391 (21)	-127/-106 (21)
3	-11395/-11379 (16)	-159/-147 (12)
4	-11332/-11313 (19)	-165/-156 (9)
5	-11280/-11262 (18)	-131/-144 (17)
6	-11188/-11144 (44)	-89/ -82 (7)

5 Discussion

Indubitably, the results of computer simulations are affected by the quality of the potential energy functions employed. Empirically derived potential functions represent only an approximation to the true potential energy surface. However, it is difficult to obtain quantitative estimates for the errors involved. The only facts that are known concern the overall accuracy of the **computer simulations** and **not** the accuracy of potential functions. In this regard our views are almost diametrically opposed to those, for example, recently expressed by van Gunsteren (1988). The overall accuracy of a simulation depends only partly on the accuracy of the potential functions. Various approximations made to reduce the size of the actual model system employed, such as the use of united atom approximation and the exclusion of crystal and solvent environments, also make a difference. Considerable contribution to the accuracy of a simulation also comes from the limitations of the simulation techniques employed. Since computer simulation techniques are limited to sampling only a small fraction of the potential energy surface, such simulations cannot be used to attain the ergodic average and consequently cannot be used to assess the accuracy of the potential functions by comparing with experimentally determined structures (which are temporal and spatial averaged structures). Our results from a number of 1000 ps md simulations show that convergence to a true average would require several orders of magnitude longer simulations than are currently possible, even if such simulations are started from the experimentally determined structures. Our results also suggest that the converged structures would have smaller rms differences with experimentally determined structures. Thus to ascribe rms differences between *ca.* 10 ps time averaged and experimentally determined structures to errors in potential functions can only be warranted by **totally** ignoring the many limitations of computer simulation techniques.

While empirical potential functions may provide unreliable estimates for total conformational energy of macromolecules, differences due to the replacement of a single buried residue should be more dependable. This is indeed found to be the case - we have used a robust molecular mechanics technique to successfully establish the basis of much higher affinity of the antibody D1.3 for lysozymes with Gln at position 121 over lysozymes with His at 121. The same technique has been used to study the binding of two inhibitors, methotrexate and trimethoprim, to vertebrate (Mouse) and bacterial (*L Casei*) DHFR's. The results from this technique are very encouraging, and suggest that molecular mechanics can be used to study a wide range of molecular recognition processes, and that such studies can be used to make useful predictions about the effects of site-directed mutations, and also to determine the relative binding energies between a number of similar drugs to the same protein molecule.

The task of understanding many particle systems is far from trivial, even if the interactions between individual atoms were known exactly. The problem is not just one of carrying out complicated computations on bigger and better computers. The main aim should, instead, be to use one's knowledge of basic physical laws to develop new concepts which can illuminate the essential characteristics of such complex systems and thus provide sufficient insight to facilitate one's thinking, to recognize important relationships, and to make useful predictions. When the systems under consideration are not too complex and/or when the desired level of description is not too detailed, considerable progress can indeed be achieved by relatively simple methods.

Acknowledgements

Many of the ideas developed here have arisen through arguments over several years while the author was working at Birkbeck college (University of London) and more recently at the University of Leeds. I am grateful to my colleagues, past and present, for useful criticism, comments and stimulating discussions. My work and discussions with Anne-Marie Treharne are particularly acknowledged; her 1000 ps md simulation (A C Treharne, PhD Thesis, Uni of London) of nona-peptide, deamino-oxytocin, stimulated much of my current research. The responsibility for any errors or incorrect diagnoses rests solely with the author.

References

Amit A G, Mariuzza R A, Phillips S E V and Poljak R J, Science **233** (1986) 747
Bernardi A and Spahr P F, Proc Natl Acad Sci **69** (1972) 3033
Blundell T L, Pitts J E, Tickle I J, Wood S P and Wu S P, Proc Natl Acad Sci **78** (1981) 4175
Blundell T L and Sternberg M J E, Trends Biotech **3** (1985) 228
Boyd and Lipkowitz, J Chem Ed **59** (1982) 269
van Gunsteren W F, Berendsen H J C, Hermans J, Hol W G J and Postma J P M, Proc Natl Acad Sci **80** (1983) 4315
van Gunsteren W F, Protein Engineering **2** (1988) 5
Haneef I, PhD Thesis (1985), Uni of London
McQuarrie D A, 'Statistical Mechanics' (1976), Harper and Row, NY
Mehrotra P K, Mezei M and Beveridge D L, J Chem Phys **78** (1983) 3156
Rafferty J B, Phillips S E V, Rojas C, Boulot G, Saint-Girons I, Guillou Y and Cohen G N, J Mol Biol **200** (1988) 217
Roberts V A, Dauber-Osguthorp P, Osguthorp D J, Levin E and Hagler A T, Israel J of Chemistry **27** (1986) 198
Weiner S J, Kollman P A, Case D A, Singh C, Ghio C, Alagona G, Profeta S and Weiner P, J Am Chem Soc **106** (1984) 765
Winter G and Fersht A R, Trends Biotech **2** (1984) 115

Conformational Variability of Insulin: A Molecular Dynamics Study

Leo Caves
Department of Chemistry
University of York

1 Introduction

The three-dimensional atomic structure of insulin is amongst the most studied of all protein structures. X-ray crystallographic analysis has revealed many structures of the native molecule at high resolution (Derewenda et al, 1989), and has shown that the insulin molecule can adopt a number of similar, but distinct conformations (Dodson et al, 1980a, 1980b; Cutfield et al, 1981). Studies of modified insulins have indicated that the potential for conformational change may be important for the biological activity of the molecule (Dodson et al, 1983; Baker et al, 1988). Thus the observed conformational states of insulin in the crystal may be important to our understanding of how (and where) such conformational changes occur.

Aspects of the conformational variability of insulin are examined using the molecular dynamics (MD) simulation method. The questions addressed are the nature of the simulated dynamics of the insulin molecule (its intrinsic *flexibility*), and how these relate to experimental findings. In addition, the relationship of distinct insulin conformers both as observed in the crystal and as a result of simulation are examined. The study focuses on the structures of pig insulin as observed in three different crystal forms (2Zn, 4Zn and cubic - see Table 1). From these structures five conformational states of the molecule can be distinguished. Some of the conformers from different crystals are found to be very similar, others exhibit large differences (Cutfield et al, 1981). The differences between conformers have been attributed to the effects of crystal packing (Baker et al, 1988) or medium effects (Bentley et al, 1978).

In this study, the conformers are removed from their crystal environments and, through the molecular dynamics simulation method, allowed to traverse their intrinsic potential energy surfaces with kinetic energy appropriate to room temperature. From the resultant trajectories the dynamics of the conformers, both at the atomic and secondary structure level are analysed and compared. The results are related wherever possible to data derived from X-ray crystallography. Before presenting aspects of the analysis of the trajectories there is a brief description of how the molecular dynamics simulations were performed and of the methods used in their analysis.

2 Materials and Methods

2.1 Atomic Coordinates

The atomic coordinates used as the basis of this study are those of 2Zn, 4Zn and cubic pig insulin (see Table 1). These structures contain five distinct conformers of insulin (two each for 2Zn and 4Zn, one for cubic). As each distinct monomer and dimer from each form was considered, eight simulations were performed (three dimer, five monomer). This resulted in a total of eleven distinct simulated monomeric conformers to be considered (six from the dimer simulations, five from the monomer). All solvent molecules were excluded.

2.2 Computational Details

The simulations were performed using the CHARMM potential and molecular simulation package (Brooks et al, 1983). The *polar-hydrogen* representation was used as the molecular model (i.e.. only hydrogen atoms that are potentially capable

of participating in hydrogen bonding are included, the remainder being incorporated into *extended* heavy atoms). The hydrogen atom positions required to fulfil the representation were placed using the method of Brünger and Karplus (unpublished results). For the electrostatic terms of the potential, a distance-dependent dielectric constant was employed (where ε_r is set to be numerically equal to the distance between two atoms in Å), as a correction for the shielding effect of the neglected solvent (Weiner et al, 1984). The non-bonded interactions were truncated at 9.0Å, by means of a smoothing function acting between 8.5 and 9.0Å on an atom-pair basis (Brooks et al, 1983).

Table 1: Pig Insulin Crystal Structures used in this Study.

code	crystal system	level of assembly	assym. unit	space group	res (Å)	R (%)	solv (%)	comment	references
2ZX	rhombo-hedral	hexamer	dimer	R3	1.5	15.3	31	2Zn	Adams et al, 1969; Baker et al, 1988.
4ZX	rhombo-hedral	hexamer	dimer	R3	1.5	18.0	33	4Zn	Bentley et al, 1976; Derwenda et al, 1988.
CBX	cubic	dimer	monomer	$I2_13$	1.7	17.0	61	Zn-free	Dodson et al, 1978; Badger, 1986.

For a full list of the principal insulin structures see Derewenda et al (1989), Table 1.

2.3 Molecular Dynamics Simulation Protocol

The models resulting from the refinement of the crystallographic data were subjected to energy minimization prior to molecular dynamics simulation. A cursory run of 10 steps of the steepest descents method was performed on the structures (both isolated and associated), which resulted in backbone atom (N,Cα,C) root mean square differences in atomic positions (RMSD's) of less than 0.1Å. The MD simulations were performed at constant energy and volume, by solving the Newtonian equations of motion for the systems on the potential surface described by the CHARMM interaction function and parameters (Brooks et al, 1983). The Verlet algorithm (Verlet, 1967) was used with all bond lengths involving hydrogen atoms constrained with the SHAKE procedure (Ryckaert at al, 1977; van Gunsteren and Berendsen, 1977), allowing an integration timestep of 1×10^{-15}s. The simulations were performed in three phases: heating, equilibration and production. In the heating phase, velocities were assigned to the atoms from Gaussian distributions corresponding to successively increasing temperatures until reaching 300K. In the equilibration period the temperature of the system was monitored and if it exceeded a window of 10 degrees from 300K the atomic velocities were reassigned at 300K. This period lasted 20ps. In the production period, no temperature checking was performed and the integration proceeded uninterrupted up to 100ps. The simulations were performed on a CYBER-205 supercomputer at the John von Neuman Computer Centre at Princeton University. Execution times were of the order of 1 c.p.u hour per 10ps dynamics of the dimer.

2.4 Methods of Analysis

2.4.1 Measures of Structural Similarity

The index of conformation similarity used in this study is the r.m.s difference in the interatomic distance matrix of two structures (R_d):

$$R_d = \left(\frac{1}{N(N-1)} \sum_{j<i}^{N} (r_{ij}^{1} - r_{ij}^{2})^2 \right)^{0.5}$$

where N is the number of atoms
and r_{ij}^{1} is the interatomic distance between atoms i and j in structure 1.

In addition the components of R_d, that comprise a *difference-distance* matrix (D) were examined to identify the specific structural origins of R_d:

$$D = [r_{ij}^1 - r_{ij}^2]$$

2.4.2 Atomic Fluctuations

(a) Temporal Correlation of Fluctuations

The collective behaviour of atomic displacements was investigated in the trajectories. The method chosen was the temporal correlation of the displacement of the residue or backbone atom centroids (McCammon, 1984):

$$R_{ij} = \frac{\langle (r_i - \langle r_i \rangle)\,(r_j - \langle r_j \rangle) \rangle}{\langle (r_i - \langle r_i \rangle)^2 \rangle^{0.5}\, \langle (r_j - \langle r_j \rangle)^2 \rangle^{0.5}}$$

where: r_i is the instantaneous position of the centroid of the atoms of residue i
$\langle\rangle$ denotes time average.

This analysis yields a residue-residue correlation matrix which may be interpreted as follows. A high correlation coefficient (close to 1.0) infers that two residues are displaying collective character in their motions. A low correlation coefficient (close to zero) infers that there is little coupling in the motions of the residues.

(b) Variation in Interatomic Distances

In this study, the method of examining the intrinsic structural integrity of protein secondary structure uses a distance matrix approach (Havel et al, 1983; Elber and Karplus, 1987). The distances between particles i and j , r_{ij} , in a given conformation make up a matrix R. Of interest is the time averaged fluctuation of these distances in the simulation:

$$DD = \langle (R_t - \langle R \rangle)^2 \rangle^{0.5}$$

where $\langle\ \rangle$ denotes time average, and subscript t denotes value at time t.

This analysis yields a residue-residue variance (actually r.m.s) matrix which may be interpreted as follows. A low element (close to zero) means there is little variation in the specific interatomic (eg. Cα) distance, thus the structural element is relatively rigid. A high value means that there is a large variation in the interatomic distance and that the structural element is flexible in nature.

2.4.3 Defining a Structural Axis

In order to follow the relative orientation of structural elements (such as α-helices and β-sheets), one needs to define accurately their axes. In this study, the axes of regions were calculated by finding the principal components of the inertia tensor for the backbone (N,Cα,C) atoms, by diagonalization of the second moment matrix of the centre-of-mass atomic cartesian position vectors.

3 Results

3.1 Overall Structural Behaviour

In order to gauge the overall structural consequences of the simulations, the deviations of the structures from their initial (crystal) conformations were monitored throughout the trajectory. As an example, the plot of the time series of $R_d(t)$ for the 2Zn dimer simulation is shown in Figure 1. There is a clear difference in behaviour for the two constituent monomers when one examines the overall (both chains) deviations. At approximately 11ps there is a steep transition in the time series for molecule 2 that develops a difference over molecule 1 of 0.5Å (0.6Å ->1.1Å) over a period of approximately 6ps. The individual chain components of this

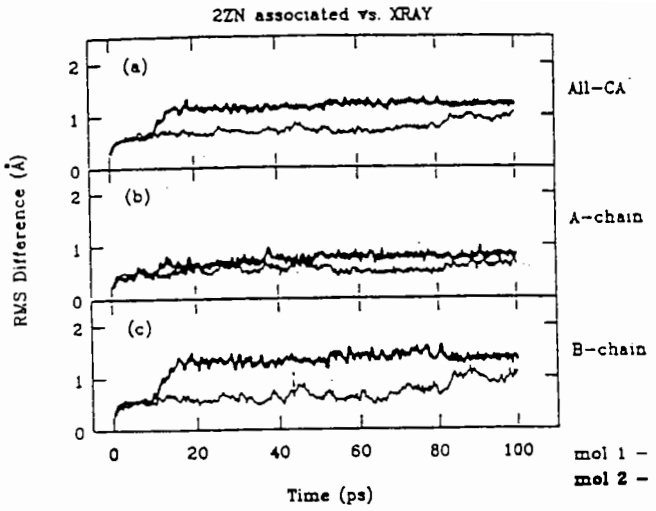


Figure 1

Deviation of individual molecules from their initial conformation as the trajectory proceeds. The measure used for comparison is the r.m.s difference in inter Cα distances. 2Zn trajectories: Associated species (a) All atoms, (b) A-chain, (c) B-chain.

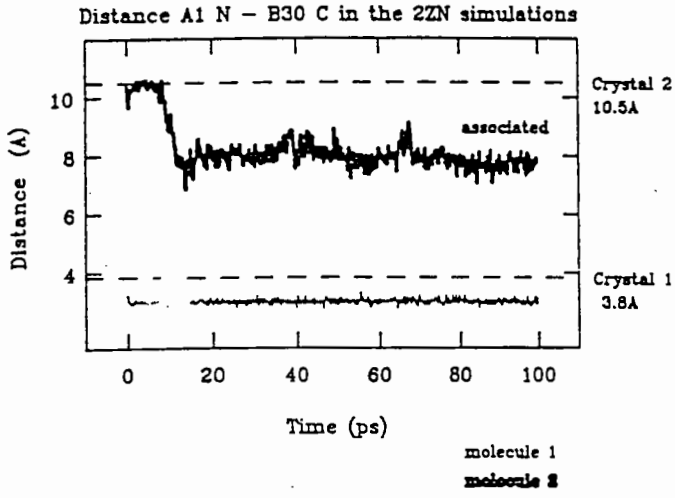


Figure 2

Time series of the distance between A1 peptide nitrogen and B30 carbonyl carbon 2Zn trajectories associated molecules, The distances found in the 2Zn crystal conformers of molecule 1 and 2 are indicated by dashed lines.

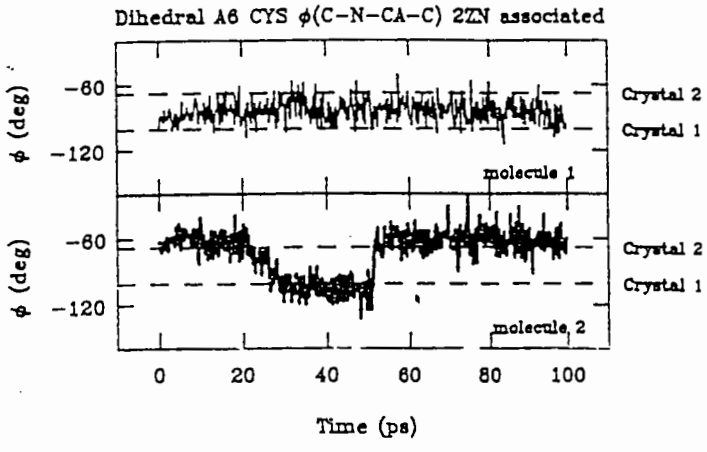


Figure 3

Time series of the peptide backbone dihedral φ, in A6 cystine in the 2Zn trajectories. (a) associated molecules, The angles found in the 2Zn crystal conformers are indicated by dashed lines.

value (Figures 1b and 1c), reveal that the change occurs predominantly in the B-chain. The Cα difference-distance matrix (D) of the 17ps vs 11ps transient structures for molecule 2 revealed the last three residues of the B-chain (B28, 29, 30) moving away from the C-terminal A-helix. Also evident was a closing up of the upper strand region of the B-chain (B25, 26, 27) against the B-helix. These changes are consistent with the closing up of the N-terminus of the A-chain and the C-terminus of the B-chain seen in molecule 2 between 7 and 11ps (Figure 2). This movement is illustrative of the salt-bridge formation between the A-chain N-terminal NH_3^+ -group and the B-chain C-terminal COO^- -group. Large positional deviations in the A-chain were traced to A5 glutamine, as a result of a change in the main chain dihedral ϕ of A6 cystine (Figure 3) between 13 and 17ps (which moves from approx. -65° to -115°). This change, which is subsequently reversed, is analogous to a change from the crystal conformation of molecule 2 (-66°) to molecule 1 (-101°) (Dodson et al, 1980b).

The analysis has revealed that the 2Zn *molecule 2* species undergoes greater deviations from its initial structure than does *molecule 1*. The specific nature of the structural changes relate to the major differences between the conformers observed in the crystal and have been observed in other simulation studies of 2Zn insulin (Wodak et al, 1984; Krüger et al 1987).

3.2 Atomic Fluctuations

3.2.1 Trends of Simulated Atomic Fluctuations versus those in the Crystal

The atomic fluctuations found in the simulations were compared to those in their respective initial crystal conformers (derived from crystallographic B-values). In general the agreement was poor (Table 2). It was expected that for a given initial crystal conformer, the agreement with the simulation results would be better for the simulations of the conformers as associated species, rather than as isolated species (the dimer association state is closer to that found in the crystals). In the trends of the residue-averaged backbone atomic fluctuations this is not generally found, however there are such cases, such as in the A-chain of 2Zn *molecules* 1 and 2. Interestingly, there is better agreement of the cubic conformer with its crystal trends in simulation as an isolated species rather than as a dimer, despite the crystal form being dimeric.

Table 2: Correlation of Atomic Fluctuations in the Simulations with those of the Initial Crystal Conformer.
Residue averaged backbone results.

	Correlation Coefficient			
	A-chain (n=21)		B-chain (n=30)	
	Simulation State		Simulation State	
CRYSTAL CONFORMER	isolated	associated	isolated	associated
2Zn mol 1	0.14	0.51	0.67	-0.01
2Zn mol 2	-0.44	0.24	0.35	0.19
4Zn mol 1	0.05	0.30	0.33	0.56
4Zn mol 2	0.47	-0.59	0.60	-0.47
Cubic monomer	0.80	-0.03/-0.13	0.67	0.72/-0.22

3.2.2 Collective Nature of the Atomic Fluctuations

The backbone (N,Cα,C) centroid displacement cross-correlations in the 2Zn dimer trajectory are shown in Figure 4 as a two-dimensional contour plot. The range of correlation coefficients is 0.72 -> 1.0, indicating that the motion of the

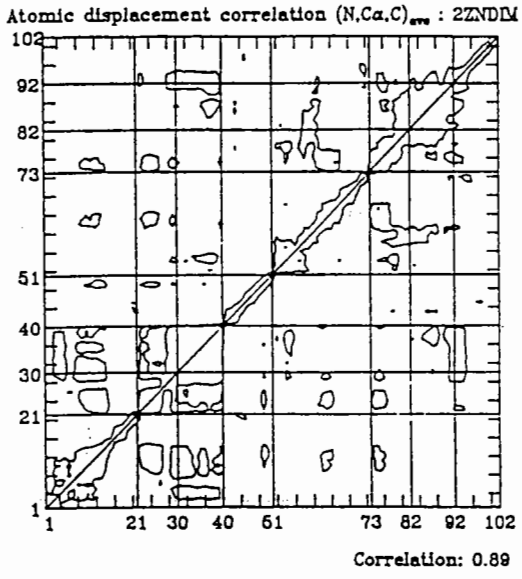


Figure 4

Representation of the collective motion in the 2Zn dimer simulation via the cross-correlation of backbone (N,Cα,C) centroid displacements. The range of absolute correlation coefficients is 0.7 to 1.0. A single contour is drawn at a value of 0.89 which corresponds to one standard deviation above the mean. Larger correlation occur along the diagonal and within the closed off regions off the diagonal. The correlations are calculated over the period 20-100ps.
Residue numbering: Molecule 1, A: 1 -21, B: 22 - 51; Molecule 2, A: 52 - 72, B: 73- 102.

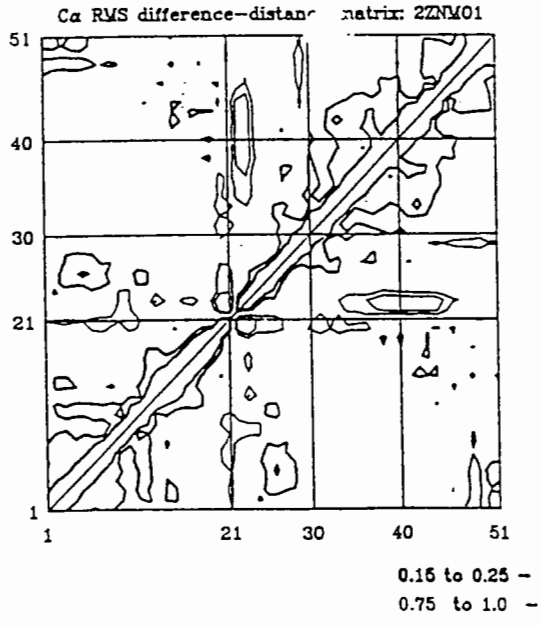


Figure 5

The r.m.s inter-Cα difference-distance matrix calculated from the 2Zn molecule 1 isolated trajectory over the period 20-100ps. Low variation occurs mainly along the diagonal and is more extensive in helical regions. High variation is found to occur at loop regions and chain termini. Contours are drawn at 0.15 and 0.25Å (thick lines) and 0.75 and 1.0Å (thin lines).
Residue numbering: A-chain: 1 - 21, B-chain: 22 - 51.

whole molecule is highly coupled. Higher correlations occur along the diagonal, reflecting coupling between adjacent neighbours. High correlations are also found to occur in off-diagonal positions, reflecting coupling with non-adjacent or non-contiguous residues. More extensive regions with high correlations are found to occur, for example in region 29 - 39 (B18 - B29) - the molecule 1 B-helix. This reflects the collective character of the motion of this helix.

These results may be compared with the findings of an examination of the thermal diffuse X-ray scattering (T.D.S) from insulin crystals (Caspar et al, 1988). T.D.S can give information on coupled correlated motion in crystals (eg. Doucet and Benoit, 1987). In the study by Caspar et al, the diffuse scattering was considered to arise from two main components, a variational term from internal protein motions and a lattice term from the rigid-body displacements of neighbouring molecules in the crystal lattice. The diffuse scattering was modeled by the use of two parameters, δ, the average magnitude of fluctuation, and γ the mean coupling distance of atoms, which can be independently estimated from experimental measurements. Caspar et al found that the variational scattering was best described by atomic fluctuations of the order of 0.69-0.78Å coupled over an average distance of 4-8Å. The lattice term was found to have fluctuations of approximately 0.25Å over a distance of 20-30Å. Thus it was found that the intramolecular atomic displacements account for 60-80% of the overall observed displacements.

Data from the simulation of the 2Zn dimer may be compared to this experimental study. Overall (average of both molecules and both chains), it was found that the magnitude of the atomic fluctuations in the simulations accounted for 70-75% of those found in the crystal. From the analysis of the temporal correlation of the centroid fluctuations, the average distance over which residues are coupled is estimated to be 6-10Å. The the agreement with experiment is encouraging.

