THE STRUCTURE AND FLEXIBILITY OF MYOGLOBIN: MOLECULAR DYNAMICS AND X-RAY CRYSTALLOGRAPHY

by

John Kuriyan

B.S., Juniata College

Huntingdon, PA., 1981

Submitted to the Department of Chemistry in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Physical Chemistry at the Massachusetts Institute of Technology

February, 1986

Copyright (c) 1986 Massachusetts Institute of Technology

Signature of Author Signature redacted
Department of Chemistry
February 17, 1986

Certified by Signature redacted
Prof. Gregory A. Petsko
Thesis Supervisor

Certified by Signature redacted
Prof. Martin Karplus
Thesis Supervisor
Harvard University

Accepted by Signature redacted
Prof. Glenn A. Berchtold
Chairman, Departmental Graduate Committee

MASSACHUSETTS INSTITUTE
OF TECHNOLOGY

JUN 05 1986

LIBRARIES

ARCHIVES

MIT Libraries

77 Massachusetts Avenue
Cambridge, MA 02139
http://libraries.mit.edu/ask

DISCLAIMER NOTICE

MISSING PAGE(S)

p.244

This doctoral thesis has been examined by a Committee of the Department of Chemistry as follows:

Professor Robert A. Alberty Signature redacted
Chairman

Professor Gregory A. Petsko Signature redacted
Thesis Supervisor

Professor Martin Karplus Signature redacted
(Harvard University) Thesis Supervisor

Professor Wayne A. Hendrickson Signature redacted
(Columbia University) Outside Examiner

THE STRUCTURE AND FLEXIBILITY OF MYOGLOBIN: MOLECULAR DYNAMICS AND X-RAY CRYSTALLOGRAPHY

Research Supervisors: John Kuriyan

Gregory A. Petsko (M.I.T.)

Martin Karplus (Harvard)

February, 1986

ABSTRACT

Molecular dynamics (M.D.) simulations of myoglobin are used to evaluate the effect of anisotropy and anharmonicity on the results of X-ray refinement of protein structures. Coordinates sampled from M.D. trajectories are used to calculate time-averaged X-ray structure factors. These structure factors are then treated as experimental data and the positions and temperature factors of all the atoms in the protein are refined against these data. The results are compared with the exact average positions and temperature factors obtained from the simulation. It is found that the refinement consistently underestimates the temperature factors for atoms with large mobility. Such atoms have multiple peaks in their probability distribution functions and the refinement fits only the major peak(