3.3 Secondary Structure Dynamics

3.3.1 Characterisation of the Dynamic Behaviour of Secondary Structure

The integrity of the secondary structure was examined by monitoring the variation of interatomic distances through the trajectories. The DD matrix (the r.m.s difference-distance matrix) for the Cα atoms was calculated from each of the simulations over the period 20-100ps at a time resolution of 0.1ps In Figure 5 the DD matrix for the isolated 2Zn molecule 1 simulation is shown as a two-dimensional contour plot. Along the leading diagonal are found the self-self (zero) terms. As expected, low fluctuations (<0.2Å) are found for residues adjacent in the sequence (commonly extending up to 2 residues either side), reflecting the constraints of covalent linkage. These regions manifest themselves as low contours about the leading diagonal. There are also extensive contiguous regions of secondary structure that may be regarded as semi-rigid in nature. These regions include the observed secondary structure elements and reflect the spatial restraints of the interactions that give rise to them. Non-contiguous regions may also be constrained. These may be explained via covalent (S-S bridge) or non-covalent (hydrogen-bond) interactions. Flexible regions are revealed clearly, and tend to display flexibility towards the whole of the rest of the molecule.

3.3.2 Relative Motion of Secondary Structure Regions

Certain regions were selected to serve as probes of the internal dynamics of the molecule via the monitoring of their relative motion. The regions selected were AN,AC,BH,BS (as defined in Figure 6) which coincide approximately with the N-terminal A-helix, C-terminal A-helix, B-helix and C-terminal B-chain strand respectively. The time series of the relative displacements of the centroids of the four selected structural regions show r.m.s fluctuations of 0.2-0.3Å and ranges of 1.0-1.5Å. Statistics for the time series of the relative orientations of the structural regions reveal that the motions typically have r.m.s fluctuations of 2-3° with ranges of 15-20°. An example of the time series for the BH,AC displacement and orientational motion for the cubic isolated simulation is given in Figure 7.

CODE	TYPE	REGION	LENGTH (residues)
AN	helix	A2-9	8
AC	helix	A13-19	7
BH	helix	B12-20	9
BS	strand	B24-28	5

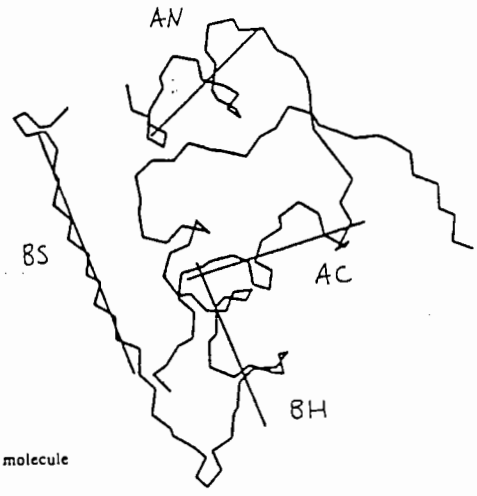


Figure 6

Axes defined for selected secondary structure regions. Illustrated for the 2Zn molecule 1 crystal conformer.

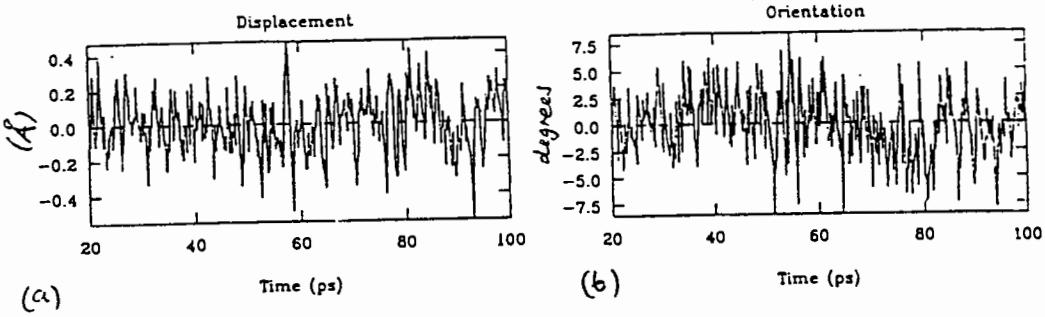


Figure 7

The relative motion of the BH and AC regions (defined in Fig 6) Time series of their relative (a) displacement, (b) orientation in the cubic isolated trajectory over the period 20-100ps. (Fluctuations from mean.)

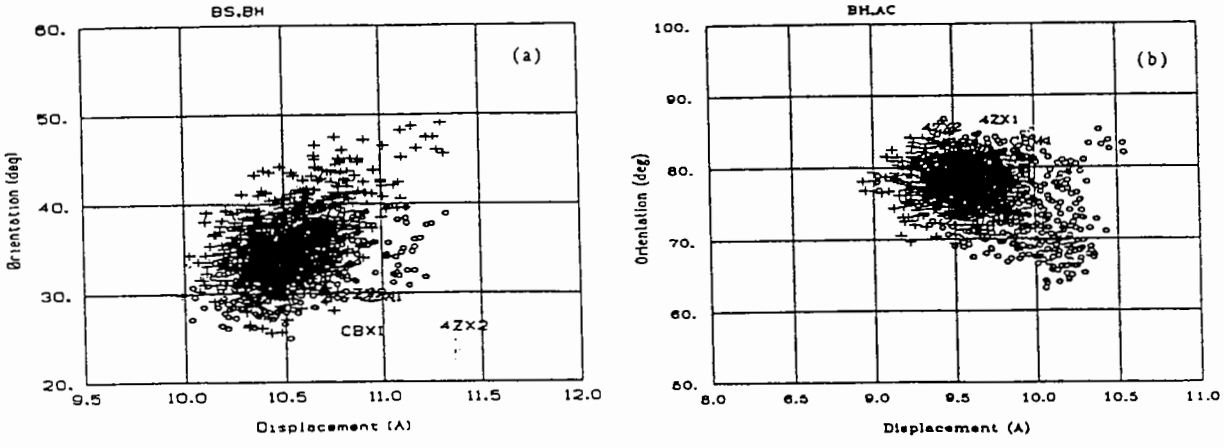


Figure 8

Relative motions of selected secondary regions. The instantaneous orientation versus displacement is plotted throughout the 20-100ps regions of the trajectories. Data from the cubic isolated (symbol: o) and 2Zn molecule 1 isolated (symbol: +) trajectories over the period 20-100ps.

(a) BS,BH motions

(b) BH,AC motions

The relative orientations and displacements of these regions in the five crystal conformers are shown by their code (Table 1). If the code is not visible it is because the trajectories have encompassed the corresponding packing of the regions in the crystal. The plots show the extensive overlap of the secondary structure packing between individual trajectories, as well as the overlap of individual trajectories with the range of crystal conformations.

It was found that there is a high degree of overlap of the secondary structure motions between the trajectories (see Figure 8). Thus the molecular dynamics are sampling very similar regions of the available secondary structure packing space. Further the range of motions found in one trajectory were found to encompass the range of secondary structure packing observed in the different crystal structures of pig insulin. An exception is that the dynamics of *molecule 1* conformers does not extend to motions such as those observed between the A and B-chain in the 4Zn molecule 2 crystal conformer. This can be contrasted with the study of the dynamics of myoglobin by Elber and Karplus (1987) .

4 Summary

Molecular dynamics simulations of distinct conformers of pig insulin have been performed. The resulting trajectories have been analysed in terms of atomic fluctuations and secondary structure dynamics. The main points emerging from the analyses are now given

• Overall structural behaviour in the simulations revealed that prominent changes occur to conformational features in the initial structures that are considered to result from specific crystal packing of medium effects. Examples are the salt-bridge formation between chain termini in 2Zn *molecule 2* and the contraction of the A and B-chain separation in 4Zn *molecule 2*. Structural regions without specific conformation defining interactions (such as the N-terminus of the B-chain) were also found to differ from their initial conformation as a result of simulation.

• The magnitude and spatial extent of correlated fluctuations in the trajectories is in agreement with results inferred from a study of X-ray thermal diffuse scattering (T.D.S) data on 2Zn insulin (Caspar et al, 1988). Analysis of the simulations revealed short range (4-6Å) collective behaviour occurs mainly between adjacent residues. More extensive regions (10-20Å) with a collective character were found to coincide with (or occur at the interface between) defined secondary structure regions (such as the helices).

• An analysis of the relative motion of certain structural regions (within the rigid-body limit) allowed a mapping of the overall motions of the molecule in terms of secondary structure dynamics. The results revealed that there is a large degree of conformation overlap between the distinct conformers in terms of secondary structure packing. The range of motions (displacements, orientations) of a given conformer, in a 100ps room temperature *in vacuo* trajectory, is found to encompass the range of packing observed in the different crystal forms. However conformational changes of the magnitude of the difference of 4Zn *molecule 2* from the rest of the conformers (which involve large displacements in the relative position of A and B-chains) are found not to be sampled in the trajectories.

Acknowledgements

Thanks to Prof. Martin Karplus and Dr. Dzung Nguyen for their help in generating the trajectories and to Prof. Guy Dodson and Dr. Rod Hubbard for helpful discussions.

References

Adams, M.J., Blundell, T.L., Dodson, E.J., Dodson, G.G., Vijayan, M., Baker, E.N., Harding, M.M., Hodgkin, D.C., Rimmer, R. and Sheet, S. *Nature* **224** (1969) 491-496.
Badger, J. Ph.D. Thesis (1986) University of York.

Baker, E.N., Blundell, T.L., Cutfiled, J.F., Cutfield, S.M., Dodson, E.J., Dodson, G.G, Crowfoot Hodgkin, D.M., Hubbard, R.E., Isaacs, N.W., Reynolds, C.D., Sakabe, K., Sakabe, N. and Vijayan, N.M.*Philos. Trans. R. Soc. Lond [Biol]* **319** (1988) 369-456.
Bentley, G.A., Dodson, E.J., Dodson, G.G., Hodgkin, D.C. and Mercola, D.A. *Nature* **261** (1976) 166-168.
Bentley, G.A., Dodson, G.G. and Lewitova, A.*J. Mol. Biol* **126** (1978) 871-875.
Brooks, B.R., Bruccoleri, R.E., Olafson, B., Stales, D.J., Swaminathan, S. and Karplus, M. (1983) *J. Comput. Chem.* **4**(2), 187-217.
Caspar, D.L.D., Clarage, J., Saluke, D.M. and Clarage, M. (1988) *Nature* **332**, 659-62.
Cutfield, J.F., Cutfield, S.M., Dodson, E.J, Dodson, G.G, Reynolds, C.D. and Valley, D. in *Structural Studies on Molecules of Biological Interest* ed Dodson, G.G., Glusker, J.P. and Sayre, D. (1981) Oxford University Press.
Derewenda, U., Derewenda, Z. and Dodson, G.G. (1988) manuscript in preparation.
Derewenda, U., Derewenda, Z., Dodson, G.G., Hubbard, R.E. and Korber, F. *British Medical Bulletin* **45(1)** (1989) 4-18.
Dodson, E.J., Dodson, G.G. and Hodgkin, D.C. in *Frontiers of Bio-organic Chemistry and Molecular Biology* ed Ananchenko, S.N. (1980a) Pergamon Press.
Dodson, E.J., Dodson, G.G., Lewitova, A., Sabesan, M. *J. Mol. Biol.* **125** (1978) 387-396.
Dodson, E.J., Dodson, G.G., Reynolds, C.D. and Valley, D. in *Insulin, Chemistry, Structure and Function of Insulin and Related Hormones* eds Brandenburg, D. and Wollmer, A. (1980b) de Gruyter, New York.
Dodson, G.G., Hubbard, R.E. and Reynolds, C.D. *Biopolymers* **22** (1983) 281-292.
Doucet, J. and Benoit, J.P. *Nature* **325** (1987) 643-646.
Elber, R. and Karplus, M. *Science* **235** (1987) 318-21.
Havel, T.F., Kuntz, I.D. and Crippen, G.M. *Bull. Math. Biol.* **45** (1983) 665.
Krüger, P., Straßburger, W., Wollmer, A., van Gunsteren, W.F. and Dodson, G.G. *Eur. Biophys. J.* **14** (1987) 449-59.
McCammon, J.A. *Rep. Prog. Physics* **47** (1984) 1-46.
Ryckaert, J.P., Ciccotti, G. and Berendsen, H.J.C. *J. Comput. Phys* **23** (1977) 327-341.
van Gunsteren, W.F. and Berendsen, H.J.C. *Mol. Phys.* **34**(5) (1977) 1311-1327.
Verlet, L. *Phys. Rev.* **159** (1967) 98.
Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S. and Weiner, P. *J. Am. Chem. Soc.* **106** (1984) 765-784.
Wodak, S.J., Alerd, P., Delhause, P. and Renneboog-Squilbin, C. *J. Mol. Biol.* **181** (1984) 317-22.

Structure Determination from NMR Conformational Data by Molecular Dynamics Calculations

Michael Nilges, Angela M. Gronenborn and G. Marius Clore

Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA

Introduction

NMR spectroscopy has been used successfully to solve three-dimensional structures of proteins (see Wüthrich, 1986; Clore & Gronenborn, 1987; for reviews). Both model calculations and comparisons with X-ray structures (e.g. Kline et. al., 1986; Clore et. al., 1987a,b) indicate that protein structures are well determined by the experimental NMR data in the form of approximate distances between protons close to each other in space (< 5Å), and torsion angles (ϕ, χ_1). In order to convert the measured distances and angles into three-dimensional structures they are combined with the known stereochemistry (bond lengths, angles, planarity, chirality) and packing restraints (van der Waals radii) into a target function; solving the structure amounts to locating the global minimum region of this target function. One can distinguish two aspects in this problem: structure determination involves finding the global conformation (i.e. the correct fold of the polypeptide chain, in the case of proteins), while the local conformation is then improved by refinement. This is illustrated in fig. 1, where in a hypothetical energy surface of a protein the three large minima represent different folds of the chain, while the sub-minima correspond to different local conformations.

A variety of mathematical tools has been developed for determination and refinement of structures with NMR data (see Braun, 1987; for a review), such as metric matrix distance geometry (DG) and restrained minimization in torsion angle space. Restrained molecular dynamics was shown to be a powerful tool for refinement of structures generated with other methods (Kaptein et. al., 1985), and for structure determination itself, as was demonstrated for peptides (Clore et. al., 1985), nucleic acids (Nilsson et. al., 1986), and with a folding strategy also for small proteins (Brünger et. al., 1986; Clore et. al. 1986a). The basic idea behind the use of MD as a minimization technique is that the kinetic energy of the atoms at a high temperature allows to escape local minima in which the structure would easily be trapped with conventional minimization techniques (figure 1). The methodology was extended to crystallographic refinement by Brünger et. al. (1987) who used RD with very high temperatures to overcome even very high energy barriers. Clearly, these methods are closely related to simulated annealing (Kirkpatrick et. al., 1983), which uses the Metropolis algorithm (Metropolis et. al., 1953) to raise the "temperature" in a system that is then minimized by cooling it slowly.

In this paper, we do not try to give an overview over the whole field of RD. We concentrate on efforts to improve the efficiency and power of the method, by simplifying the force field employed in the structure determination phase, and using novel dynamics protocols which take the analogy to simulated annealing more seriously. All dynamics calculations were carried out with the molecular dynamics program X-PLOR (Brünger, 1988).

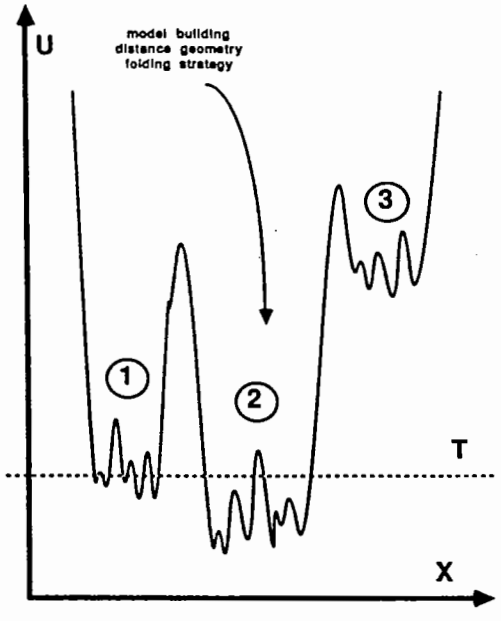


Figure 1: Hypothetical potential energy surface (U) of a protein. The three large minima (numbered 1, 2, 3) indicate different folds of the polypeptide chain, the subminima different local conformations. Generally, RD and related methods try to improve a structure generated with the help of a different method locally by raising the temperature: Potential energy barriers can be overcome due to the kinetic energy (T) (see text).

The Target Function

The target function consists of terms which contain the *a priori* knowledge of the molecule and are derived from the MD empirical energy function, and terms comprising the experimental information:

$$F_{total} = F_{empirical} + F_{experiment}$$

The form of $F_{empirical}$ used in the calculations described here reflects the fact that the aim is not to obtain dynamical information on the protein but to find a conformation which satisfies the geometric restraints imposed by the stereochemistry and packing requirements. For angle, bond, planarity and chirality restraints the standard analytical form of the X-PLOR force field is used (which is harmonic in the deviation from their ideal values) but the force constants are all set to uniform values:

$$F_{covalent} = k_{bond} \sum_{bonds} (b - b_o)^2 + k_\theta \sum_{angles} (\theta - \theta_o)^2 + k_\phi \sum_{impropers} (\phi - \phi_o)^2$$

The exact values of these force constants do not influence the structure greately (as long as they are large enough). Additionally, all dihedral angle potentials are removed. More important, some or all of the force constants (or better, weight factors) may be varied; their values in the final stages of the calculations are set to 500 kcal mol^{-1} $Å^{-2}$ for the bonds and 500 kcal mol^{-1} $radian^{-2}$ for the angle terms.

Modifications of the non-bonded energy terms are crucial for improving efficiency. The electrostatic interaction is removed, and the van der Waals (vdW) interaction is represented by a simple quartic

repulsion term similar to the one employed in PROLSQ (Konnert & Hendrickson, 1980):

$$F_{nb} = k_{repel} (r_{min}^2 - r^2)^2$$

where r is the actual distance between two atoms, r_{min} the separation at the minimum of the standard Lennard-Jones potential multiplied by a scaling factor s, and k_{repel} the force constant for the non-bonded interaction. This vdW repulsion term is finite at r=0 and thus allows atoms to pass each other. Due to the lack of an attractive component in this form of non-bonded interaction the vdW radii have to be reduced. For this reason, the scaling factor s is set to 0.8 in the final stages of dynamics and minimization; the resulting effective vdW radii are similar to those used in the various distance geometry programs. The final value of k_{repel} is set to 4 kcal mol^{-1} $Å^{-4}$ as this was found to be sufficient to ensure that no unduly close nonbonded contacts occur in the final structures.

The second part of the target function, $F_{experiment}$, introduces the experimental data. The principal source of information is the nuclear Overhauser effect (NOE) (Noggle & Schirmer, 1971). The initial buildup rate of the NOE is approximately proportional to $\tau_c r^{-6}$, where τ_c is the correlation time which depends on the overall tumbling rate of the molecule in the solvent but also on local mobility. The precision with which the distances can be determined is affected by two factors: τ_c is not uniform and generally not known, and the NOE intensity measured for a certain proton pair can be influenced by indirect NOEs involving other neighbouring protons. In a conservative analysis, one would therefore classify the distances only roughly into a few classes depending on the intensity of the NOE: Strong, medium and weak NOEs correspond to distances between 1.8 and 2.7 Å, 1.8 and 3.3 Å, and 1.8 and 5.0 Å, respectively. This particular classification does not introduce any lower bound estimates on the distances other than those given by the sum of the vdW radii. The distance restraints are incorporated into the target function in the form of a square well potential with harmonic walls (Clore et. al., 1986b):

$$F_{noe} = k_{noe} \begin{cases} (r - r^l)^2 & \text{if } r < r^l \\ 0 & \text{if } r^l \le r \le r^u \\ (r - r^u)^2 & \text{if } r > r^u \end{cases}$$

where r^l and r^u are the lower and upper limits of the distance, and k_{noe} the force constant. The uncertainty in the distance is directly reflected in the potential form, in contrast to a simple harmonic potential (Kaptein et. al., 1985) where the force constant is adjusted in such a way that the NOE contribution to the energy is 1/2 k_bT at r^l or r^u (k_b is the Boltzmann constant). Since in a square well potential the value of F_{noe} at r^l or r^u is zero, the force constant can be substantially larger than in a harmonic potential. Like the other force constants, k_{noe} can be varied during the calculation; the final value is set to 50 kcal $mol^{-1}Å^{-2}$ which is ten times smaller than the force constants employed for the terms determining the stereochemistry.

In addition to the distances, torsion angle restraints for ϕ and χ_1 can be obtained from coupling constant measurements via Karplus relations (see Wüthrich, 1986, and original literature cited there); these are included into the target function also in the form of a square-well potential in angle terms.

The final form of the "force field" resembles the target functions in various distance geometry programs. Independently of the work described here, target functions (with fixed force constants) incorporating geometric restraints rather than energy terms were used in MD calculations by other groups to refine structures generated with a DG algorithm (Nerdal et. al., 1988) and to improve the sampling properties of DG (Scheek and Kaptein, 1988).

Annealing Dynamics With a Variable Target Function

Similar to simulated annealing, the target function is minimized by heating the system to a high temperature and cooling it slowly. This is achieved here by solving Newton's equations numerically for all n atoms using a MD algorithm:

$$m\ddot{\underline{r}}_i = -\underline{\nabla}_i F_{total} .$$

To increase the stability of the calculations, the atom masses have been set to a uniform value here (10 a.m.u.) as they do not contribute at all to the conformation which satisfies the geometric restraints of the target function. The temperature is related to the kinetic energy of the system:

$$T = \frac{2}{k_b (3n - 6)} \sum_i (m v_i^2 / 2)$$

with the Boltzmann constant k_b and velocities v. As pointed out before, the target function is not constant during the calculation; the way it is varied is described in more detail below and elsewhere (Nilges et. al., 1988 abc). Common features of the minimization schemes are the following:

1) initial conjugent gradient minimization;
2) high temperature stage (1000 - 1200 K)
   variable target function (potential form and/ or force constants);
3) cooling stage;
4) final conjugate gradient minimization.

The temperature is controlled either by simple velocity scaling or by coupling to a heat bath (Berendsen et. al., 1984). It should also be noted, however, that varying force constants in the target function has a similar effect to changing the temperature directly. Heating the system corresponds to reducing the force constants. Instead of raising the kinetic energy, the potential energy barriers are reduced. Cooling, on the other hand, can be achieved by slowly increasing the force constants. Thus, several ways of varying the annealing "temperature" can be distinguished. It can be changed directly by scaling the velocities, the relative weights of the different terms in the target function can be altered, or the potential energy can be scaled by varying all the force constants in the target function at the same time. The first of the two annealing schedules described below uses a combination of the first two methods, while the second relies mainly on scaling the potential energy.

Efficient Refinement of Bad Initial Structures

The aim of the minimization scheme described in this paragraph is to calculate protein structures from NMR data in a very CPU efficient way. The dynamics program is used for the refinement of a crude starting structure obtained with a DG algorithm (DISGEO, Havel 1986). DG algorithms obtain the coordinates by a projection from the space of distances between all pairs of atoms to three-dimensional Euclidean space (Crippen & Havel 1978). The general protein fold can be obtained reliably as there is no danger of the polypeptide chain getting "entangled" during a minimization and thus trapped in a deep false minimum. The projection, called "embedding", exploits the fact that the three-dimensional coordinates are the eigenvectors of the metric matrix, which can be calculated directly from the distances. The theory requires exact distances between all pairs of atoms. However, in reality only distance ranges are known. The algorithm selects distances randomly within these ranges. The distances generated in this way are usually not consistent with a structure in three dimensions. As a consequence the embedded structures have a large number of violations of the distance bounds, even of those which are well known, like bond lengths. DG structures therefore always have to be refined. This is usually done by conjugent gradient minimization. The resulting structures still exhibit a number of problems: 1) There usually remain

distance bound violations; 2) the quality of the structures (in terms of non-bonded contacts and stereochemistry) depends on the amount of experimental data; 3) the sampling of the allowed conformational space consistent with the data is relatively poor and 4) depending on the algorithm used, DG can be CPU expensive.

For these reasons, we have chosen DG only to locate the correct protein fold in a "hybrid" approach. Instead of embedding all atoms present in the molecule, the problem is reduced in size by embedding only about 1/3 of the atoms (using a built-in feature of DISGEO). Such a "substructure" can be generated in a few minutes on a μVax III computer for a small protein. The remaining atoms are added arbitrarily, that is, without taking non-bonded contacts or experimental restraints into account [Holak et. al. (1988) use similar starting structures, employing energy minimization rather than dynamics for refinement]. This very crude starting structure is then refined as follows. After some initial unrestrained conjugate gradient minimization which serves to improve the covalent structure it is heated to 1000 K. During the approximately 4 ps dynamics at 1000 K the force constants for vdW and experimental restraints are increased from their low starting values to the final values mentioned above. The stereochemistry terms are kept at their high values throughout the calculation. The structure is then cooled down to 300 K and minimized. A part of the X-PLOR input file is shown below.

```
!-------------------------------------------------------
! annealing dynamics at 1000 K, variable target function

evaluate ($1 = 0.001)                    ! initial vdw force constant
evaluate ($3 = 0.5)                      ! initial NOE force constant

while ($1 < 0.25) loop anneal

  evaluate ($1 = min($1*1.125, 0.25))    ! increase vdw constant to 0.25
  parameters nbonds
    repel     = 1.0                      ! full L-J radii (s = 1.0)
    rconstant = $1                       ! set force constant to current value
  end  end

  evaluate ($3 = min ($3*2.0, 50.0))     ! increase NOE constant to 50
  noe
    sqconstant NOE $3                    ! set force constant to current value
  end

  dynamics verlet
    nstep     = 75                       ! 75 steps dynamics
    timestep  = 0.001                    ! with time step 1 fs
    iasvel    = current                  ! initial velocities from prev. cycle
    tcoupling = true                     ! Berendsen method
    tbath     = 1000                     ! at 1000 K
    nprint    = 75    iprfrq    = 75     ! output statistics at end of cycle
  end

end loop ( $1 ) anneal

!-------------------
! cool down to 300 K

parameter  nbonds
    repel     = 0.8                      ! reduced radii (s = 0.8)
    rconstant = 4.0                      ! final force constant
end  end

evaluate ($2 = 1000.0)                   ! starting temperature

while ($2 > 300.0) loop cool
  evaluate ($2 = $2 - 25.0)              ! reduce temperature
  dynamics verlet
    nstep     = 50    timestep  = 0.001
    iasvel    = current
    tcoupling = true                     ! Berendsen method
    tbath     = $2                       ! at current temperature
    nprint    = 50    iprfrq    = 50
  end
end loop ( $2 ) cool

!-----------------------
! Final energy minimizing

minimize powell
    nstep = 200
    drop  = 10.0
end
```

Structures calculated with this approach exhibit none of the problems associated with DG mentioned above. Test calculations were carried out for Crambin, using a reasonable set of model NOE distance restraints derived from the crystal structure (Hendrickson & Teeter, 1981), and the globular domain of Histone H5, which is almost twice the size of Crambin and has relatively few distance restraints (Clore et. al., 1987c). The sampling is improved with respect to structures calculated with DG alone, especially in the case with fewer NOE distances. The quality of the structures is nearly "regularized", independently of the amount of the experimental data; the structures show almost no violations of experimental restraints. The non-bonded contacts are also good which can be assessed by calculating the standard Lennard-Jones potential (this is not part of the minimized target function); it is invariably found to be negative. In addition, the hybrid method is even more CPU efficient than DISGEO alone, making it possible to calculate a large number of structures.

Structure Determination With Annealing Dynamics

The minimization scheme described in this section exploits the analogy with simulated annealing further and tries to determine a structure from NMR data directly without using a starting structure generated with a different program or folding strategy. In order to achieve this, the temperature has to be raised to high enough values that the kinetic energy is comparable to the potential energy barriers between different folds of the polypeptide chain. Equivalently, the potential energy can be scaled down by reducing all force constants; this latter method is actually used here (see figure 2).

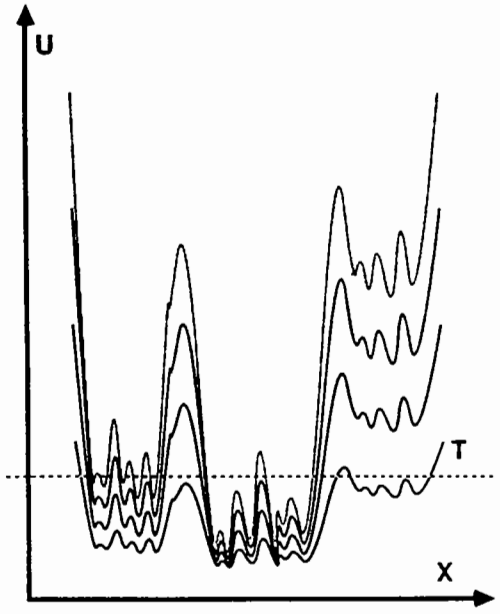


Figure 2: Instead of raising the temperature to very high values, the calculation is performed at a constant temperature. All force constants are reduced to very low values initially and increased slowly during the calculation.

The starting point of the calculations can be a "high temperature conformation" (such as a random distribution of atoms, figure 3) or in fact any conformation which is then heated up. The annealing schedule, which is very straightforward, is illustrated in figure 4. The very low starting values for all force constants are chosen in such a way that the total potential energy is approximately equal to the kinetic energy at 1000 K for a random starting structure. Every 1000 steps the force constants are increased by multiplying them with a constant factor (here 1.25). The only force constant which is treated differently is that of the vdW repulsion term: It is initially kept constant and is only increased after the conformation has converged to the correct fold. Identical to the procedure in the hybrid method, the heating stage is followed by a cooling stage and final minimization.

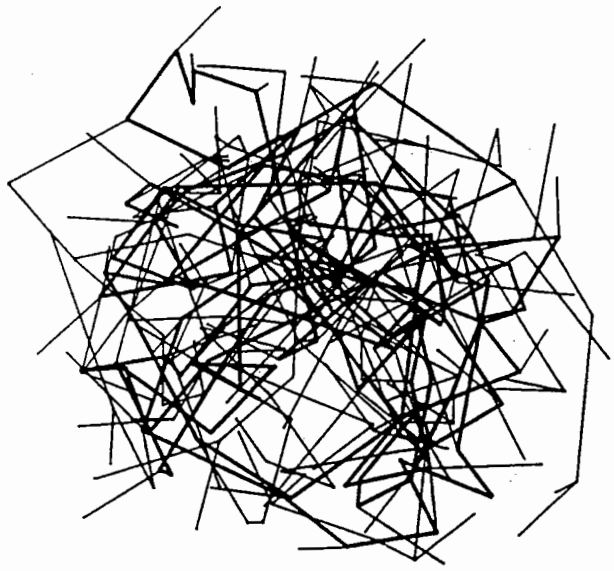


Figure 3: A "high temperature" starting conformation for crambin generated by choosing random x, y and z coordinates. The diameter of the structure is approximately 20 Å. From Nilges et. al., 1988c.

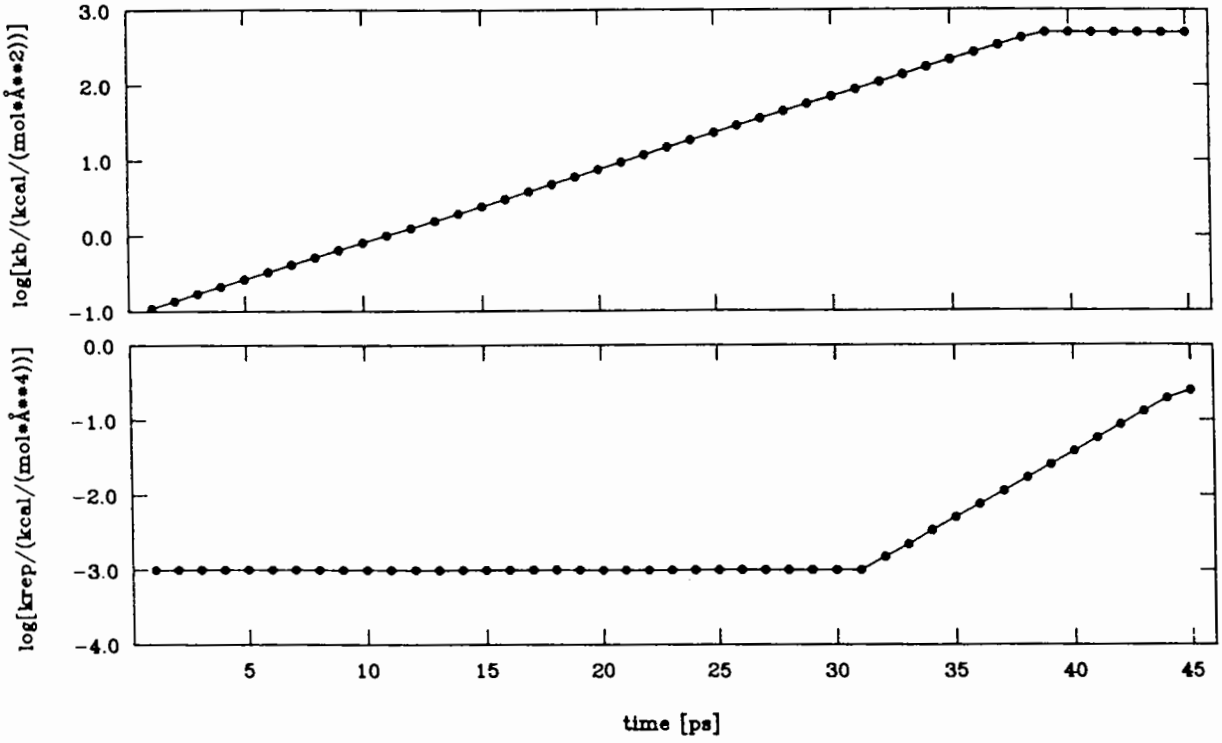


Figure 4: Time dependence of the bond force constant k_b and the vdW repulsion force constant k_{rep} during the course of a dynamics calculation on crambin. The other force constants are scaled simultaneously with the bond force constant, but k_{noe} is only increased to a maximum value of 100 kcal mol^{-1} $Å^{-2}$. From Nilges et. al., 1988c.

Figure 5 shows a trajectory from a test calculation with the same set of model distance restraints as in the previous model calculations. In this particular calculation, the starting structure is an incorrect fold, namely, the complete mirror image of the crystal structure (containing D-amino acids). All distance restraints are optimally satisfied in this structure; the only term which distinguishes it from the correct fold is the chirality restraint at the asymetric carbon atoms. The starting conformation is thus in a deep false minimum. Additional test calculations were carried out for Crambin, starting from a variety of incorrectly folded structures; the method invariably found the correct fold. It was successfully applied to a few other small proteins and peptides.

Figure 5 also demonstrates that the size of the molecule remains approximately constant throughout the calculation. In contrast to methods that vary torsion angles like DISMAN (Braun & Go, 1985), all structures in the trajectory are compact and "globular" with this method.

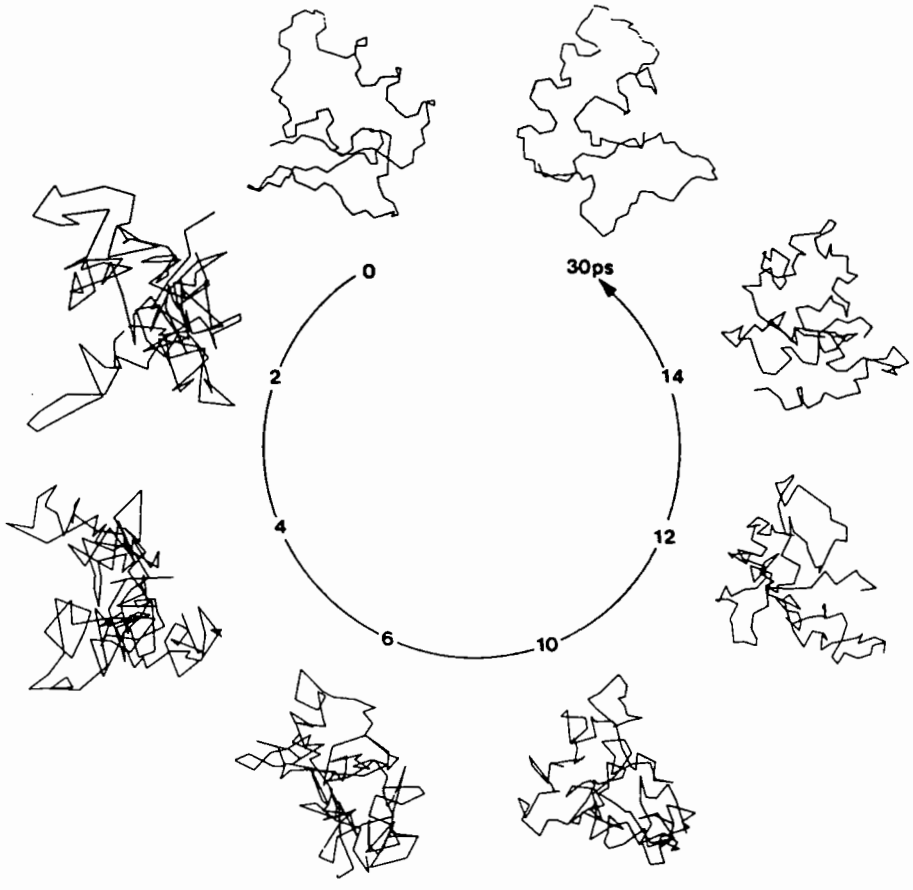


Figure 5: Path of the dynamics calculation for crambin starting from the global mirror image of the crystal structure. Only the N, C^{α} and C backbone atoms are shown. From Clore et. al., 1989.

While this method is not as CPU efficient as the hybrid method described in the previous section (around 20 hours for Crambin on a μVax III computer), the CPU costs are not prohibitive. This method may be especially valuable for small peptides where DG algorithms, in our experience, perform poorly.

Concluding Remarks

The success of the calculational strategies described in this paper relies mainly on 1) the inherent power of molecular dynamics calculations to overcome energy barriers, 2) the way the energy barriers themselves are varied in the course of the calculations by either changing the relative weights of the single terms in the target function or scaling the potential energy. Strategies similar to the one described in the last section may prove useful for other minimization problems, such as protein folding or three-dimensional structure prediction studies. The experimental part of the target function could then incorporate a hydrophobic interaction.

Ackgnowledgements: This work was supported by the Max-Planck Gesellschaft and grant 321/4003/0318909A from the Bundesministerium für Forschung and Technologie (to AMG and GMC) and by the Intramural AIDS Targeted Antiviral Program of the Office of the Director, NIH (AMG and GMC). We would like to thank A. Brünger for a very stimulating collaboration on the subject of RD and B. Brooks for useful discussions. MN would like to thank T. Holak and P. Kraulis for useful suggestions and discussions.

Present Address (MN): The Howard Hughes Medical Institute, and Department of Molecular Biophysics and Biochemistry, Yale University, 260 Whitney Avenue, New Haven, CT 06511, U.S.A.

References

Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A., & Haak, J.R., J. Chem. Phys. 81 (1984), 3684.

Braun, W., Quarterly Reviews of Biophysics 19 (1987),115.

Braun, W., & Go, N., J. Mol. Biol. 186 (1985), 611.

Brünger, A.T., Clore, G.M., Gronenborn, A.M., & Karplus, M., Proc. Natl. Acad. Sci. USA 83 (1986), 3801.

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

Brünger, A.T., X-PLOR Manual, Yale University, 1988.

Clore, G.M., & Gronenborn, A.M., Protein Engineering 1 (1987), 275.

Clore, G.M., Gronenborn, A.M., Brünger, A.T., & Karplus, M., J. Mol. Biol. 186 (1985), 435.

Clore, G.M., Brünger, A.T., Karplus, M., & Gronenborn, A.M., J. Mol. Biol. 191 (1986a), 523.

Clore, G.M., Nilges, M., Sukumaran, D.K., Brünger, A.T., Karplus, M., & Gronenborn, A.M., EMBO J. 5 (1986b), 2729.

Clore, G.M., Gronenborn, A.M., Nilges, M., & Ryan, C.A., Biochemistry 26 (1987a), 8012.

Clore, G.M., Gronenborn, A.M., James, M.N.G., Kjær, M., McPhalen, C.A., & Poulsen, F.M., Protein Engineering 1 (1987b), 313.

Clore, G.M., Gronenborn, A.M., Nilges, M., Sukumaran, D.K., & Zarbock, J., EMBO J. 6 (1987c), 1833.

Clore, G.M., Nilges, M., & Gronenborn, A.M., in: Computer-Aided Molecular Design (Graham Richard, ed), IBC Technical Services, London, 1989, in press.

Crippen, G.M., & Havel, T.F., Acta Crystallogr. A 34 (1978), 282.

Havel, T.F., DISGEO, Quantum Chemistry Program Exchange Program No. 507, Indiana University, 1986.

Hendrickson, W.A., & Teeter, M.M., Nature 290 (1981), 107.

Holak, T.A., Kearsley, S.K., Kim, Y., & Prestegard, J.H., Biochemistry 27 (1988), 6135.

Kaptein, R., Zuiderweg, E.R.P., Scheek, R.M., Boelens, R., & van Gunsteren, W.F., J. Mol. Biol. 182 (1985), 179.

Kirkpatrick, S., Gelatt, C.D., & Vecchi, M.P., Science 220, (1983), 671.

Kline, A.D., Braun, W., & Wüthrich, K., J. Mol. Biol. 189 (1986), 377.

Konnert, J.H., & Hendrickson, W.A., Acta Crystallogr. sect. A 36 (1980), 344.

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

Nerdal, W., Hare, D.R., & Reid, B.R., J. Mol. Biol. 201 (1988), 717.

Nilges, M., Gronenborn, A.M., & Clore, G.M., FEBS Lett. 229 (1988a), 317.

Nilges, M., Gronenborn, A.M., Brünger, A.T., & Clore, G.M., Protein Engineering 2 (1988b), 27.

Nilges, M., Clore, G.M., & Gronenborn, A.M., FEBS Lett. 239 (1988c), 129.

Nilsson, L., Clore, G.M., Gronenborn, A.M., Brünger, A.T., & Karplus, M., J. Mol. Biol. 188 (1986), 455.

Noggle, J.H., & Schirmer, R.E., The Nuclear Overhauser Effect, Academic, New York, 1971.

Scheek, R.M., & Kaptein, R., in: NMR in Enzymology (Oppenheimer, N.J., & James, T.L., eds.), Academic, New York, 1988.

Wüthrich, K., NMR of Proteins and Nucleic Acids, J. Wiley, New York, 1986.

The Calculation of Protein Structure Using NMR Data

Timothy. S. Harvey,
Department of Biochemistry,
University of Oxford,
South Parks Road,
Oxford.
OX1 3QU

The use of Nuclear Magnetic Resonance (NMR)-derived information (NOEs, $^3J_{NH\text{-}\alpha CH}$ coupling constants and amide proton exchange data) to investigate the structures and properties in solution is now well established (1). I have studied a wide range of protein structures and their dynamic properties using such data in restrained molecular dynamics (RMD) simulations, starting from random conformations, crystal structures, predicted structures, or those calculated by distance geometry-based methods. I will illustrate the use of RMD refinement in all these cases using a range of protein sizes from melittin (26 residues), the epidermal growth factor (EGF) family of proteins (50 residues) to insulin-like growth factor-1 (IGF-1, 70 residues) and the different types of information that can be gained from such work.

(1) Melittin.

The structure of the transmembrane polypeptide melittin, the major lytic component of honey bee venom, has been investigated by high-resolution ^1H NMR (2). The X-ray coordinates (3) were used as a starting structure for RMD refinement. An interesting feature of this study was the apparent flexibility of the molecule during restrained molecular dynamics simulations using the GROMOS package (4), as seen in Figure 1.

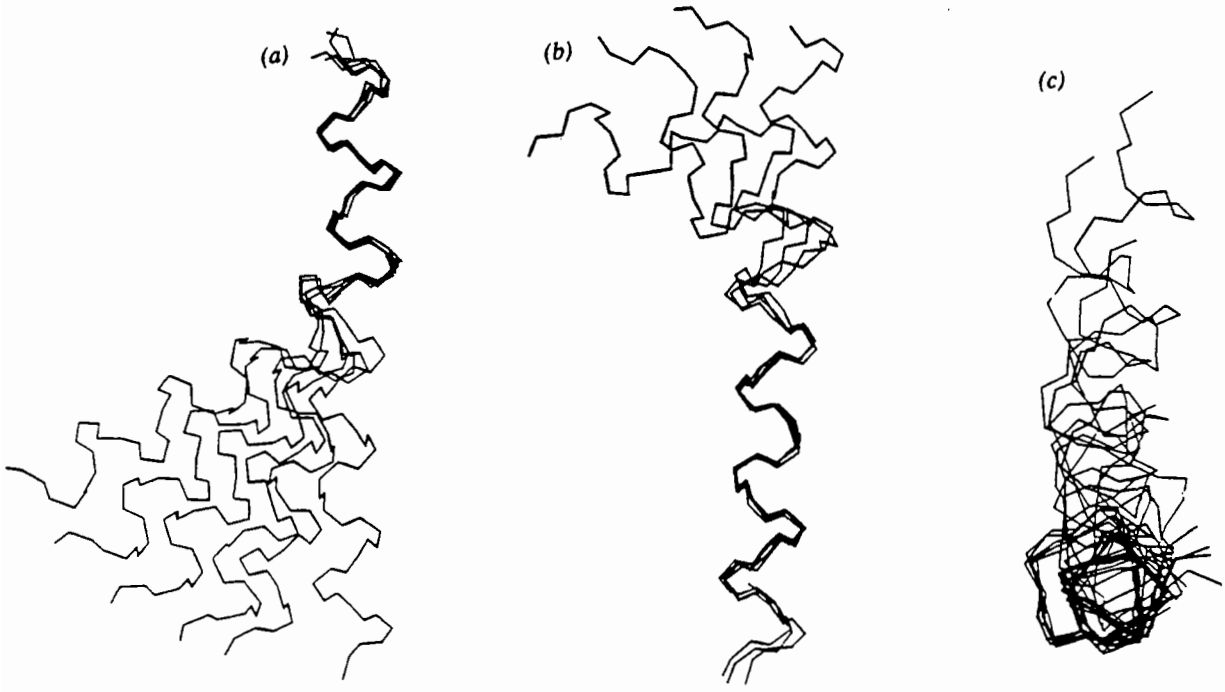


Figure 1. Melittin structures from RMD simulations (a) overlaid over the backbone atoms of residues 2-11, (b) overlaid over residues 13-22 and (c) viewed down the axis of the helix, showing the in-plane nature of the movement.

To investigate the dynamic properties of this molecule more closely, the angle between the two stable helical sections was followed as a function of time (Figure 2). The large variation in this angle becomes apparent. In earlier work on δ-lysin, no such flexibility was observed (5).

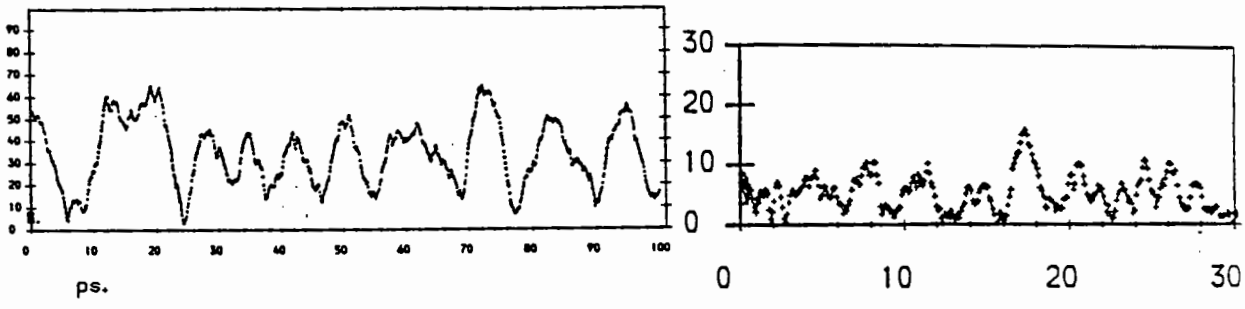


Figure 2. A plot of the angle between the two helical sections for (a) melittin and (b) δ-lysin. Note the change in scale of the ordinate in (b).

A comparison of the sequences and hydrogen-bonding patterns of melittin and δ-lysin shows a gap in the latter in melittin due to the presence of a proline at residue 14. It seems reasonable to assume that this is the cause of the flexibility associated with this region.

In addition to the NOE data, $^3J_{NH\text{-}\alpha CH}$ coupling constants and amide exchange rate data were collected. These were not included in the RMD runs, but were used as independent checks of the resulting structure as done previously (5). A good agreement between the measured $^3J_{NH\text{-}\alpha CH}$ was found. As observed in Figure 3, a good correlation between the calculated H-bond lifetime from RMD trajectories and the rate of amide exchange exists (6). These would be expected to have an inverse relationship.

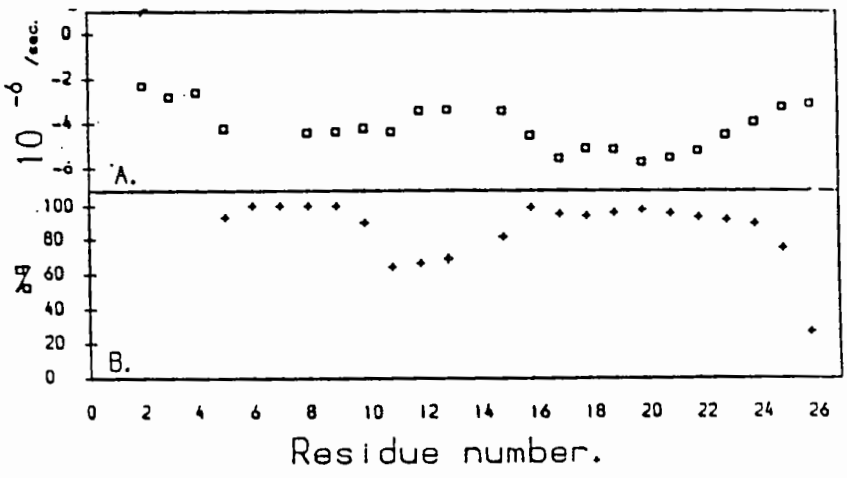


Figure 3. (a) The measured amide exchange rate and (b) the calculated H-bond lifetime from RMD runs. The ordinate in (a) is logarithmic.

A similar investigation has been carried out on a mutant of melittin in which proline 14 has been replaced by an alanine. A reduction in the flexibility, together with reduced amide exchange rates in the previously flexible region around residue 14 would be expected if the presence of the proline is responsible for the observed flexibility, and this is what is seen (Figure 5).

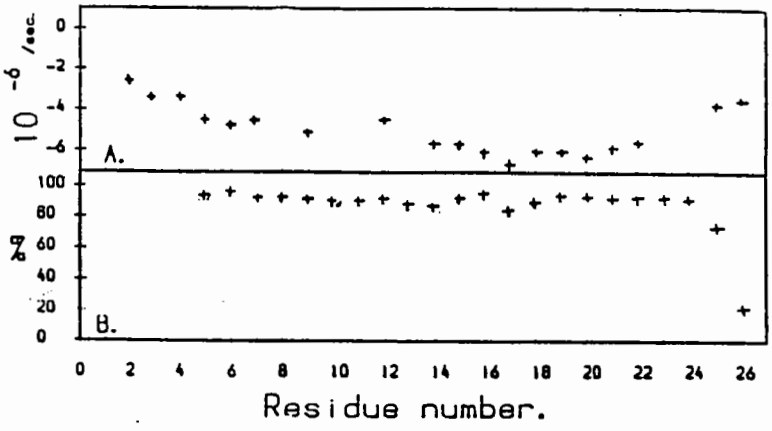


Figure 5. (a) The amide exchange rate as measured by NMR for the Ala-14 mutant of melittin and (b) the calculated H-bond lifetime from RMD simulations.

We are currently investigating the role of other residues in the hinge region, specifically glycine 12, which we have replaced by alanine, valine and phenylalanine residues in computer-based models.

The Epidermal Growth Factor (EGF) family of proteins.

The growth factors are a family of polypeptides which bind to receptors causing a cascade of intracellular responses leading to cell growth and division. We have studied two of these proteins, human Epidermal Growth Factor (hEGF) and human Transforming Growth Factor α (hTGFα).

hEGF was first studied as the 1-48 fragment. Initial structure calculation was done using DISGEO (7) and a data set comprising 224 NOEs, 10 H-bonds and 8 J-coupling constant restraints (8). The resulting structures were then refined using RMD (Table 1).

Table 1. The structure calculation of hEGF 1-48.

Structure	*After DISGEO*		*After EM*		*EM after MD*	
	P. E.	*R. E.*	*P. E.*	*R. E.*	*P. E.*	*R. E.*
I	*> 10⁵*	*63*	*-1424*	*211*	*-2222*	*118*
II	*> 10⁵*	*33*	*-1097*	*214*	*-1981*	*140*
III	*> 10⁵*	*25*	*-1128*	*187*	*-2262*	*79*
IV	*> 10⁵*	*41*	*-1386*	*154*	*-2142*	*52*
V	*> 10⁵*	*36*	*-1494*	*150*	*-2146*	*125*

Energy units kJ mol^{-1} with force constant 1000 kJ mol^{-1} nm^{-2}

Figure 6(a) shows five DISGEO structures overlaid. The position of the N-terminal region appears to be well defined, although there are few restraints defining its position relative to the rest of the molecule. It would appear that DISGEO has not searched conformational space well here, since more variability might have been expected. hEGF 1-53 is currently being calculated using DISMAN (7). This program appears to give a truer representation of the experimental data (Figure 6(b)).

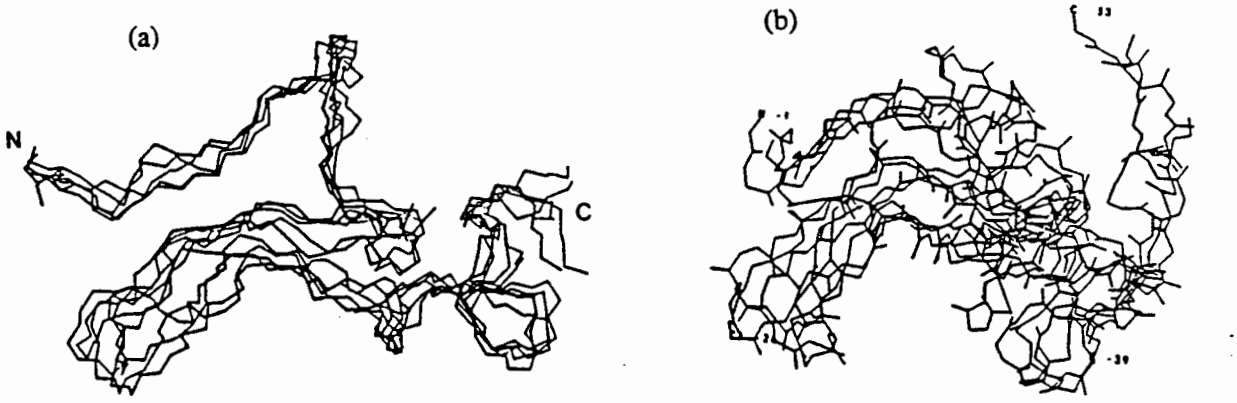


Figure 6. Five hEGF structures overlayed from (a) DISGEO (1-48) and (b) DISMAN (1-53) showing the different extent to which each programme has searched conformational space.

hTGFα provided more restraints for initial distance geometry-based calculations using both DISGEO and DISMAN (383 NOE restraints) (10). We also model-built one structure (I) on our previously-calculated hEGF structure. The resulting structures were then refined by RMD. Table 2 summarises the work.

Table 2. The structure calculation of hTGFα.

Structure	*Distance Geometry*		*Restrained MD*		*Restrained EM*		*Largest*
	P.E.	*Sum viol.*	*P.E.*	*Sum viol.*	*P.E.*	*Sum viol.*	*Violation (nm)*
I	*>10⁵*	*25.9*	*-1330*	*0.26*	*-2319*	*0.27*	*0.049*
II	*>10⁵*	*3.7*	*-1250*	*0.41*	*-2202*	*0.37*	*0.039*
III	*>10⁵*	*8.4*	*-1218*	*0.80*	*-1592*	*0.73*	*0.071*
IV	*>10⁵*	*6.4*	*-1380*	*0.68*	*-1824*	*0.65*	*0.056*
V	*>10⁵*	*3.7*	*-1432*	*0.34*	*-2070*	*0.40*	*0.050*

Potential energy (P.E.) units kJ mol⁻¹, sum of violations (Sum viol.) units nm.

A problem sometimes encountered when calculating NMR-based structures *ab initio* is that structures may be produced whose chain folds are mirror images of each other. Obviously this does not occur for α-helices. The problem of deciding which is the correct one may be resolved using signed distance maps (11). Comparison of such plots of known secondary structure from the NMR-derived structures with those obtained from X-ray data allows this problem to be overcome if it is assumed that the handedness is the same as those in the database.

Both hTGFα and hEGF appear to be quite flexible proteins. The variability of the resulting structures is relatively high despite their good agreement with experimental data. An example of this for hTGFα is shown in Figure 7. Other NMR evidence exists to show that these are flexible proteins in solution; the slowly exchanging amide protons in hEGF are still 1000 times faster than in BPTI (12).

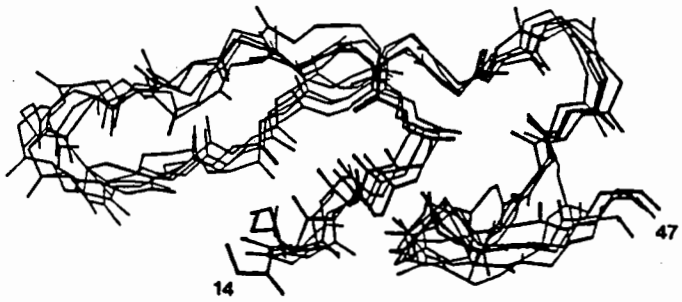


Figure 7. Five energy minimised hTGFα structure overlayed over the backbone atoms of residues 14-46. The average rmsd is 2.0 A.

The knowledge of both the structures of hTGFα and hEGF and their sequence homologies with other proteins allows us to classify their conserved residues into those which are structurally and those which are functionally important (13). In our hEGF structure, these residues lie on one face of the molecule, and are shown in Figure 8.

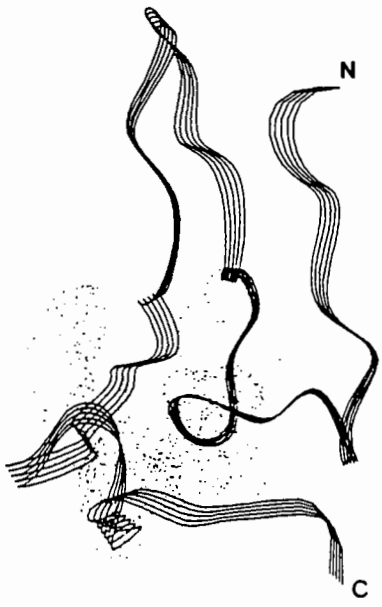


Figure 8. hEGF (1-53) with residues 13, 15, 16, 41, 43 and 47 which are believed to be directly involved in receptor-binding, shown as surfaces.

Insulin-like Growth Factor-1.

Another molecule where knowledge of the 3-dimensional structure would provide a valuable insight into the nature of its receptor binding surface is IGF-1. A modelled structure for IGF-1 exists (14), having been built on the homologous insulin crystal structure. Since its secondary structural elements seemed in quite close agreement with those seen in our NMR studies, this was used as a starting point for RMD refinement using the 367 NOE restraints. No H-bond data was obtained because of the high temperature used in the experiments, and no $^3J_{NH-\alpha CH}$ coupling constant data because of broad NH. (15). This work is still in progress, and it seems that some differences do exist, both in the secondary structure and in the orientation of certain important side-chains. The refinement is summarised in Table 3.

Table 3. The structure refinement of IGF-1.

	Total potential energy	*Sum of violations*
	(kJ mol^{-1})	*(nm)*
Modelled structure (based on Insulin)	*>10^{12}*	*3.5*
Restrained energy minimisation	*-1460*	*2.3*
Restrained molecular dynamics (45 picoseconds)	*-2370 (av.)*	*1.3 (av.)*
Restrained energy minimisation	*-2479*	*0.69*

The IGF-1 structures calculated so far show good convergence. The most significant difference when compared to the starting structure would appear to be around the short helix from residue 54 to 60 (figure 9).

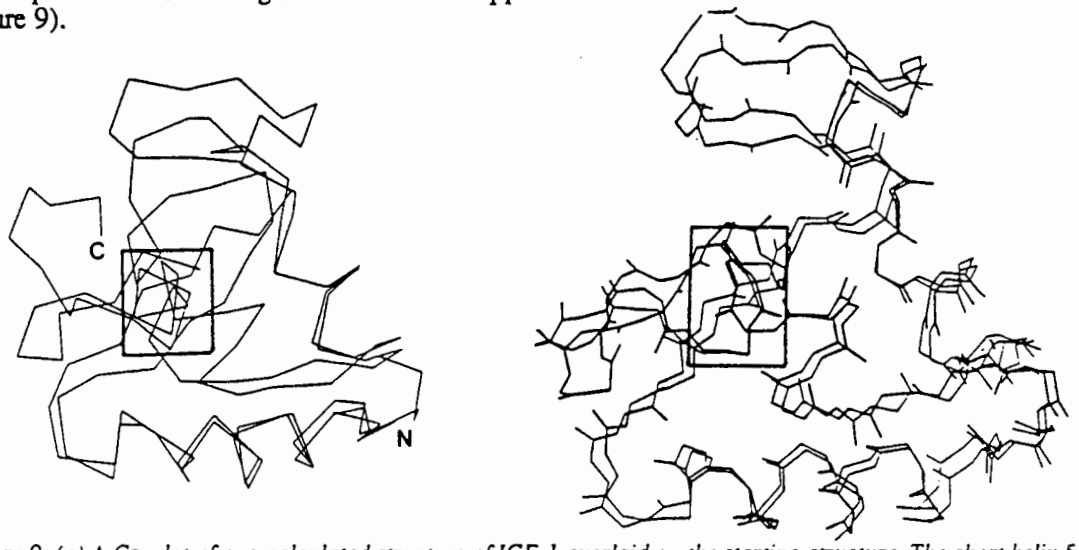


Figure 9. (a) A Cα plot of one calculated structure of IGF-1 overlaid on the starting structure. The short helix 54-60, where there appears to be a difference, is shown boxed, (b) two calculated IGF-1 structures overlaid over backbone atoms showing the convergence obtained.

One difference which may have important implications is that the the orientation of the sidechains of residues Phe 23 and Phe 25 are different in the calculated structures when compared to the model, as shown in Figure 10.

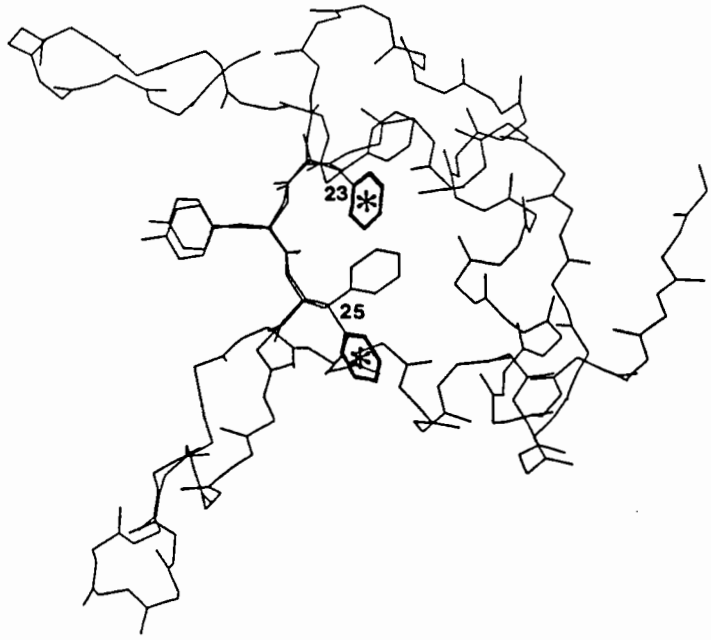


Figure 10. The calculated conformation of the region 23-25 overlaid on to the corresponding region in the starting structure showing the change in position of sidechains 23 and 25. The calculated positions are marked with an asterisk.

Summary.

The type of information that can be obtained by the combination of NMR and RMD is limited by the size of the protein. For melittin which is relatively small, the simulations are accurate enough to allow prediction of experimental observables with reasonable confidence, because of the large amount of NMR data available. For proteins such as the hEGF family which are flexible and so do not give as full a data set, we are still able to produce structures which give us a valuable insight into important features such as the receptor binding site. Larger proteins such as IGF-1 provide less information again, but we are still able to produce meaningful models of their structure.

Acknowledgements.

I would like to thank Iain Campbell, Rob Cooke, Annalisa Pastore, Renzo Bazzo, Chris Dempsey, Tony Wilkinson (ICI Pharmaceuticals), Mike Tappin and Charles Courquin for their invaluable contributions to this work, and ICI and Monsanto for providing proteins.

References.

(1) Wuthrich, K. (1986) *"NMR of Proteins and Nucleic Acids."* Wiley Interscience, New York.

(2) Bazzo, R. , Tappin, M. J. , Pastore, A. , Harvey, T. S. , Carver, J. A. and Campbell, I. D. (1988) *Eur. J. Biochem.* **173**, 139-146.

(3) Terwilliger, T. C. and Eisenberg, D. (1982) *J. Biol. Chem.* **257**, 6010-6022

(4) Van Gunsteren, W. F. and Berendsen, H. J. C. (1987) *"Groningen Molecular Simulation (GROMOS) Library Manual."* BIOMOS B. V. , Nijenborgh 16, Groningen, The Netherlands.

(5) Tappin, M. J. , Pastore, A. , Norton, R. S. , Freer, J. H. and Campbell, I. D. (1988) *Biochemistry* **27**, 1643-1647.

(6) Pastore, A. , Harvey, T. S. and Campbell, I. D. (1989) *Eur. Biophys. J.* **16**, 363-367.

(7) Havel, T and Wuthrich, K (1984) *Bull. Math. Biol.* **46**, 673-698.

(8) Cooke R. M. , Wilkinson, A. J. , Baron, M. , Tappin, M. J. , Campbell, I. D. , Gregory, H. and Sheard, B. (1987) *Nature* **327**, 339-341.

(9) Braun, W. and Go, N. (1985) *J. Mol. Biol.* **186**, 611-626.

(10) Tappin, M. J. , Cooke, R. M. , Fitton, J. E. and Campbell, I. D. (1989) *Eur. J. Biochem.* , **179**, 629-637.

(11) Braun, W. (1983) *J. Mol. Biol.* **163**, 613-321.

(12) Richarz, R. , Sehr, P. , Wagner, G. and Wuthrich, K (1979) *J. Mol. Biol.* **130**, 19-30.

(13) Campbell, I. D. , Cooke, R. M. , Baron, M. , Harvey, T. S. , and Tappin, M. J (1989) *Prog. Growth Factor Res.* **1**, 13-22.

(14) Blundell, T. L. , Bedarkar, S. , Rinderknechet, E. and Humbel, R. E. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 180-184.

(15) Cooke, R. M. , Harvey, T. S. , Likos, J. J. and Campbell, I. D. (1989) manuscript in preparation.

PROTEIN STRUCTURE FROM SIMULATED N.M.R. DATABASES

by
Robert M. Esnouf

Laboratory of Molecular Biophysics,
Rex Richards' Building,
South Parks Road,
Oxford OX1 3QU.

Protein structure determination from N.M.R. data usually proceeds in two stages. First, an initial model (or set of models) is generated which is then further refined by Molecular Dynamics using restraining potentials derived from the N.M.R. data. Several approaches have been used to generate the starting model, the most common being from distance geometry algorithms (Crippen and Havel, 1978) and in particular the programs DISMAN and DISGEO (Havel and Wüthrich, 1984 and 1985; Havel, 1986). In other protocols deliberately simple starting models are used and restrained Molecular Dynamics at elevated temperatures (simulated annealing) is relied on to fold the protein correctly. Examples of these simple starting models include a fully extended polypeptide chain; an extended chain with α–helices preformed; and even a random array of atoms (Brünger *et al.*, 1986; Clore *et al.*, 1986; Nilges, Clore and Gronenborn, 1988). With these methods it is found that the way in which the restraints are introduced during the simulated annealing run is of paramount importance in 'driving' the protein towards the correct final conformation, and it is possible to generate structures with the wrong global fold or lacking secondary structural elements. An alternative way of generating the initial model is to build it up from fragments of known protein structures which obey the N.M.R. restraints well. This is the approach we have been using, and it has also been explored by Jones and Thirup, 1986. The rationale behind this approach is that not all of the structures which obey a set of N.M.R. restraints are actually adopted by proteins. A knowledge of structures known from crystallography could be used to add extra information to the N.M.R. data and hence provide a better starting model.

N.M.R. can give three types of information relevant to the fold of a protein backbone. The ${}^3J_{\mathrm{NH}\alpha}$ coupling constant can be measured from COSY experiments and this is related to the backbone ϕ angle by the Karplus equation (using parameters from Pardi *et al.*, 1984):

$$ {}^3J_{\mathrm{NH}\alpha} = 6.4\cos^2\theta - 1.4\cos\theta + 1.9 $$

where $\theta = | \phi - 60^\circ |$. This gives ${}^3J_{NH\alpha}$ values in the range 2 to 10 and can be measured to an accuracy of about ± 0.5. The ϕ angle in extended β–structures is about -120° and this coincides with the maximum value of ${}^3J_{NH\alpha}$. However, values from 2 to 7 occur for four different ranges of ϕ angles and one of these ranges includes the α–helix angle ($\phi = -60^\circ$) so the predictive power of the ${}^3J_{NH\alpha}$ value is limited. Amide proton exchange rates may also be measured, and this gives an indication of the lifetime of hydrogen bonds to the backbone. Strong hydrogen bonds are expected in both helices and sheets, so the information is again of little use in structure prediction. The major source of information is interproton distance restraints derived from NOESY experiments. The strength of the signal obtained diminishes with the sixth power of the interproton distance, so only distances up to around 5Å give rise to observable signals. There are also other restrictions which fall into three types. Firstly, no signals are observed between protons attached to adjacent atoms, so, for example, no signals are observed between the amide proton and the Cα proton in the same residue. Secondly, rapidly exchanging protons will not give rise to signals, thus NOEs are not observed to protons attached to oxygen (and some nitrogen) atoms. Finally, some regions of the spectrum are very crowded and weak signals in this area may be obscured. This crowding makes it uncommon to observe weak NOEs between protons both attached to carbon atoms and also obscures interactions along a single sidechain. Another consideration is that the derived distances are only approximate and so are usually divided into classes such as weak (<2.8Å), medium (<3.4Å) and strong (<4.1Å) with errors of 0.5Å. These are the divisions used in the programs discussed later. Finally, stereospecific assignments between individual Cβ protons are rarely made and in this paper NOEs involving Cβ protons are calculated to the mid–point of the two protons.

Using these rules it is not too difficult to write a program to simulate N.M.R. data for a protein whose structure has been determined crystallographically. Such a program was written and used to create a database of N.M.R. data from 34 well–refined protein structures. The information stored was based on the positions of the N, Cα, Cβ, C and O atoms of each residue. It consisted of the residue names, coordinates of the atoms, their secondary structure assignment using the program DSSP (Kabsch and Sander, 1983), ${}^3J_{NH\alpha}$ values and NOE interproton distance data. In early versions of the program the NOE data were stored as no NOE or as Weak, Medium or Strong NOE depending on the interproton distance. In later versions, just the actual interproton distances were stored as this gave a better way to measure the error in the fit between fragments of structure. By storing approximate distances in Fortran BYTE variables the whole database for the 5410 residues processed only occupied 3MBytes and so could easily be held in memory, which decreased the programs' running time.

The programs were tested using simulated data for crambin, a 46 residue protein from Abyssinian cabbage seed (Hendrickson and Teeter, 1981). The molecule is shown in Figure

1. It consists of two α–helices – one from residues $7-18$ which is distorted towards a 3_{10}–helix at one end, and the other from residues $23-30$. The helix–turn–helix motif has a particular handedness as is shown in the figure. There is also a short stretch of antiparallel β–sheet between residues $2-4$ and $32-34$ and possibly one turn of a 3_{10}–helix between residues 42 and 44. There are three disulphide bridges in the molecule – between residues 3 and 40, residues 4 and 32 and finally residues 16 and 26. From this structure theoretical $^3J_{NH\alpha}$ values were calculated and a set of 263 NOE restraints were derived, divided into three classes as above. This set of restraints was very similar to those used by Brünger *et al.*, 1986. A breakdown of the restraints is given in Table 1 where a short range NOE is one between residues less than five apart in the primary sequence and a long range NOE is between more distant residues.

The original program to generate a model structure for crambin took all the NOE data related to a short fragment of crambin (between four and six residues) and, by various criteria, tried to match it to all similarly sized fragments in the database. The fragment which gave the best fit was selected as the model for that fragment of crambin. Then a new fragment, overlapping with the previous one by three residues was taken and a best fit found for that piece and so on. At the end of the program all the selected fragments were superimposed on each other to produce a complete model structure. The prediction was done in sequential fashion from the N–terminus to the C–terminus. The resulting structure had good local fit but the relative positioning of pieces of secondary structure was poor. This is shown by the RMS errors in the fit of the Cα coordinates of the model to the crystal structure of crambin for various residue ranges:

Residues 1 to 18	0.85Å
Residues 1 to 32	2.57Å
Resiudes 1 to 46	8.70Å

Despite the poor RMS fit of the model, it was then used as the starting point for restrained Molecular Dynamics using GROMOS87. After the simulations at various temperatures the RMS error in the fit improved to 5.0Å but would get no better. The reason for this is shown in Figure 2, which shows the model before restrained Molecular Dynamics from the same viewpoint as Figure 1. The area of β–sheet isn't well formed, but more importantly, the handedness of the helix–turn–helix motif is incorrect. The energy barrier stopping the helices crossing each other to obtain the correct handedness is obviously too great to be surmounted even by dynamics at 6000K.

One of the major shortcomings of this method is the treatment of the long range NOEs, as these are vital for the positioning of pieces of secondary structure relative to each other. With this program, however, the long range NOEs are not considered until several

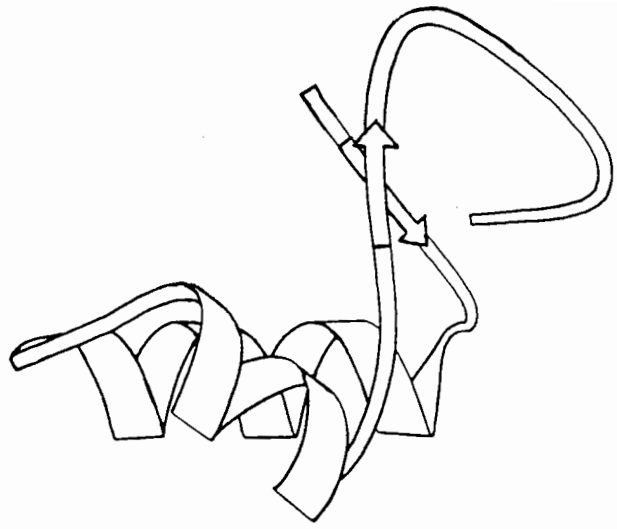


Figure 1: *Crystallographic Structure of Crambin.*

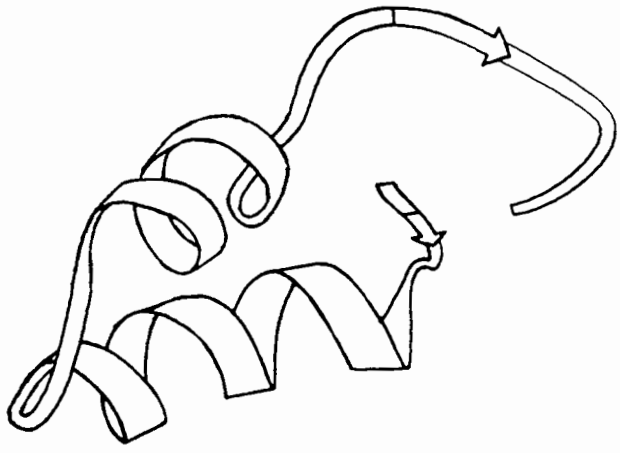


Figure 2: *Original Model Structure of Crambin.*

	Intra–Residue	Short Range	Long Range	Totals
Strong	9	48	8	65
Medium	37	79	34	150
Weak	0	43	5	48
Totals	46	170	47	263

Table 1: *Breakdown of Simulated NOEs for Crambin.*

short fragments have been selected and joined together and by this stage it is too late for the long range NOEs to alter the shape of the molecule significantly. It would be better to have some idea of the overall shape of the molecule first, by considering the long range NOEs, and then to build a more detailed structure based on the short range NOEs.

It has been shown that the different secondary structural elements can be distinguished by characteristic NOE patterns (Wüthrich, Billeter and Braun, 1984). Table 2 shows the NOEs for crambin residue by residue, where the number shown is the number of NOEs between those two residues (*e.g.* 3 between residues 13 and 2), and is divided into two halves. The bottom–left shows NOEs between all protons and the top–right is restricted to those involving only backbone amide, Cα and Cβ protons. From this table the gross secondary structure can be seen by inspection. α–helices give NOEs between a residue and the ones three and four further along in the sequence as there are about $3\frac{1}{2}$ residues per turn. This shows up as a band of NOEs 3 and 4 off the diagonal and can been seen in the table from residues 6 to 17 and from residues 23 to 30, though in the latter case the pattern isn't so perfect. β–sheet structures are shown by long range NOEs between the two strands and these appear as diagonal lines away from the leading diagonal in Table 2. If the line is parallel to the leading diagonal it signifies a parallel sheet, if perpendicular then an antiparallel sheet is indicated. Thus it can be seen that there is a short antiparallel sheet between residues $2-4$ and $32-34$. Finally, the mutual orientation of the two helices is indicated by NOEs between residues 9 & 30, 13 & 26/27 and 17 & 23. These occuring every three to four residues, *i.e.* after every complete turn of each helix.

A program searching for these NOE patterns using the database has been written which produces two outputs: a diagram of the β–sheet regions of the molecule and a summary of the secondary structure based on the Kabsch and Sander, 1983 definitions. These definitions are based on hydrogen bonding. A hydrogen bond is defined to exist between an amide on one residue and a carbonyl group on another if

$$27.888\left[\frac{1}{r_{ON}}+\frac{1}{r_{CH}}-\frac{1}{r_{OH}}-\frac{1}{r_{CN}}\right]<-0.5,$$

where r_{ON} is the distance between the carbonyl oxygen and the amide nitrogen in Ångstroms *etc.* Using this definition, the following secondary structural elements are defined:

T: Turn with H–bond to residue 3, 4 or 5 ahead in the sequence,
G: At least two repeated 3–turns giving a 3_{10}–helix,
H: At least three repeated 4–turns giving an α–helix,
I: At least four repeated 5–turns giving a π–helix,

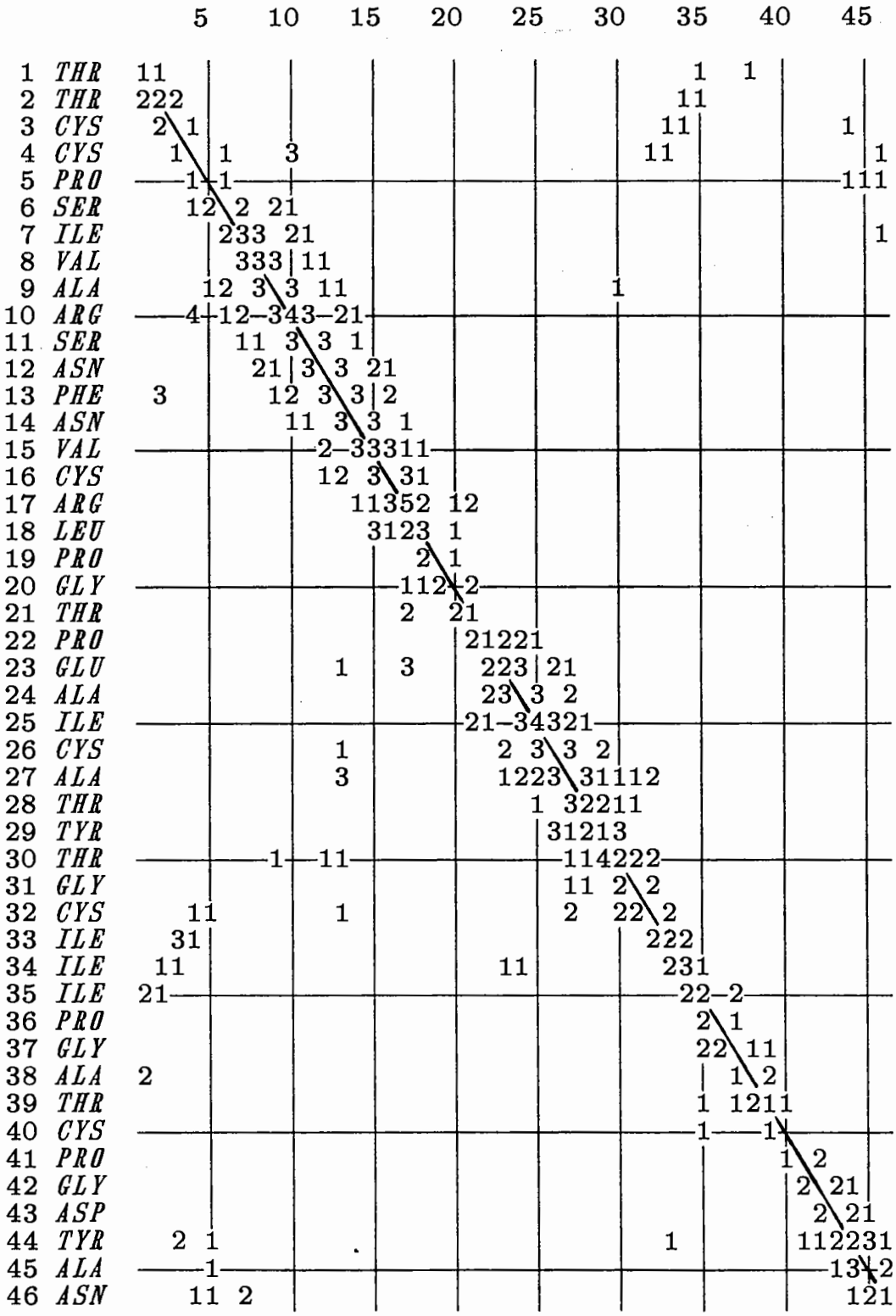


Table 2: *Summary of NOE restraints for Crambin.*

B: A hydrogen bond bridge between residues distant from each other in the sequence,

E: Repeated bridges forming an extended β–strand,

S: A bend with curvature of at least 70° between residues $i-2 \rightarrow i$ and $i \rightarrow i+2$.

These definitions are good for describing the overall secondary structure, but are not necessarily accurate in locating the ends of pieces of structure. In particular, helices followed by turns and the exact end of β–sheets may not be well defined. These difficulties are also expected to be present in any secondary structure prediction based on the definitions. Also, a bend has no necessary implications for interproton distances, so the detection of this feature relies on similar bends being present in the database.

The prediction of the β–sheet region does not use the database, but searches for the characteristic long range NOEs which are summarised in Table 3. In this table the distances expected between protons are shown, for example $\mathbb{D}_{N\alpha}(i,j)$ is the distance between the amide proton on residue i and the Cα proton on residue j (Wüthrich, Billeter and Braun, 1984). It should be noticed that there is a very short $\mathbb{D}_{\alpha N}(i,i+1)$, *i.e.* between consecutive residues within a strand. The inter–strand NOEs are the more useful ones, and it should be noticed that short Cα–amide and amide–Cα distances alternate along the strands for parallel sheets, whilst short Cα–Cα and amide–amide distances alternate in antiparallel strands. The program searches for these patterns (which may, of course, be incomplete) and indeed locates the β–sheet in crambin (residues $2-4$ and $32-34$) correctly. However, this is only a simple case. As a sterner test, NOEs were simulated for

Parallel Sheets		Antiparallel Sheets	
Intra–strand:		*Intra*–strand:	
$\mathbb{D}_{\alpha N}(i,i+1)$	2.2Å	$\mathbb{D}_{\alpha N}(i,i+1)$	2.2Å
$\mathbb{D}_{N\alpha}(i,i+1)$	4.8Å	$\mathbb{D}_{N\alpha}(i,i+1)$	4.7Å
$\mathbb{D}_{NN}(i,i+1)$	4.2Å	$\mathbb{D}_{NN}(i,i+1)$	4.3Å
$\mathbb{D}_{\alpha\alpha}(i,i+1)$	4.3Å	$\mathbb{D}_{\alpha\alpha}(i,i+1)$	4.3Å
Inter–strand:		*Inter*–strand:	
$\mathbb{D}_{\alpha N}(i+1,j+1)$	3.0Å	$\mathbb{D}_{\alpha\alpha}(i,j)$	2.2Å
$\mathbb{D}_{N\alpha}(i,j)$	3.0Å	$\mathbb{D}_{\alpha N}(i,j+1)$	3.2Å
$\mathbb{D}_{NN}(i,j-1)$	4.0Å	$\mathbb{D}_{N\alpha}(i+1,j)$	3.2Å
$\mathbb{D}_{\alpha\alpha}(i+1,j)$	4.8Å	$\mathbb{D}_{NN}(i+1,j-1)$	3.3Å

Table 3: *Short interproton distances in β–sheets.*

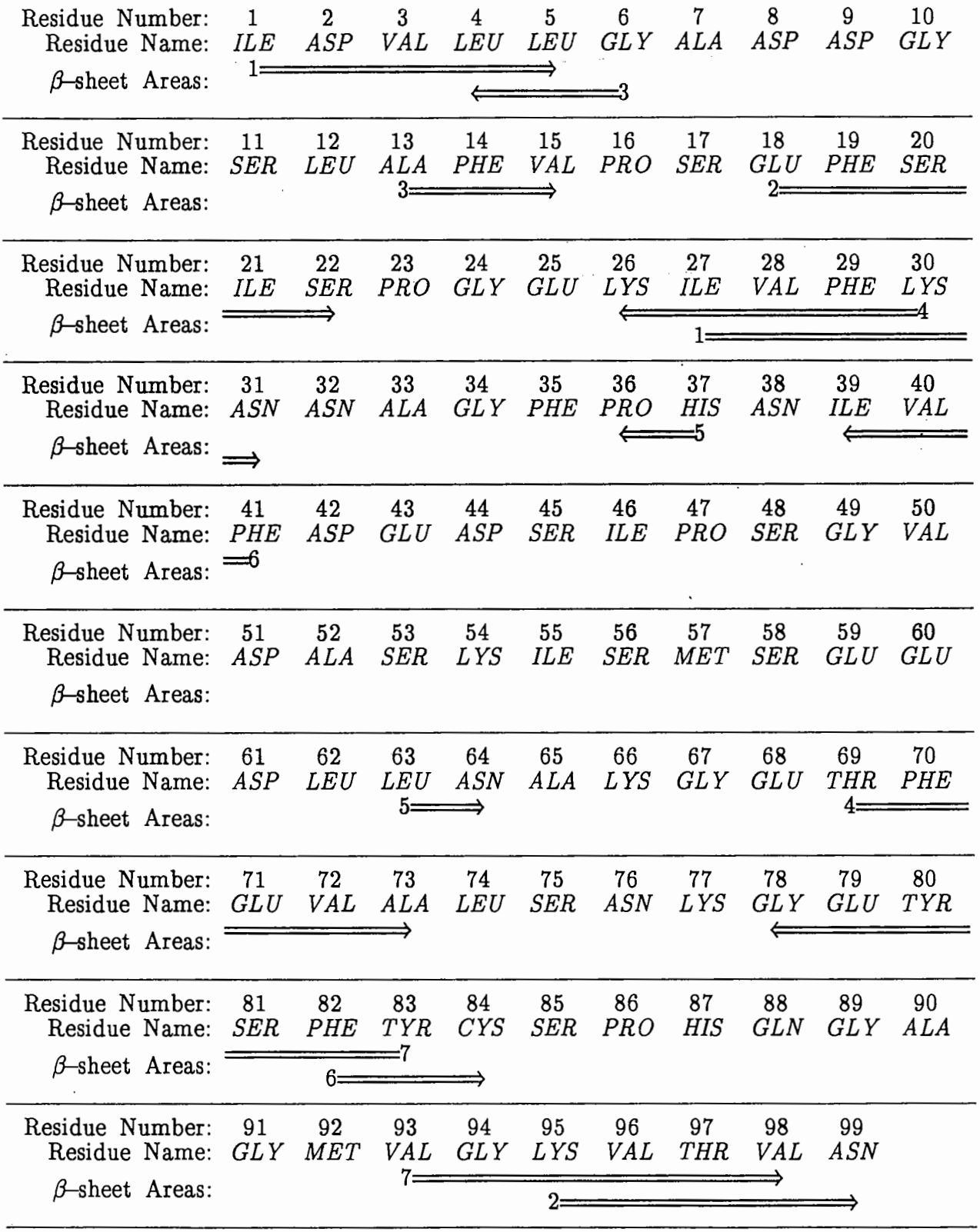


Table 4: *β–sheet Prediction for Plastocyanin from Long Range NOEs.*

the 99 residue electron transport protein plastocyanin (Guss and Freeman, 1983) which has an extended β–structure. The output summary of the β–structure prediction is shown in Table 4. For example, arrow 1 means that there are parallel strands from residues $1-5$ and $27-31$, and arrow 4 shows antiparallel strands from residues $26-30$ and $69-73$ *etc.* This prediction is almost entirely correct, the only possible errors being in the exact length of the features.

The other part of the secondary structure prediction uses the database and proceeds by matching short fragments from the database to the NOEs – this process essentially just uses the short range NOEs. If the fit is better than a predefined cutoff then the secondary structure of that fragment is stored and from all fragments with suitable fits, the most likely secondary structural element is used as the predicted secondary structure for that residue of the unknown. Using this method good secondary structure predictions have been made for both crambin and plastocyanin, summarised in Table 5. For crambin, 40 out of the 46 residues have secondary structural elements assigned in agreement with DSSP, whilst for plastocyanin, 86 out of 99 residues are assigned in agreement. It should be noted that of the 6 differences for crambin, 2 are due to β–sheet length and 1 is due to the difference in the exact length of a helix immediately followed by a turn – both areas where DSSP is likely to be a poor predictor. Of the 13 differences with plastocyanin, 9 are due to differences in length of one in β–sheet areas. The program has been run using different fragment lengths and different cutoffs for matching the fragments and the results show that the predictions are fairly constant over a range of values and that the best fragment length

Residue	1 2 3 4 1234567890123456789012345678901234567890123456
2° Str.	EEESSHHHHHHHHHHHHTTS HHHHHHHHSEEE SSS SSG
DSSP	*EE SSHHHHHHHHHHHTTT HHHHHHHS EE SSS GGG*

(a) Crambin

Residue	1 2 3 4 5 12345678901234567890123456789012345678901234567890
2° Str.	EEEEE TT SEEESTEEEEETT EEEEEE SS EE EEESTTSS TT
DSSP	*EEEES TT S EESSEEEE TT EEEEEE SS B EE TTSS TT*
Residue	6 7 8 9 123456789012345678901234567890123456789012345678
2° Str.	HHHHS TT EESTT EEEEE S EEEEEEE GTTTT EEEEEE
DSSP	*HHHHS TT B STT EEEEE S EEEEEE GGGTTTT EEEEEE*

(b) Plastocyanin

Table 5: *Secondary Structure Predictions using NMR Database.*

is five or six residues. If a shorter fragment is used then there are too few NOEs relevant to that fragment to define the structure uniquely, on the other hand, if the fragment is too long then it is unlikely that suitable pieces of structure exist in the database.

The knowledge of the secondary structure allows the starting model to be built in a more logical way than just from one end to the other. The program which does this is, as yet, incomplete and so the discussion will be limited to the prediction of a model structure for the two helices in crambin. Fragments from the database are matched against the NOEs for crambin for sections which are known to be in an α–helix. For the helix starting at residue 17 four fragments were required, whilst for the helix starting at residue 23 three were needed. Unlike the original program, the best 100 fits were stored for each fragment along with their scores. Also stored were scores based on the RMS error in the fit between all possible consecutive fragments. For the four fragments with 100 possibilities for each fragment there are 10^8 possible helices that can be constructed. The program knows that not only should the predicted helix fit the NOEs, but also the true helix is a continuous piece of structure and so the fits between fragments should be good. With this in mind, the program calculates the total error made up from the NOE fit error and the overlap fit error and selects the helix with the smallest overall error. The weighting of the two sorts of error can be adjusted, and provides a good check for the program. Biasing towards the NOE fit produces a helix made up from the fragments with the best fits, biasing towards the overlap produces the best fitting single piece of helix actually found in a protein. The ideal weighting is obviously between these two extremes. Using this program, the helix between residues 7 and 18 was predicted with an RMS error in fit of the backbone atoms to the crystal structure of only 0.20Å. For the helix between residues 23 and 30 the error was 0.25Å. These compare with values of 1.06Å and 1.08Å for the model predicted previously.

	Helix 1 (7–18)		Helix 2 (23–30)	
	ϕ	ψ	ϕ	ψ
Energy minimised Distance Geometry Structure	26°	24°	30°	29°
DG followed by Restrained Dynamics	17°	15°	30°	28°
Model built from NMR database fragments	**4.3°**	**5.1°**	**4.2°**	**9.3°**

Table 6: *ϕ,ψ angular RMS differences for helices in model crambin structures.*

The reasons for this improvement are the knowledge that it is a helix that is being predicted and the fitting scheme outlined above, which means that the selected fragment is not necessarily the one with the best score.

The quality of the predicted helices can also be assessed by considering the RMS errors in the backbone ϕ and ψ angles. Clore *et al.*, 1987 give these data for a study of crambin. A comparison with their data is shown in Table 6 and shows that the helix structures generated from database fragments are significantly better than those generated from distance geometry. Much work is still needed to give a complete model structure based on databases, but the results so far are encouraging that good model structures can be constructed.

ACKNOWLEDGEMENTS:

I would like to thank S.E.R.C. for financial support and Dr. Garry Taylor for help and advice during the last eighteen months.

REFERENCES:

Brünger, A. T., Clore, G. M., Gronenborn, A. M. and Karplus, M. *Proc. Natl. Acad. Sci. U. S. A.* 83 (1986) 3801.

Clore, G. M., Brünger, A. T., Karplus, M. and Gronenborn, A. M. *J. Mol. Biol.* 191 (1986) 553.

Clore, G. M., Nilges, M., Brünger, A. T., Karplus, M. and Gronenborn, A. M. *FEBS Lett.* 213 (1987) 269.

Crippen, G. M. and Havel, T. F. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* A34 (1978) 282.

Guss, J. M. and Freeman, H. C. *J. Mol. Biol.* 169 (1983) 521.

Havel, T. F. DISGEO, Quantum Chemistry Program Exchange Program No. 507, Indiana University, Bloomington, IN. (1986).

Havel, T. F. and Wüthrich, K. *Bull. Math. Biol.* 46 (1984) 673.

Havel, T. F. and Wüthrich, K. *J. Mol. Biol.* 182 (1985) 281.

Hendrickson, W. A. and Teeter, M. A. *Nature (London)* 290 (1981) 107.

Jones, T. A. and Thirup, S. *EMBO J.* 5 (1986) 819.

Kabsch, W. and Sander, C. *Biopolymers* 22 (1983) 2577.

Nilges, M., Clore, G. M. and Gronenborn, A. M. *FEBS Lett.* 239 (1988) 129.

Pardi, A., Billeter, M. and Wüthrich, K. *J. Mol. Biol.* 180 (1984) 741.

Wüthrich, K., Billeter, M. and Braun, W. *J. Mol. Biol.* 180 (1984) 715.

On The Treatment of Disorder in Protein Refinement: Some Preliminary Results

John Kuriyan
The Rockefeller University
1230 York Avenue
New York NY 10021, U.S.A.

Abstract:
The success of the simulated annealing method in speeding up the refinement process leaves two major remaining problems. One is the estimation of the errors in protein structures, now especially important since the automated molecular dynamics refinement process undoubtedly leaves some regions uncorrected, while improving the overall structure. The other problem is the treatment of dynamics and disorder in X-ray refinement. Inadequacies of the refinement model are probably responsible for our current inability to reduce the R-factors to much below 15%, whereas the intrinsic errors in the data are not likely to be more than 5-10% for the best data sets (Wlodawer et al., 1988). Some preliminary results of the application of simulated annealing to evaluating the results of high-resolution refinement, and searching for conformational disorder, are discussed.The need for incorporating anisotropic temperature factors is demonstrated. The rigid-body model of Schomaker and Trueblood (1968) is shown to be an effective treatment for protein temperature factors at low resolution.

Treatment of Discrete Disorder:

Crystallographers currently rely on visual inspection of electron density maps to locate and model discrete conformational disorder (Smith et al. 1986, Svensson et al., 1988).This process is simplest when the electron density consists of well separated peaks. At 1.5Å, atomic positions separated by less than ~1.2Å are unlikely to result in distinct peaks (Swanson, 1988). For sidechains with widely separated conformations (for example, alternative conformations involving rotation around the sidechain dihedral angle χ_1), this does not pose a serious problem. For more subtle disorder, all conformers may be within a continuous peak of density, and the modelling requires considerable effort at interpretation. An automated procedure for locating and modelling discrete disorder would therefore be helpful, and simulated annealing, with its intrinsic ability to search conformational space, is an obvious choice. We have carried out, in collaboration with S.K. Burley, A.T. Brunger, M. Karplus and W.A. Hendrickson, some simulated annealing runs on RNase (data from Burley and Petsko) and crambin (in collaboration with Wayne A. Hendrickson). Following a suggestion of Martin Karplus, two structures for the protein were simultaneously included in the refinement process, as explained below. Conformational disorder in ribonuclease has recently been described by Wlodawer and co-workers (Svensson et al., 1988). Each of the 13 sidechains reported to be disordered by them is also found to be disordered by our SA approach, even though the solvent conditions used for crystal growth are different in the two cases.

Estimation of Errors in Refined Structures:

We have previously suggested that the uncertainties in refined parameters may be estimated by perturbing the structure and re-refining the parameters (Kuriyan et al., 1987). In our first study, energy minimization (without reference to the X-ray data) was used to gently perturb the refined structure of CO-myoglobin, and two re-refinements were done to arrive at estimated standard deviations of 0.1Å for backbone atoms and 0.2Å for sidechain atoms. Large deviations were found for atoms surrounding a discretely disordered sidechain, Arg 45, suggesting that a residue existing in two very different conformations might lead to slight, but measureable, disorder in the residues it interacts with.

Energy minimization leads to relatively small shifts from the original X-ray structure, and generally cannot escape from local minima in the potential energy surface. One might therefore expect that the errors estimated in this way would be unrealistically low. It is preferable to move the structure well away from the local minimum of the X-ray structure and allow it to find a new minimum. However, the shifts introduced to perturb the structure must not be too large, or else the new refinement will not be able to converge to an acceptable structure with low R-factor. The SA method provides a powerful means of perturbing refined structures while keeping reasonable stereochemistry and not moving too far away from the X-ray minimum. This is closely related to the search for disorder, and our application of SA refinement to error analysis is also discussed below.

being suggested as an improved refinement *model*, but only as a search procedure for characterizing the errors and disorder. We limit ourselves to 2 structures because the inclusion of any more makes the system underdetermined and potentially unstable.

Discriminating Between Errors and Disorder: We use two criteria for distinguishing between displacements in structure due to errors or uncertainty, and those due to discrete disorder or anisotropic motion. The first is the reproducibility of the shifts: we expect that displacements due to coordinate uncertainty will be rather randomly distributed, whereas structural changes due to disorder will show systematic features. The second criteria is the fit to electron density. This can eventually be automated, but our intial studies rely on visual examination of density maps.

Methods: A starting model for twin SA runs is generated by removing all alternative conformations, and duplicating the structure. The current version of X-PLOR does not allow for multiple occupancy, so the "twin" structures are translated by 2 or more unit cell lengths along any of the crystallographic axes (suggested by Axel T. Brunger). Because of the non-bonded cut-offs used in the calculations (8Å), the two molecules are now invisible to each other in terms of the empirical energy; however, the translational symmetry of the crystal guarantees that they are completely equivalent in terms of the X-ray structure-factor calculation. This initial system corresponds to two identical molecules with 0.5 occupancy each. The second molecule is translated back to the primary unit cell for analysis.

Simulated annealing runs are carried out as described by Kuriyan et al. (1989), except that no inter-molecular crystal packing forces are computed (the use of two molecules in symmetry related positions prevents this). This does not lead to any problems, because of the high quality of the starting model. Trial annealing runs showed that the solvent molecules move out of electron density during the higher temperature stages and, due to lack of covalent attachment, they do not always refine back to reasonable locations. Harmonic constraints of 200 Kcal/mole/Å are therefore placed on all the non-hydrogen solvent atoms.

Application to Ribonuclease:

The initial model used is that of S.K. Burley and G.A. Petsko (in preparation) for RNase at room temperature, crystallized from ammonium sulfate. 14937 reflections between 8Å and 1.5Å are used in the current refinements. The refined model of Burley and Petsko has an R-factor of 15.6%, with 129 residues (8 with alternative conformations), 165 water molecules and two sulfate ions.

Two independent trajectories, using X-ray data to 2Å resolution, are run at 2000K for 3 picoseconds resulting in the A-1 and A-2 structures. This is followed by 1 picosecond of dynamics at 300K. The resulting twin structures are subjected to least-squares optimization against data to 1.5Å resolution. The scale factor of the X-ray term is set to 40000 during the dynamics stages and is increased to 180000 during the optimizations, resulting in the B-1 and B-2 structures. These twin structures are then separated into two single structures each, which are independently optimized, resulting in the C-1 to C-4 structures. All the optimized structures reported here have rms deviations of bond lengths and angles from ideality of ~0.02Å and $<4^0$, respectively. One SA run takes approximately 20-40 minutes of CPU time on a Cray XMP, depending on the resolution and temperature. Isotropic temperature factors have been optimized for all the structures except C.

The electrostatic term in the potential function is very poorly modelled, and perhaps biases the results, especially at the surface. Also, the high temperature and reduced X-ray resolution (2Å) of the dynamics stage, followed by abrupt quenching of the system, might cause the system to get trapped in false minima. A third trajectory has therefore been generated, at a lower temperature (500K) for 3 picoseconds, using data to 1.5Å, and resulting in the A-3 structure. The system is gradually cooled from 500K to 0K over 2 picoseconds. Least-squares optimization of the twin structures is then carried out, as before, yielding the B-3 structures. The electrostatic term in the potential is turned off completely for this run. The results of this run are very similar to those obtained from the earlier runs.

In order to evaluate the effect of the dynamics stages, two additional least-squares optimizations are carried out. In one, the starting twin model is simply subjected to least-squares optimization, resulting in the D structure. For the second calculation, the atomic positions of the starting structure are perturbed by introducing random displacements (a Gaussian distribution of standard deviation 0.5Å is used), followed by least-squares optimization (E structure).

Results

The lowest R-factors are obtained with the twin B structures (13%), but these are only somewhat lower than that for the original single structure with 8 disordered sidechains (15.5%). The twin refinement, as such, is *not a very good model* because it introduces too many parameters without a great reduction in R-factor. Nevertheless, when we compare the results of the various twin refinements, we find a number of features that are reproduced in all three independent runs; these provide useful suggestions for improving the refinement model.

Molecular Dynamics Simulations and X-ray Refinement:

Molecular dynamics simulations provide an extremely detailed, though approximate, picture of protein motions, and analyses of MD simulations of several proteins have revealed features of the internal motions with important consequences for the interpretation of structures obtained by refinement against X-ray data (Mao et al 1982; Yu et al. 1985; Kuriyan et al.1986a; Ichiye and Karplus 1987,1988). These studies have not involved direct comparison with experiment, but rather have used the simulations as closed systems in which to study the effect of various approximations.

The atomic probability distributions are seen to be strongly anharmonic and anisotropic (Mao et al. 1982). MD studies on lysozyme and myoglobin (Ichiye and Karplus 1987,1988; Kuriyan et al.,1986a) showed that the most important anharmonic effect is the existence of multiple peaks in atomic distribution functions, rather than third or fourth order corrections to the harmonic function. If the peaks are well separated, refinement of single site models leads to the atoms moving away from the mean position and towards the major peak in the distribution. The mean-square displacements of the atoms are underestimated because the refinement tends to fit only the peaks closest to the atoms' final position.

Ichiye and Karplus (1988) suggest that if the refinement model is to be extended beyond the isotropic harmonic one, for many atoms the most effective improvement would be to include two positions, rather than one position and a harmonic anisotropic tensor. It is quite common to find surface sidechains in more than one widely separated conformation (Smith et al., 1986). In the case of buried residues, however, examination of experimental electron density maps suggests that alternative conformations, if present, would be close together (<1Å). If the electron density has such closely spaced multiple peaks, it may be possible to refine different structures that fit the X-ray data equally well, but represent different conformations adopted by the molecule. At least two structures would have to be used in the refinement, otherwise the refinements would collapse to the mean position (Ichiye and Karplus, 1988). This was the motivation for starting our "twin" refinement experiments, using two structures simultaneously in the calculations.

Rigid-body Motions of Protein Molecules:

In a pioneering study, Phillips and co-workers analyzed the temperature factors of lysozyme in terms of rigid-body motions of the whole molecule (Sternberg et al. 1979). They showed that the temperature factors of the backbone atoms of the protein can be explained as arising from rigid-body librations (TLS model of Schomaker and Trueblood, 1968) of a pair of molecules in the crystal, about a common axis. Several sidechains, particularly on the surface, were found to have B-factors higher than that predicted by the TLS model, implying contributions from internal motion. A practical application of the results of Sternberg et al., which does not concern itself with the physical significance of the TLS model, is the refinement of B-factors for proteins at low resolution. We describe below our success in modelling the B-factor variation in influenza virus hemaglutinin, where we reduce the number of B-factor parameter from ~4000 to 10.

RESULTS AND DISCUSSION

(i) Searching for Errors and Disorder by Simulated Annealing:

The program X-PLOR, developed by A.T. Brunger (1988b), based on CHARMM (Brooks et al., 1983) is used, with appropriate additions, for all the calculations described here.

Errors: Our idea for locating errors is quite simple, and involves running a number of SA refinements, starting from the current refinement model, but with different initial velocities. When the refinement model has large errors in it (~1-5Å), we have shown that deviations between structures obtained from independent annealing runs correlate well with the errors in the starting model (Kuriyan et al., 1989). Our aim in the present work is to see whether this approach can also be applied to highly refined structures (errors of ~0.1-0.5Å), and to see if the more extensive conformational sampling made possible by the SA method results in structures that are significantly different. This approach is similar in spirit to the generation of a large number of structures that fit the NMR constraints equally well (Brunger et al., 1986, Wuthrich, 1989).

Disorder: When multiple peaks in the probability distribution of an atom are separated by less than the sum of their half-widths, refinement using single site models will fit the mean position of the peaks (Ichiye and Karplus, 1988). Thus, if only one structure is used in the SA refinements, only structures close to the mean position will be obtained, and information about the conformational heterogeneity will be lost. To get around this problem, we use two structures simultaneously in the search procedure. A danger is that random shifts may be introduced into the structures, because of the reduction in the ratio of observations to parameters by a factor of 2. We therefore only consider features that are consistently reproduced between independent annealing runs. We emphasize that this procedure is not

Table 1

Structures	R-Factor
Initial	15.6%
A	~30%
B	~13%
C	~17%
D	~15%
E	~15%

Table 1 summarizes the R-factors of the various structures. The A structures are distorted due to the high temperature of the dynamics run, and have the highest R-factor. The individual structure refinements (C) have somewhat higher R-factors than the starting structure because alternative conformations have not been included, and also because the solvent model is somewhat damaged by the simulated annealing and needs to be improved by manual intervention.

On simply optimizing the starting twin model, very little movement away from the initial structure is observed (Table 2, structure D) and there is little improvement in the R-factor. This is because the structure is already in a local minimum, and simply duplicating it leaves it in the same minimum.The extent of conformational space sampled by the dynamics can be estimated from the deviations between each of the two molecules in the A runs, and is on the order of 0.5Å-1Å (Table 2). Applying random shifts (σ=0.5Å) to all the atoms, followed by optimization, does cause the atoms to move away from the initial structure (Table 2, E). However, the R-factor of the resulting twin structure is high (15.5%). This is because the random shifts applied to each atom are often in opposite directions for atoms in the same sidechain, and the resulting stereochemistry is very bad. Some sidechains become inverted at centers such as Cγ atoms, and the minimization is not able to correct these sorts of errors. The advantage of the simulated annealing method is that while random velocities are used to start the dynamics, the stereochemistry and fit to electron density are maintained at reasonable values during the simulation.

	δ_1	δ_2	δ_3	δ_4
Structure	Å	Å	Å	Å
A-1	0.63	1.02	0.42	0.83
A-2	0.68	1.10	0.45	0.87
A-3	0.37	0.67	0.26	0.50
B-1	0.43	0.87	0.23	0.59
B-2	0.40	0.88	0.21	0.67
B-3	0.45	0.75	0.24	0.47
C-1/C-2	0.11	0.51	0.10	0.56
C-3/C-4	0.09	0.71	0.09	0.55
D	0.03	0.05	0.05	0.07
E	0.28	0.59	0.16	0.23

Table 2: rms deviations between structures. Four types of deviations are given: (1) δ_1: rms deviations between backbone atoms of two molecules that are part of the same twin structure. For the A, B, D and E structures, both molecules were simultaneously present in the refinement model. For the C structures, these molecules were refined independently, but their deviations are reported for comparison. (2) δ_2: as in δ_1, but the deviations are computed over all sidechain atoms and carbonyl oxygen atoms that have an accessible surface area of less than 3.5Å^2 using a water sized probe. This results in 2/3rds of the sidechain atoms being included in the calculation. (3) δ_3 : deviation of backbone atoms from the original X-ray structure. (4) δ_4, as for δ_3, for the buried sidechain atoms. For the C structures, the deviations have been averaged over C1,C2 and C3,C4 in pairs.

The deviations between the *single* molecule refinements (C1-4) represent our estimate of the errors in the structure. These deviations are ~0.1Å for the backbone atoms (Table 1), which is consistent with other estimates of coordinate errors in RNase (Wlodawer et al., 1986) and myoglobin (Kuriyan et al., 1987) at comparable resolutions. It is significant that the SA refinement at this resolution (1.5Å) has not resulted in any differences in the structure of backbone. This emphasizes that the X-ray term in the potential function is a powerful restraint on the dynamics; without this term, backbone deviations as large as 1-2Å from the X-ray structure can be introduced within 3 picoseconds of dynamics.

A surprising result is that when two molecules are retained in the optimizations (the B structures), they show significantly greater deviations (backbone deviations of ~0.4Å see Figure 1). In Run 3, for example, the structures actually diverge *further* on optimization of the heated structure (Table 2). In Figure 1 it can be seen that some of the backbone atoms deviate by as much as 1Å. The larger shifts in the backbone in many regions are highly reproducible on repeating the simulations and are therefore a manifestation of systematic features in the data. In Figure 2a we superimpose six structures, from B1, B2 and B3, for residues 39 to 43 (which have the largest backbone displacements in Figure 1). It is seen that the shifts in backbone positions are highly clustered, rather than random. Note that when single structures are refined independently, this region has small backbone displacements (0.1Å instead of 1Å). Such clustered displacements between the two structures are also observed for a number of sidechains, especially the those of tyrosines, phenylalanines and histidines. In Figure 2b, we show the systematic displacement (by about 0.8Å) between six structures for Tyr 97.

We have quantified the extent to which the structures obtained by the twin refinement are clustered. For each atom, the two positions obtained from each annealing run are included in one or the other of two groups. In Figure 3 we plot the distribution of the ratios of distances between positions in different groups, over distances between positions in the same group. 50%, 30% and 10% of the atoms have ratios greater than 2.0, 3.0 and 5.0, respectively, indicating that a large number of atoms are significantly clustered. From Figure 3 it can be seen that the carbonyl atoms and aromatic residues especially tend to be clustered. For example, sidechain atoms of Tyr 97 have ratios between 4 and 6.

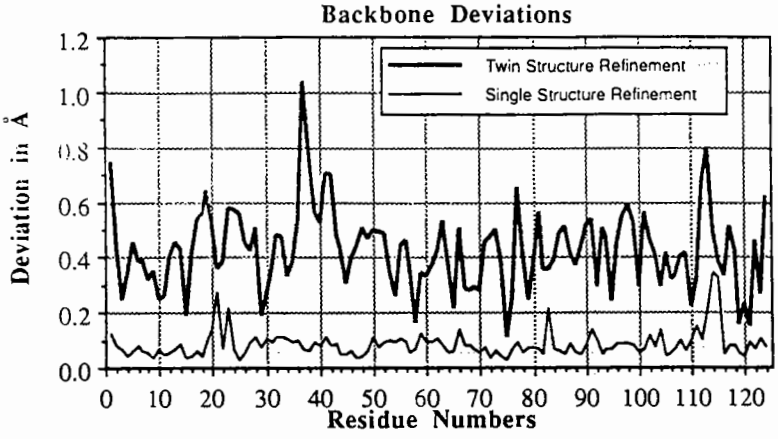


Figure 1. Backbone deviations between the two molecules in B3 that were simultaneously present during the optimization (thick line), and two molecules (C1 and C2) that were refined independently as single structures (thin line). The deviations between the two B structures are not changed on least-squares superpositioning, and similar shifts are obtained with the other B structures.

Examination of the regions that show these reproducible displacements reveals three categories of clustering. (1) The two groups of structures are in different torsional minima. These are mostly sidechain atoms on the surface. Some of these had been located in the original crystallographic model, but others, such as Gln 34 (Fig. 4) have been newly located by the twin refinement without manual intervention. An encouraging result is that many of the alternate conformations found in our twin refinements of the Burley & Petsko data (for crystals grown from high concentrations of sulfate) correspond to those reported by Wlodawer et al.(1988), even though their structure is for a salt-free protein crystallized from alcohol. For example, the two conformations for Asn 34 shown in Figure 4 have torsion angles of (χ_1 =-56 & -159) and (χ_2 = 114 &38), respectively. Wlodawer et al. report (χ_1 = -55 & -148) and (χ_2 = 132 & 43), for the same residue, close to the conformations found by simulated annealing. In Figure 4 we also show one of the structures obtained by random displacement followed by least-squares optimization and it is found to be trapped in an intermediate conformation. Another example of alternative torsional minima found by the simulated annealing method is Val 34, in Figure 2a. However, the twin refinement was unable to locate the alternative conformation for the active site histidine, 119. This residue has two conformations, separated by 180^0 on χ_1 (also reported by Borkakoti et al. 1982), and the barrier between the two is too high to be overcome during the short dynamics run. Neglect of the second conformation leads to a very strong density feature in difference maps, which is easy to identify. The SA method is most effective at locating alternative conformations that are relatively close together, and therefore separated by small barriers, and these are often the most difficult to model manually.

(2) The second class of clustering is when the groups of structures have fairly close torsional angles, but exhibit distinct hydrogen bonding. We find such alternative H-bonding schemes to be rare in the absence of torsional disorder. One example is at the active site of RNase, where Gln 11, Ala 41, Lys 41, a sulfate ion and two water molecules show mutually exclusive H-bonding in the two groups of structures. Lys 41 is in Figure 2a, where it can be seen that even though most of the sidechain is randomly distributed between the structures, the terminal N atoms fall into two clusters.

(3) The most common reproducible shift involves no large changes in torsional angles or H-bonding, but rather is the result of displacements of atoms that are correlated over several residues (Figure 2) and is seen throughout the protein, even in the interior. Since the displacements are about 1Å or less, they generally cannot be resolved in the 1.5Å electron density maps as separate peaks (Swanson, 1988). The magnitude of these displacements is consistent with the temperature factors in the interior of the protein, and suggest that these correlated shifts may be due to the neglect of anisotropy in the B-factors. In fact, anisotropic B-factors computed from the twin structures (by calculating the second moments of the bimodal distribution) are reasonable (Figure 5), and further suggest that the displacements might be modelled well by the TLS method. As a preliminary step we have refined overall translational tensors (no libration) for segments and residues (Figure 5), and the results encourage us to include the effects of rigid body libration as well.

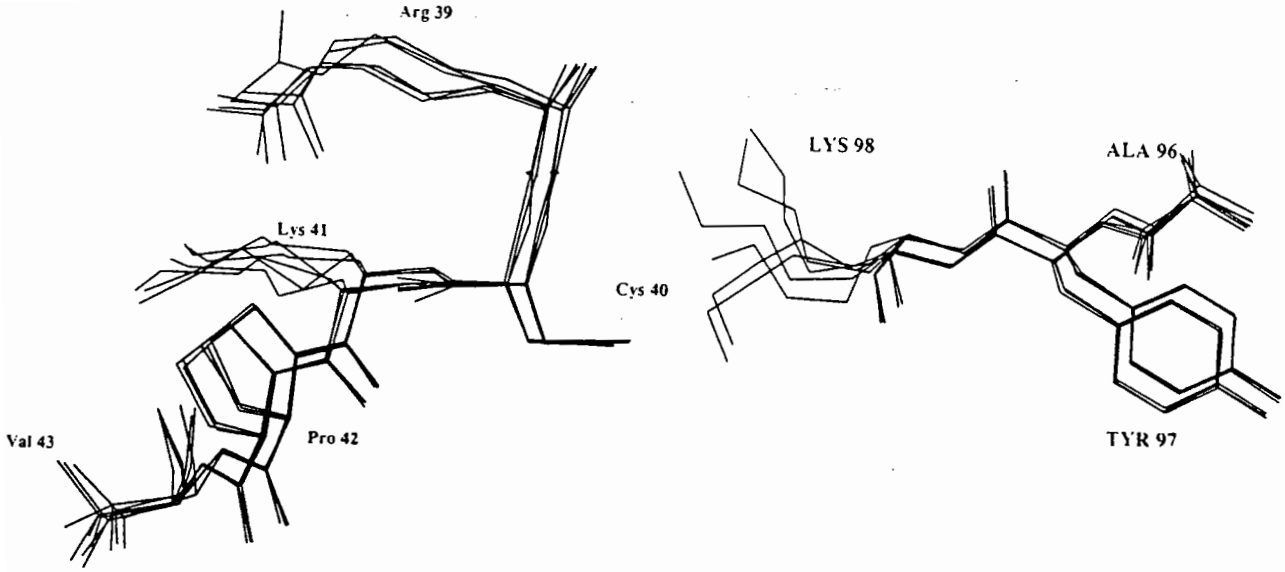


Figure 2. Refined structures from three independent runs are superimposed. Each run had two molecules simultaneously present, so there is a total of six molecules in the figures. **2a(left)** Residues 39 to 43. The backbone and the proline ring are displaced by about 1Å in the two groups of structures, but the deviations within each group are very small (<0.1Å). Note that the sidechain of Val 43 has two conformations, which are as reported by Wlodawer et al. (1988). The disorder seen for Lys 41 and Arg 39 are typical for surface sidechains. **2b(right)** As in 2a, for residues 96 to 98.

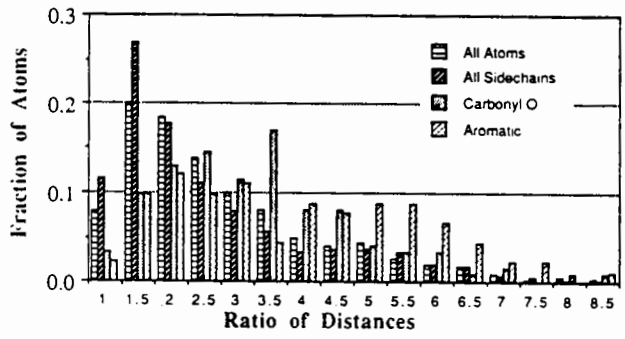


Figure 3. Histogram showing the ratio of d_1 and d_2. For each protein atom, the two positions in each of B1, B2 and B3 were classed as being in group A or group B. d_1 is the rms distance between a positions in group A and one in group B (9 distances for each atom) and . d_2 is the rms distance between positions internal to A and to B (6 distances). The histograms are normalized to unity.

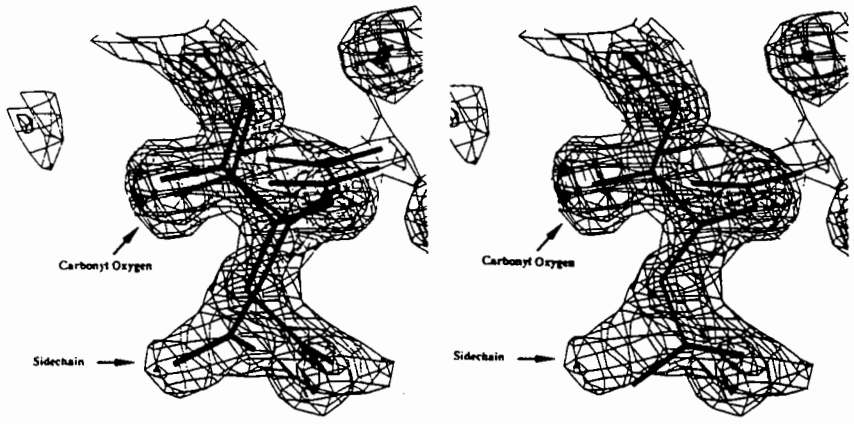


Figure 4: (2Fo-Fc) electron density maps for Asn 34. Left, two molecules from the B1 twin refinement. Right, one molecule from the random displacement / refinement.

Comparison with Crambin:

The resolution of the RNase data (1.5Å) makes it difficult to distinguish between the effects of anisotropy and discrete disorder. We therefore turned to crambin, for which data to 0.9Å are available (Teeter & Hendrickson, 1979) and for which the discrete disorder has been described (Smith et al., 1986). The "twin" annealing procedure is repeated for this small protein, including data to 0.9Å resolution. Very similar results to the RNase case are obtained. For example, Tyr 29 is found to occupy two sites, analogous to the displacements found in RNase (Figure 6). This residue has been modelled with two conformations by Smith et al. (1986) and even though the displacements are only on the order of 1Å, at

this resolution the electron density clearly shows that a two site model is better than a single site with anisotropy (Figure 6a). Another example is Pro 36, for which Smith et al. report two puckerings of the ring, with N-CA-CB-CG dihedrals of -18^0 and 28^0. Our twin refinement also produces two puckerings of the ring, with this dihedral at -22^0 and 28^0, obviously very close to the manually refined result.

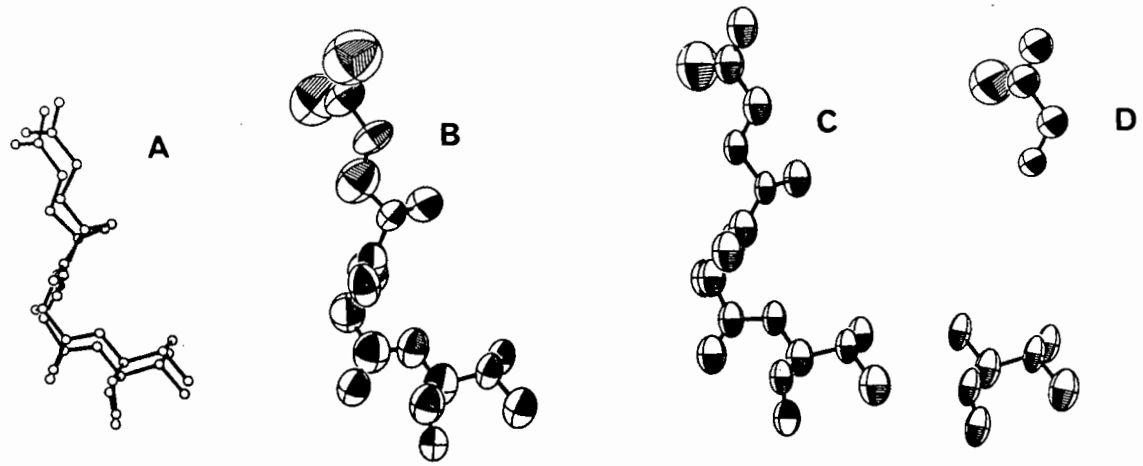


Figure 5. Residues 55 to 58.in ribonuclease (a) Two structures in B1. (b) The thermal ellipsoids calculated from the isotropic B-factors of the two B1 structures (c) Overall anisotropic thermal factor, refined against X-ray data, for all 4 residues (d) Overall anisotropic B-factors, refined for residues 55 and 58 independently. The effects of libration are not included in these rigid-body tensors.

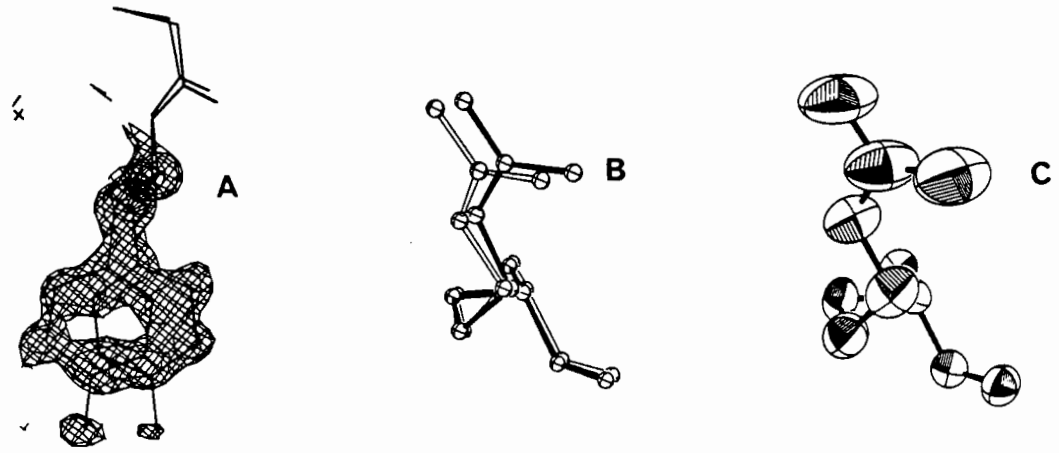


Figure 6. Crambin (a) Fo-Fc map for Tyr 29 in crambin, with the residue omitted from the phase calculation. The two structures shown are from the twin refinement. Note the well separated peaks for the terminal OH group. (b) Two structures for Arg 10, from the twin refinement. (c) Anisotropic B-factors for Arg 10.

On the other hand, a number of residues in crambin also show significant deviations in the "twin" refinements, and yet are within a continuous envelope of density with no evidence for discrete disorder (Figure 6b). We can compare these displacements with the unconstrained atomic anisotropic B-factors (Kuriyan, unpublished). The tensors for Arg 10 are shown in Figure 6(c), and comparison with the twin structures (Fig. 6b) shows that the displacements in the latter are along the principal axes of the thermal ellipsoids. Further comparison of the RNase and crambin results suggests that the displacements of the type shown in Figures 2 and 6(b) are best modelled by anisotropic B-factors.

(ii) Rigid-body Motions

The refinement of isotropic B-factors by the TLS method is described in detail by Sternberg et al. (1979). In this case, only the trace of the mean-square displacement tensor is considered, and the number of parameters reduces to 10: the trace of the translational tensor, 6 components of the libration tensor, which is retained in full, and 3 combinations of the screw tensor elements. These 10 parameters specify

the isotropic temperature factors of the entire protein molecule and can be determined either by refinement against X-ray data, or by refinement against individual isotropic B-factors, considered as the data. We have taken both approaches. The equations are given by Sternberg et al. (1979), and they will not be discussed here, except to state that the B-factor of an atom is dependent on its distance from three libration axes, that are not required to intersect each other or pass through the centroid of the molecule.

A simplified treatment of B-factors is most likely to benefit large proteins that only diffract to moderate resolution, so we chose to see how the method would work on influenza virus hemagglutinin. This structure has two chains in the monomer, with a total of about 4000 atoms. We treated both chains as part of the same rigid body, so there are 10 TLS parameters specifying the B-factors of the 4000 atoms. For a first test of the method we used the structure in the protein data bank, with the refined isotropic B-factors of Knossow et al. (1986) as the target "data". The fit for the A chain backbone is given in Figure 7(a), and is really quite striking. The 10 parameter model is able to reproduce the average B-factor curve quite accurately, given the resolution of the data used to derive the target B-factors (3Å).

For the B-chain (Figure 7b), the situation is somewhat different. For most of the structure the TLS model does well, except for residues 50-60. Here the TLS model predicts very low B-factors, whereas the refined individual B-factors are very high. This region turns out to be the only extended region in the molecule where the electron density is discontinuos and ambiguous, and the structural model has never been satisfactory (D.C. Wiley & W. Weis, personal communication). Poor electron density in this region persists even in data for mutant proteins (Wiley & Weis, personal communication). This is a very significant result, because it shows that comparison of individually refined B-factors with the TLS model is able to pick out the region of highest uncertainty in the model. The individual B-factors "blew up" in this region because of the poor fit to density.

There are several advantages to using a TLS model for refinements of such proteins. The number of parameters would be greatly reduced, and regions of coordinate error might be identified by refining individual temperature factors for a few cycles and comparing them with the TLS B-factors. The interpretation of electron density maps will also be improved, because very high temperature factors, such as obtained for the 50-60 region using individual B-factors, tend to obscure peaks in difference maps that migh otherwise signal errors. We are therefore continuing our modelling of hemaglutinin, and plan to use the observed structure factors as the target data rather than the refined individual B-factors.

For smaller proteins that diffract to high resolution, the TLS model is able to fit the general shape of the B-factor curve reasonably well. In Figure 7c, the individual B-factors of streptavidin (a β-barrel protein), refined at 2Å resolution (Pahler et al., 1989) are compared with the results of the TLS model. The characteristic pattern is reproduced well, though now there are a number of deviations, such as around residue 10, where a crystal contact damps out the individual B-factors, but the rigid-body model predicts high B-factors. In such cases, the TLS model can be used to provide a starting set of B-factors. Further cycles of refinement may be done by restraining the individual B-factors to be close to the TLS model, rather than restraining the deviations between the B-factors of bonded atoms, as is commonly done today (Konnert & Hendrickson, 1980).

The TLS B-factors shown in Fig. 11 were all obtained by treating the previously refined inidividual isotropic B-factors as the target data for the optimizations. We have refined TLS B-factors for myoglobin using the experimental X-ray structure factors as the target data, and we find that the resulting values are similar to those obtained by refinement against individual B-factors. The agreement with individual B-factors is comparable to that seen for streptavidin (Fig. 11c).

Conclusions:

The results of the "twin" refinement clearly show that there is information present in the diffraction data that is currently neglected in protein refinements at high resolution. The formidable challenge is to develop simplified models that capture the essential features of protein dynamics and yet are frugal in their use of free parameters. The powerful tool of simulated annealing is expected to play an important role in this research. At the other end of the resolution range, the TLS model is likely to provide useful starting points for isotropic temperature factor refinement. An important question is whether the success of the TLS model reflects the fact that rigid-body motions dominate the B-factors, or whether it is merely a good way of parametrizing the dependence of the B-factors on the distance from the center of mass. Diffuse scattering experiments (Doucet and Benoit, 1987; Caspar et al., 1988) may shed light on this issue.

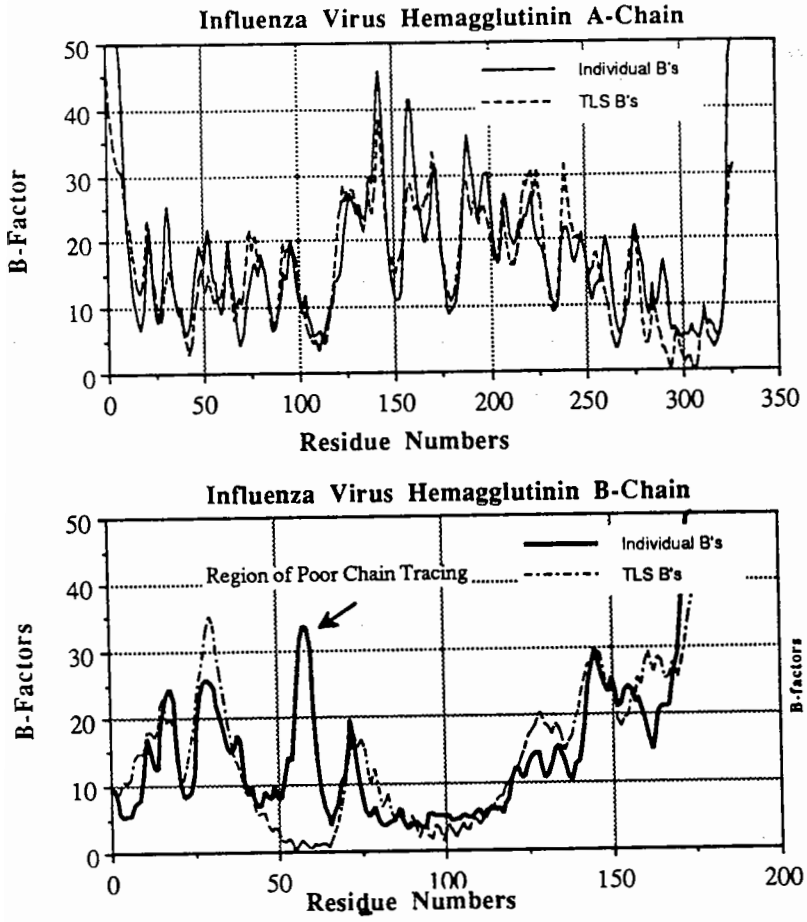


Figure 8. (a,top) Comparison of individual atomic isotropic B-factors (refined against 3Å X-ray data by Knossow et al., 1986) and TLS B-factors for the A-chain of hemagglutinin.The **(b,middle)** Same, for B-chain. A single set of TLS parameters was used to refine TLS parameters for both chains. **(c,bottom)** Same, for one monomer of streptavidin. The individual B-factors were refined against 2Å data (Pahler and Hendrickson, 1989).

Acknowledgements:

The work on the "twin" refinements of ribonuclease and crambin is being done in collaboration with Stephen K. Burley, Axel T. Brunger, Martin Karplus and Wayne A. Hendrickson, and they are thanked for many helpful suggestions. The computations described have been done on the Cray XMP at the Pittsburgh Supercomputer Center (grant no. DMB-880018P). J.K is supported by the Andre Meyer University Fellowship.

References:

1. Borkakoti, N., Moss, D.S. & Palmer, R.A. (1982)*Acta Crystallogr., Sect. B*, **38**, 2210-2217.
2. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., Swaminathan, S. & Karplus, M. (1983) *J. Comput. Chem.*, **4**, 187-217.
3. Brunger, A.T., Clore, G.M., Gronenborn, A.M.& Karplus, M. (1986) *Proc. Natl. Acad. Sci. (USA)*, **83**, 3801-3805.
4. Brunger, A.T., Kuriyan, J. and Karplus, M. (1987) *Science*, **235**, 458-460.
5. Brunger, A.T. (1988b) *X-PLOR (Version 1.5) Manual* , The Howard Huges Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University 260 Whitney Avenue, New Haven, CT 06511
6. Caspar, D.L.D., Clarage, J., Salunke, D.M. & Clarage, M. (1988) *Nature*, **332**, 659-662.
7. Doucet, J. and Benoit, J.P. (1987) *Nature* , 325, 643-646.
8. Hendrickson, W.A. & Teeter, M.M. (1981) *Nature (London)*, **290**, 107-113.
9. Ichiye, T. and Karplus, M. (1987) *Proteins: Structure,Function and Genetics*, **2**,236-259.
10. Ichiye, T. and Karplus, M. (1988) *Biochemistry*, **27**,3487-3497.
11. Knossow, M. Lewis, M., Rees, D., Wilson, I.A., Skehel, D. & Wiley, D.C. (1986) *Acta Crystallogr., Sect. B.*, **42**, 627-632.
39. Konnert, J. and Hendrickson, W.A. (1980) *Acta Crystalogr. Sect. A* , **36**,344-349.
12. Kuriyan, J., Petsko, G.A., Levy, R.M. and Karplus, M. (1986a) *J. Mol. Biol.*, **190** , 227-254.
13. Kuriyan, J. Wilz, S., Karplus, M. and Petsko, G.A. (1986b) *J. Mol. Biol.*, **192**, 133-154.
14. Kuriyan, J., Petsko, G.A. and Karplus, M. (1987) *Proteins: Structure, Function and Genetics*, **2**,1-12.
15. Kuriyan, J., Brunger, A.T., Karplus, M. and Hendrickson, W.A. (1989) *Acta Crystallographica, Sect. A*, in press.
16. Mao, B., Pear, M.R. and McCammon, J.A. (1982) *Biopolymers*, **21**,1979-1989.
17. Schomaker, V. and Trueblood, K.N. (1968) *Acta Crystallogr., Sect. B.*, **24**, 63-76.
18. Smith, J.L., Hendrickson, W.A., Honzatko, R.B. and Sheriff, S. (1986) *Biochemistry* , **25**, 5018-5027.
19. Sternberg, M.J.E., Grace, D.E.P., Phillips, D.C. (1979) *J. Mol. Biol.*, **130**, 231-253.
20. Svensson, L.A., Sjolin, L., Gilliland, G.L., Finzel, B.C. and Wlodawer, A. (1986) *Proteins: Structure, Function, and Genetics*, **1**, 370-375.
21. Swanson, S.M. (1988) *Acta Crystallogr., Sect. A*, **44**, 437-442.
22. Wlodawer, A., Borkakoti, N., Moss, D.S. and Howlin, B.(1986) *Acta Crystalogr. Sect. B*, **42**, 379-387.
23. Wlodawer, A., Svensson, L.A., Sjolin, L.and Gilliland, G.L.(1988) *Biochemistry*, **27**, 2705-2717.
24. Wuthrich, K. (1989) *Science*, **243**, 45-50.
25. Yu, H., Karplus, M. and Hendrickson, W.A. (1985) *Acta Crystallogr.sect. A*, **38**, 563-568.

An assessment of the program XPLOR as a tool for structure refinement at initial and final stages.

by

E.J. Dodson and J.P. Turkenburg
Department of Chemistry, University of York, Heslington
York YO1 5DD England

Introduction:

The refinement system XPLOR (Brunger, Kuryan and Karplus, 1987) uses energy minimisation and molecular dynamics in conjunction with diffraction data to improve the coordinates of a protein structure. XPLOR version 1.5 has been implemented on the CRAY X-MP/48 at Rutherford.

In order to assess the system, and to become conversant with its conventions, a study has been carried out on an already refined protein, bacterial ribonuclease (Sevcik et. al.). This is a molecule with two chains, each of 96 residues, which has been refined at 1.9A to a conventional crystallographic r-factor of 17%.

Cycles of refinement incorporating simulated annealing (so-called heating and cooling stages) have been run on the initial set of atomic coordinates obtained from the isomorphously phased map. These coordinates were fairly inaccurate, and contained some gross errors as well as some omissions. The procedure has improved the main chain conformation greatly in terms of rms differences with the final refined structure, although it was less successful for the side chains

Refinement employing simulated annealing has also been carried out on the final coordinate set to see how much the structure was altered by substituting the geometric restraints of PROLSQ (Konnert, 1976) by the energy parameters of XPLOR.

Crystallographic details for Ribonuclease SA:

Cell 64.9 78.32 38.79 90 90 90
Spacegroup P212121
Two molecules/asymmetric unit, each of 96 residues
Solvent fraction 0.48
R factor for 1747 atoms, 1495 protein, 252 waters 17.4%
Resolution 1.9A

Data collection
Native - Synchrotron 17202 reflections merging R 0.056

The two molecules of ribonuclease SA were built into the 2.5A isomorphously phased and solvent flattened density independently. After 7 or 8 rebuildings the structure was virtually complete. This refined structure has been used as a reference for the work

done with XPLOR. Any refinement procedure for a crystal structure should be assessed by an examination of the final difference density. This is a much more sensitive measure of a correct structure than any quoted R factor. The map based on the final coordinates was virtually flat.

Methodology:

A number of protocols have been tested on this initial model.

Protocol 1. The standard procedure as given in the Xplor manual (Brunger, 1988)

Stages:

1) Prepare: 40 cycles of energy minimisation including Xray " energy term" - harmonic restraints on CA's
2) Heat to 3000 K; time step .0005 ps; 1000 integration steps. Rescaling of velocities every 250 steps.
3) Cooling to 300K; time step 0.0005 ps; 500 integration steps.
4) Final stage; energy minimisation -80 steps- no CA restraints.
5) Coordinates transferred back to York VAX; 1 cycle of B factor refinement reduced R factor by 3% to 38.2%
   Cray allocation units used: 85 Au's

Protocol 2.

Identical to protocol 1, except that the weight for the Xray data was decreased by 1%. This gives a very different answer, worse both in R factor and rms deviation from the final model. The average rms differences between this protocol and the manually refined structure for main chain are 0.91 and for side chains 1.9

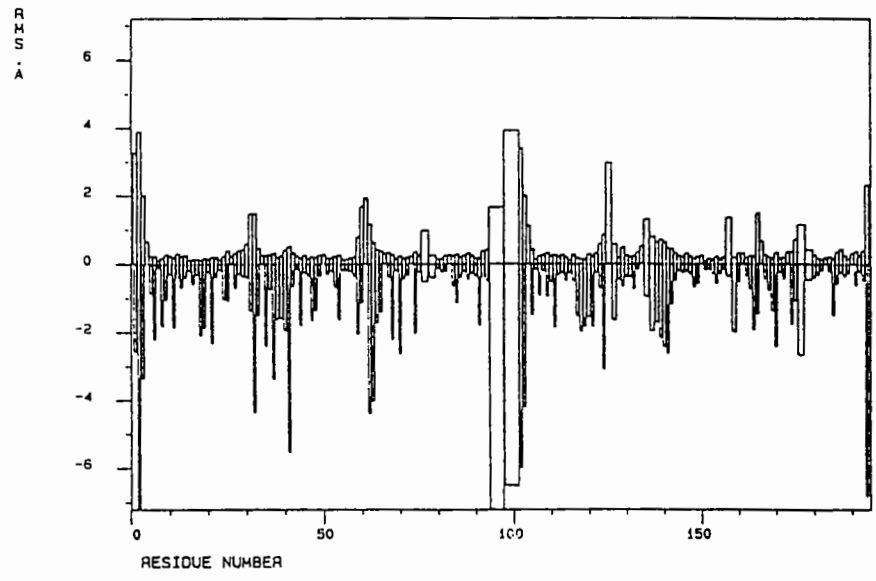


Fig. 1. Root mean square differences per residue (in Angstrom) between Xplor output and the manually refined structure. The wider peaks are the last and first residue of the A and B chain respectively. The upper half is for main chain atoms, the lower half for side chain atoms. The figures were prepared using the program SQUID (Oldfield) Protocol 1. The average rms differences are 0.78 for main chain and 1.9 for side chain atoms.

Protocol 3. To make sure the result of protocol 2 was not an artefact of the protein system with the given velocities and energy, protocols 1 and 2 were repeated with different seeds for the random assignment of the initial velocities. This ensures that the system will, at least initially, chose a different trajectory. The average rms differences between the modified protocol 1 and the manually refined structure for main chain are 1.2 and 2.0 for side chain. These values are the same for the modified protocol 2, but the individual residues are wildly different.

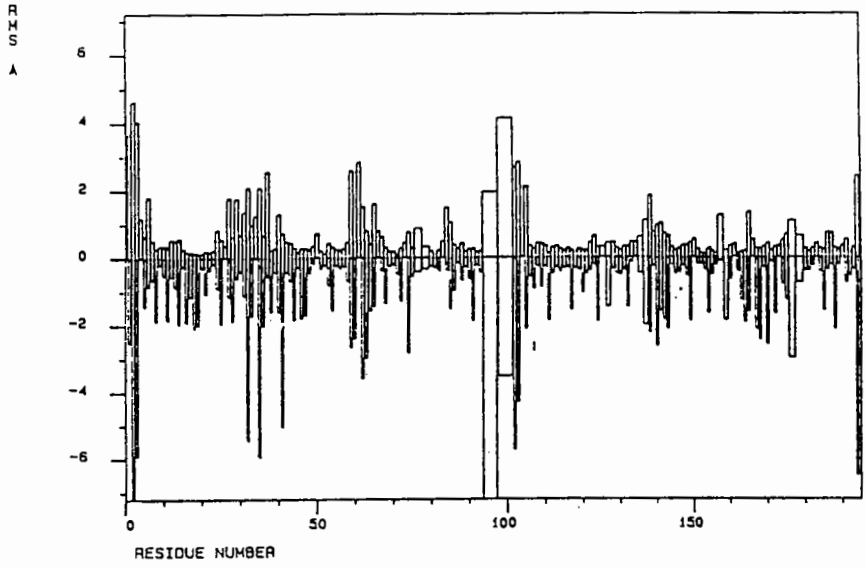


Fig. 2. Protocol 4. The average rms differences are 1.0 for main and 1.9 for side chain atoms. For description see Fig. 1.

Protocol 4. Slow cooling, described in the XPLOR manual.
Stages:
1) Prepare: cycles of energy minimisation including Xray "energy term". Harmonic restraints on CA's. No phase information included. No heat stage.
2) Slow cooling; Starting temperature of heatbath 4000K; reduce temperature by 40K every 40 steps of the dynamic simulation. Final temperature 300K.
3) Final stage; energy minimisation; 80 steps. No CA restraints. See figure 2.

Protocol 5. The initial model of a structure is often incomplete in that (parts of) residues are missing. Those residues are normally built in using interactive computer graphics, where atoms are placed in positions which are calculated applying "ideal" bond lengths and angles. To assess whether it would be possible to do this with Xplor, the SHAKE (Ryckaert,

Cicotti and Berendsen) option was used. SHAKE contraints distances between atoms to certain reference distances, which can be taken from the parameters. Bearing in mind the large radius of convergence of Xplor, one would hope that in subsequent dynamics runs those atoms are moved to their real positions.

Stages:

1) Prepare: 40 cycles of energy minimisation including Xray " energy term" - harmonic restraints on CA's SHAKE applied to incomplete residues during minimisation.
2) Heat to 3000 K; time step .0005 ps; 1000 integration steps. Rescaling of velocities every 250 steps. No SHAKing.
3) Cooling to 300K; time step 0.0005 ps; 500 integration steps. No SHAKing
4) Final stage; energy minimisation; 100 steps. No CA restraints, no SHAKing. See figure 3.

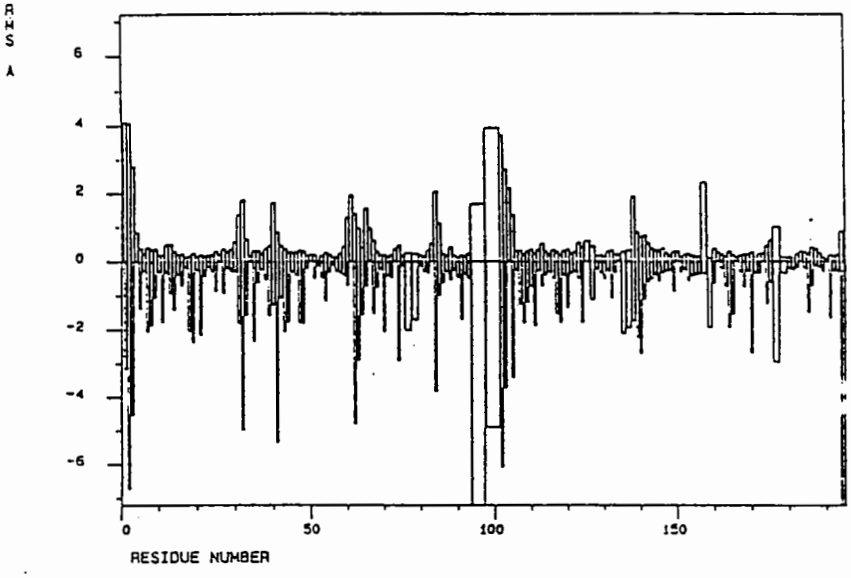


Fig. 3. Protocol 5. The average rms differences are 0.9 for main and 1.8 for side chain atoms. For description see Fig. 1.

Protocol 6. This protocol was set up to see what effect a more elaborate treatment of the system would have: the number of cycles in the initial minimisation was increased to 200 (was 40), during the heatstage the velocities were rescaled every 25 steps (was 250) and in the cooling stage coupling to a heatbath was employed. Finally, the usual 100 steps of minimisation were done.

Stages:

1) Prepare: 200 cycles of energy minimisation including Xray " energy term" - harmonic restraints on CA's.
2) Heat to 3000 K; time step .0005 ps; 1000 integration steps. Rescaling of velocities every 25 steps.
3) Cooling to 300K; coupling to heat bath, temperature decreased in steps of 25K; 50 steps of dynamics after each temperature step.
4) Final stage; energy minimisation; 100 steps. No CA restraints. See figure 4.

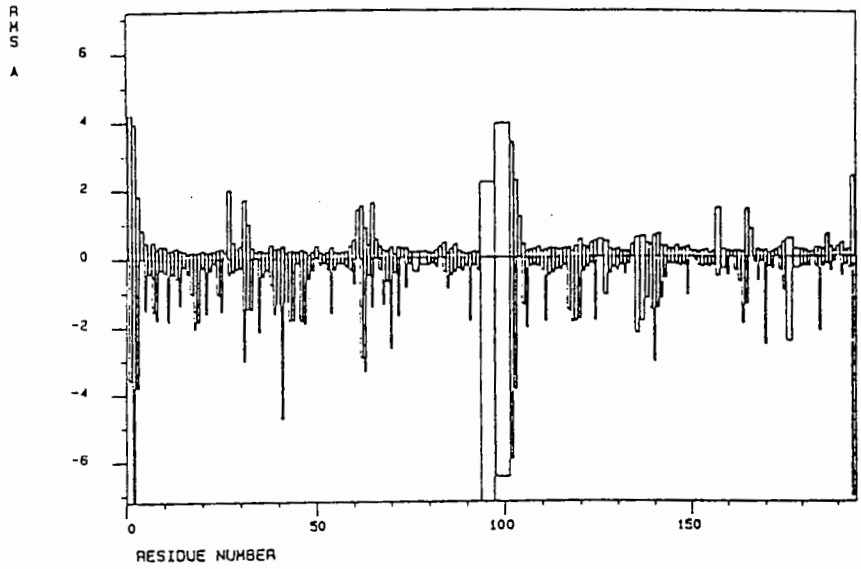


Fig. 4. Protocol 6. The average rms differences are 0.8 for main and 1.7 for side chain atoms. For description see Fig. 1.

Protocol 7. Explicit hydrogen bonds and other charges introduce forces that can either keep the system in a firm grip or make the system drift apart. An obvious remedy is to switch off the charges of Lys, Glu, Asp and Arg during molecular dynamics runs.

Stages:

1) Prepare: 200 cycles of energy minimisation including Xray " energy term" - harmonic restraints on CA's. Charges on.
2) Heat to 3000 K; time step .0005 ps; 1000 integration steps. Rescaling of velocities every 25 steps. Charges off.
3) Cooling to 300K; coupling to heat bath, temperature decreased in steps of 25K; 50 steps of dynamics after each temperature step. Charges off.
4) Final stage; energy minimisation; 100 steps. No CA restraints. Charges on.

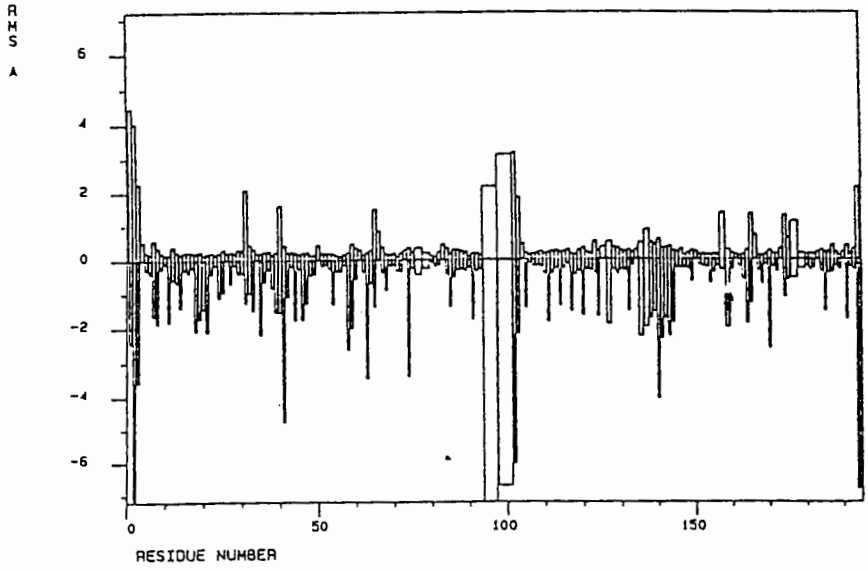


Fig. 5. Protocol 7. The average rms differences are 0.8 for main and 1.7 for side chain atoms. For description see Fig. 1.

Finally two protocols were tested on the coordinates as they were after traditional refinement. This included some 250 water molecules.

Protocol 8.
Stages:
1) Prepare: 40 cycles of energy minimisation including Xray " energy term" - harmonic restraints on CA's
2) Heat to 3000 K; time step .0005 ps; 1000 integration steps. Rescaling of velocities every 250 steps.
3) Cooling to 300K; time step 0.0005 ps; 500 integration steps.
4) Final stage; energy minimisation; 80 steps. No CA restraints.

Protocol 9.

The water molecules evaporated and moved by more than 10 Angstroms in some cases. To overcome this problem the water positions were restrained to their initial position (20 Kcal/(mole A**2)
Stages:
1) Prepare: 40 cycles of energy minimisation including Xray " energy term" - harmonic restraints on CA's and H2O.
2) Heat to 3000 K; time step .0005 ps; 1000 integration steps. Rescaling of velocities every 250 steps. Restraints on H2O.
3) Cooling to 300K; time step 0.0005 ps; 500 integration steps. Restraints on H2O.
4) Final stage; energy minimisation 80 steps; no CA restraints. Restraints on H2O.

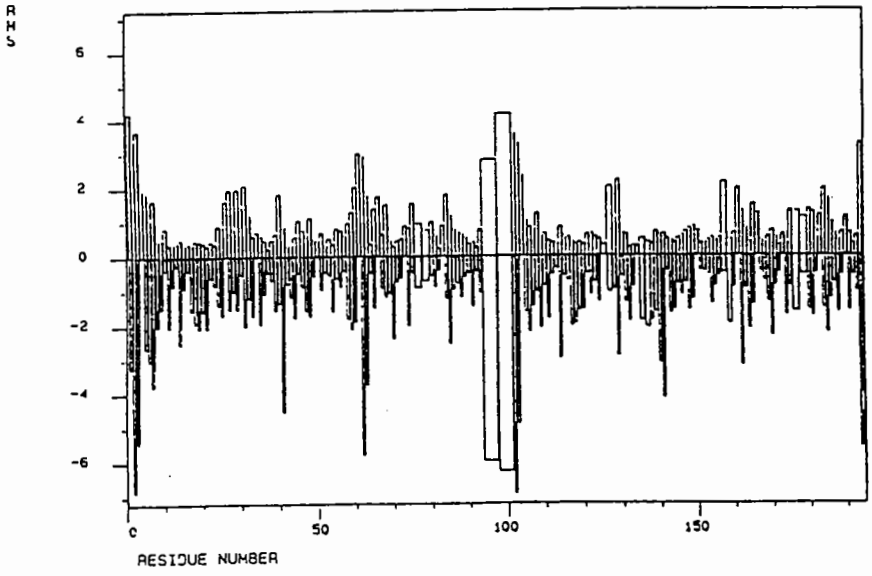


Fig. 6. For comparison the average rms differences between the initial model and the final, manually refined structure. This gives an indication of the shifts necessary. For description see Fig. 1.

Results and Discussion.

All protocols improved the main chain with shifts of up to 3.0 Angstroms. In particular 17 of the 34 main chain oxygens which were misplaced had corrected themselves automatically and the two CIS-prolines which created the most awful difficulties as TRANS conformations were forced into the density had regrouped themselves correctly.

Figure 6 shows the rms difference plot between the initial model and the final, manually refined structure. In fact, this illustrates the average shifts for each residue, going from the positions in the initial model to the "correct" position in the final structure. Figures 1 to 5 give the rms differences between several outputs of Xplor and the same manually refined structure. As can be seen from the rms plots in figure 1 to 5, the differences between protocols 1, 4, 6 and 7 are minor when taking into account the size of the shifts. It seems that for a system with a high starting R-factor a straightforward annealing procedure as described for protocol 1 is sufficient. It should be noted that for structures with R-factors as high as about 50% the system becomes unstable (Weiss, private communication). The extra cpu time used in a procedure like protocol 6 can hardly be justified by the minor improvements seen. Whether this is also true for a structure with a better starting model will be subject of further investigation.

The influence of a 1% difference in weight for the X-ray terms (protocol 2) was totally unexpected. Apparently the system follows a different trajectory and ends up in a different minimum. It might well be that there would be convergence after longer dynamics runs. Protocol 3 shows that the same happens when using a different random number seed, so it is not coincidental.

Employing the SHAKE option (protocol 5) to add missing atoms appears to give good results. This approach should also be tested in the case of residues that are missing altogether. It might be particularly useful when there is a poor homology for a molecular replacement solution and interactive model building would take a long time.

The evaporation of the water molecules in protocol 8 is obviously caused by the high temperatures employed in the dynamics, although it should be noted that the temperature has no physical meaning, but is rather a measure for the ability of the system to overcome energy barriers. Harmonic restraints as used in protocol 9 solve this problem. The actual rms differences for the main chain are very small in both cases. In protocol 8 the side chains move into the density of the evaporated waters, in protocol 9 this is not the case and the rms differences are very small for the side chains as well. This implies that there are no fundamental differences between the energy restraints in Xplor and the restraints in Prolsq.

The Fobs phase combined map calculated from the set of coordinates from protocol 1 was a great improvement on that calculated at the end of the first conventional refinement run, and it should have been possible to rebuild the structure more quickly, and to do that job fewer times; ie the procedure of refinement would have been speeded up greatly.

References

Brunger, A.T., Kuryan, J., and Karplus, M. (1987) Science 23 485

Brunger, A.T. (1988) "Xplor Manual (version 1.5)" Yale University

Oldfield, T.J. manuscript in preparation.

Ryckaert, J.-P., Cicotti, G., and Berendsen, H.J.C. (1977) J. Comput. Phys. 23 327

Sevcik, J., Dodson, E.J., Dodson, G.G., and Zelinka, J. manuscript in preparation.

List of Delegates

Dr	A	Achari	Genex Corporation, 16020 Industrial Drive, Gaithersburg, Maryland 20877, U.S.A.
Dr	R	Acharya	Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.
Mr	P D	Adams	Department of Biochemistry, University of Edinburgh, Hugh Robson Building, George Square, Edinburgh EH8 9XD.
Dr	M J	Adams	Laboratory of Molecular Biophysics, Rex Richards Building, University of Oxford, South Parks Road, Oxford OX1 3QU.
Dr	I A	Andersson	Department of Molecular Biology, Swedish University of Agricultural Sciences, Box 590, S-75124 Uppsala, Sweden.
Dr	P	Artymiuk	Department of Biochemistry, University of Sheffield, Sheffield S10 2TN.
Mr	A	Atayasiadis	Research Centre, Institute of Molecular Biology and Biotechnology, University of Crete, PO Box 1527, 71 110 Heraklion, Crete, Greece.
Ms	C B	Baguley	Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.
Dr	S	Bailey	Department of Physics, University of Keele, Keele, Staffs ST5 5BG.
Dr	P J	Baker	Department of Biochemistry, Division of Molecular Biology, University of Sheffield, Sheffield S10 2TN.
Dr	D W	Banner	Hoffmann-La Roche & Co., ZFE 65/102, Grenzacherstrasse, 4002 Basle, Switzerland.

Dr	D	Barford	Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.
Dr	J O	Baum	MRC, c/o Atlas Computer Centre, Rutherford Appleton Laboratory, Chilton, Didcot, Oxon OX11 0QX.
Mr	A J	Beavil	Department of Biophysics, Cell and Molecular Biology, King's College London, 26-29 Drury Lane, London WC2.
Dr	G	Bentley	Immunologie Structurale, Institut Pasteur, 25 rue du Dr. Roux, 75724 Paris, France.
Dr	T N	Bhat	Immunologie Structurale, Institut Pasteur, 28 rue du Dr. Roux, 75724 Paris Cedex, France.
Dr	A C	Bloomer	Laboratory of Molecular Biology, MRC, Hills Road, Cambridge CB2 2QH.
Prof.	T L	Blundell	Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Dr	M C	Bolognesi	Department of Crystallography, University of Pavia, Via Taramelli,16, I-27100 Pavia, Italy.
Prof.	C I	Branden	Department of Molecular Biology, Swedish University of Agricultural Sciences, Box 590, S-75124 Uppsala, Sweden.
Dr	P	Brick	Blackett Laboratory, Imperial College, Prince Consort Road, London SW7 2BZ.
Ms	K L	Britton	Department of Biochemistry, University of Sheffield, Western Bank, Sheffield S10 2TN.

Dr	J W	Campbell	Daresbury Laboratory.
Prof.	C R A	Catlow	Department of Chemistry, University of Keele, Keele, Staffs ST5 5BG.
Mr	L	Caves	Department of Chemistry, University of York, Heslington, York YO1 5DD.
Dr	C A	Chapman	Academic Press Ltd., 24-28 Oval Road, London NW1 7DX.
Dr	D	Clark	Chemical Physics Division, Institute of Food Research, Norwich Laboratory, Colney Lane, Norwich NR4 7VA.
Dr	A	Cleasby	Laboratory of Molecular Biology, University of Oxford, Rex Richards Buildings, South Parks Road, Oxford OX1 3QU.
Mr	I J	Clifton	Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.
Dr	S W	Cowan	Department of Molecular Biology, Biomedical Centre, University of Uppsala, Box 590, S-751 24 Uppsala, Sweden.
Dr	Z	Dauter	EMBL, c/o DESY, Notkestrasse 85, D-2000 Hamburg 52, Fed Rep Germany.
Mr	G	Davies	Department of Biochemistry, University of Bristol, University Walk, Bristol BS8 1TD.
Mrs	U	Derewenda	Department of Chemistry, University of York, Heslington, York YO1 5DD.

Dr	Z	Derewenda	Department of Chemistry, University of York, Heslington, York YO1 5DD.
Dr	R	Diamond	Laboratory of Molecular Biology, MRC, Hills Road, Cambridge CB2 2QH.
Prof.	O	Dideberg	Cristallographie, Institut de Physique B5, Universite de Liege au Sart-Tilman, B-4000 Liege, Belgium.
Dr	B W	Dijkstra	Laboratory of Chemical Physics, University of Groningen, Nijenborgh 16, 9747 AG Groningen, The Netherlands.
Ms	E J	Dodson	Department of Chemistry, University of York, Heslington, York YO1 5DD.
Dr	H P C	Driessen	Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Dr	E	Eliopoulos	Department of Biophysics, University of Leeds, Leeds LS2 9JT.
Mr	P	Emsley	Department of Chemistry, University of York, Heslington, York YO1 5DD.
Mr	R M	Esnouf	Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.
Dr	P R	Evans	Laboratory of Molecular Biology, MRC, Hills Road, Cambridge CB2 2QH.
Ms	J	Everett	Department of Physics, Liverpool Polytechnic, Byrom Street, Liverpool L3 3AF.

Dr	D	Fincham	Daresbury Laboratory.
Dr	T O	Fischmann	Immunologie Structurale, Institut Pasteur, 28 rue du Dr. Roux, 75724 Paris Cedex 15, France.
Mr	T	Flores	Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Dr	C M	Frazao	Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Dr	P S	Freemont	Imperial Cancer Research Fund, LIF 619, 44 Lincolns Inn Fields, London WC2A 3PX.
Dr	D M	Frey	Laboratory of Biological Crystallography, CNRS, Marseille, France.
Ms	E E	Fry	Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.
Dr	S	Gamblin	Department of Biochemistry, University of Bristol, University Walk, Bristol BS8 1TD.
Dr	E F	Garman	Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.
Dr	A	Geddes	Department of Biophysics, University of Leeds, Leeds LS2 9JT.
Mr	M R	Gibbs	Laboratory of Molecular Biology, MRC Centre, Hills Road, Cambridge CB2 2QH.

Mr	P	Gilchrist	Department of Biochemistry, University of Edinburgh, Hugh Robson Building, George Square, Edinburgh EH8 9XD.
Prof.	M J	Gillan	Department of Physics, University of Keele, Keele, Staffs ST5 5BG.
Dr	J	Goodfellow	Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Dr	S	Gover	Laboratory of Molecular Biophysics, Rex Richards Building, University of Oxford, South Parks Road, Oxford OX1 3QU.
Mr	P	Gros	Laboratory of Chemical Physics, University of Groningen, Nijenborgh 16, 9749 AG Groningen, The Netherlands.
Miss	A T	Hadfield	Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Buildings, South Parks Road, Oxford OX1 3QU.
Dr	J	Hajdu	Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.
Dr	I	Haneef	Astbury Department of Biophysics, University of Leeds, Leeds LS2 9JT.
Dr	L C	Hansen	Department of Protein Crystallography, University of Tromso, P O Box 953, N 9001 Tromso, Norway.
Dr	M M	Harding	Department of Chemistry, University of Liverpool, PO Box 147, Liverpool L69 3BX.
Dr	M	Harel	Department of Structural Chemistry, Weizmann Institute of Science, 76100 Rehovot, Israel.

Mr	M J	Hartshorn	Department of Chemistry, University of York, Heslington, York YO1 5DD.
Mr	T	Harvey	Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU.
Dr	S	Hasnain	Daresbury Laboratory.
Prof.	J	Helliwell	Department of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL.
Dr	K	Henrick	Daresbury Laboratory.
Dr	R	Hilgenfeld	Central Research/Biotechnology, Building G864, Hoechst AG, P.O.B.800 320, D-6230 Frankfurt 80, Fed Rep Germany
Prof.	W G J	Hol	Laboratory of Chemical Physics, University of Groningen, Nyenborgh 16, 9747 AG Groningen, The Netherlands.
Mr	P	Holden	Department of Chemistry, University of York, Heslington, York YO1 5DD.
Dr	E	Hough	Department of Protein Crystallography, University of Tromso, PO Box 953, N9001 Tromso, Norway.
Dr	R E	Hubbard	Department of Chemistry, University of York, Heslington, York YO1 5DD.
Dr	W N	Hunter	University Chemical Laboratory, University of Cambridge, Lensfield Road, Cambridge CB2 1EW.

Prof.	N W	Isaacs	Department of Chemistry, University of Glasgow, Glasgow G12 8QQ.
Dr	J	Jaegar	Department of Structural Biology, Biocenter, University of Basel, Klingelbergstrasse 70, CH-4056, Basel, Switzerland.
Mr	P	Jeffrey	Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.
Mr	H	Jhoti	Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Mr	M D	Jordan	Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.
Mrs	L	Joshua-Tor	Department of Structural Chemistry, Weizmann Institute of Science, 76100 Rehovot, Israel.
Mr	G J	Keen	Department of Biophysics, Cell and Molecular Biology, King's College London, 26-29 Drury Lane, London WC2.
Ms	S	Klupsch	EMBL, c/o DESY, Notkestrasse 85, D-2000 Hamburg 52, Fed Rep Germany.
Prof.	M	Kokkinidis	Research Centre, Institute of Molecular Biology and Biotechnology, University of Crete, PO Box 1527, 71 110 Heraklion, Crete, Greece.
Mr	P	Kraulis	Department of Molecular Biology, Biomedical Centre, University of Uppsala, Box 590, S-751 24 Uppsala, Sweden.
Dr	J	Kuriyan	The Rockefeller University, 1230 York Avenue, New York, N Y 10021, U S A.

Mr	R	Lapatto	Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Dr	J	Li	M.R.C. Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH.
Dr	P	Lindley	Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Dr	J A	Littlechild	Department of Biochemistry, University of Bristol, Bristol, Avon BS8 1TD.
Mr	D T	Logan	Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.
Dr	P R	Mallinson	Department of Chemistry, University of Glasgow, Glasgow G12 8QQ.
Mr	D J	Matthews	Blackett Laboratory, Imperial College, Prince Consort Road, London SW7 2BZ.
Mr	A S	McAlpine	Department of Biochemistry, University of Edinburgh, Hugh Robson Building, George Square, Edinburgh EH8 9XD.
Dr	P	McLaughlin	Laboratory of Molecular Biology, MRC Centre, Hills Road, Cambridge CB2 2QH.
Dr	C	Medina	Symbicom, Glunten, 751 83 Uppsala, Sweden.
Mr	J B O	Mitchell	Department of Theoretical Chemistry, University Chemical Laboratory, University of Cambridge, Lensfield Road, Cambridge CB2 1EW.

Dr	P C E	Moody	Department of Chemistry, University of York, Heslington, York YO1 5DD.
Dr	M H	Moore	University Chemical Laboratory, University of Cambridge, Lensfield Road, Cambridge CB2 1EW.
Mr	J H	Morais Cabral	Department of Biochemistry, University of Edinburgh, Hugh Robson Building, George Square, Edinburgh EH8 9XD.
Dr	P	Murray-Rust	Glaxo Group Research Ltd, Greenford Road, Greenford, Middlesex UB6 0HE
Mr	F W	Muskett	Department of Biochemistry, University of Edinburgh, Hugh Robson Building, George Square, Edinburgh EH8 9XD.
Dr	C	Nave	Daresbury Laboratory.
Mr	M P	Newman	Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Dr	M	Nilges	Laboratory of Chemical Physics, NIDDK, National Institutes of Health, Building 2, Bethesda MD 20892, USA.
Prof.	A C T	North	Department of Biophysics, University of Leeds, Leeds LS2 9JT.
Mr	B	O'Hara	Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Mr	T J	Oldfield	Department of Chemistry, University of York, Heslington, York YO1 5DD.

Dr	G	Oliva	c/o Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Ms	S	Onesti	Blackett Laboratory, Imperial College, Prince Consort Road, London SW7 2BZ.
Ms	K	Phillips	Department of Biophysics, University of Leeds, Leeds LS2 9JT.
Dr	R W	Pickersgill	AFRC, Institute of Food Research, Reading Laboratory, Reading RG2 9AT.
Dr	L	Potterton	Polygen, Department of Chemistry, University of York, Heslington, York YO1 5DD.
Dr	S L	Price	University Chemical Laboratory, University of Cambridge, Lensfield Road, Cambridge CB2 1EW.
Dr	J W	Quail	Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Dr	A	Quick	Computer Centre, University of London, 20 Guildford Street, London WC1N 1DZ.
Dr	J	Raftery	Biological N M R Centre, Medical Sciences Building, University of Leicester, Leicester.
Mr	A	Raine	Department of Chemistry, University of York, Heslington, York YO1 5DD.
Mr	J	Ren	Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.

Dr	D W	Rice	Department of Biochemistry, University of Sheffield, Western Bank, Sheffield S10 2TN.
Dr	P J	Rizkallah	Department of Chemistry, University of Liverpool, PO Box 147, Liverpool L69 3BX. and Daresbury Laboratory.
Ms	H F	Rodgers	Department of Biochemistry, University of Sheffield, Western Bank, Sheffield S10 2TN.
Dr	M	Saqi	Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Mr	R	Sarra	Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Dr	H F G	Savage	Department of Chemistry, University of York, York YO1 5DD.
Dr	L	Sawyer	Department of Biochemistry, University of Edinburgh, George Square, Edinburgh EH8 9XD
Dr	H	Scouloudi	Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.
Dr	T	Skarzynski	Blackett Laboratory, Imperial College, Prince Consort Road, London SW7 2BZ.
Mr	O S	Smart	Blackett Laboratory, Imperial College, Prince Consort Road, London SW7 2BZ.
Mr	W	Somers	Department of Biophysics, University of Leeds, Leeds LS2 9JT.

Dr	T J	Stillman	Department of Biochemistry, University of Sheffield, Western Bank, Sheffield S10 2TN.
Dr	M J	Sutcliffe	Inorganic Chemical Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QR.
Dr	B J	Sutton	Department of Biophysics, King's College London, 26-29 Drury Lane, London WC2B 5RL.
Miss	H J	Swift	Department of Chemistry, University of York, Heslington, York YO1 5DD.
Dr	G L	Taylor	Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.
Dr	P	Taylor	Department of Biochemistry, University of Edinburgh Medical School, Hugh Robson Building, George Square, Edinburgh EH8 9XD.
Ms	N	Thanki	Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Ms	P J	Thomas	Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Mr	D H	Thomas	Department of Biochemistry, University of Sheffield, Western Bank, Sheffield S10 2TN.
Dr	R	Todd	Laboratory of Molecular Biology, MRC Centre, Hills Road, Cambridge CB2 2QH.
Mr	J P	Turkenburg	Department of Chemistry, University of York, Heslington, York YO1 5DD.

Prof.	W F	Van Gunsteren	Department of Physical Chemistry, University of Groningen, Nijenborgh 16, 9747 AG Groningen, The Netherlands.
Mr	C S	Verma	Department of Chemistry, University of York, Heslington, York YO1 5DD.
Dr	B	Vessal	Department of Chemistry, University of Keele, Keele, Staffs ST5 5RG.
Dr	N	Walker	BASF AG, ZHV/W-A30, D-6700 Ludwigshafen, Fed Rep Germany.
Mr	P A	Walker	Department of Biochemistry, University of Bristol, Bristol BS8 1TD.
Dr	A	Walton	Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Dr	Z-L	Wan	Department of Chemistry, University of York, Heslington, York YO1 5DD.
Dr	H C	Watson	Department of Biochemistry, University of Bristol, Bristol BS8 1TD
Dr	W	Weis	Department of Molecular Biophysics and Biochemistry, Yale University, 260 Whitney Avenue, PO Box 6666, New Haven, Connecticut 06511, U.S.A.
Dr	D	Wigley	Department of Biochemistry, University of Leicester, University Road, Leicester LE1 7RH.
Dr	K S	Wilson	EMBL, c/o DESY, Notkestrasse 85, D-2000 Hamburg 52, Fed Rep Germany.

Mr	R	Young	Department of Biophysics, King's College London, 26-29 Drury Lane, London WC2.
Dr		Zaitsev	c/o Dr P Lindley, Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.
Mr	K J Y	Zhang	Department of Physics, University of York, Heslington, York YO1 5DD.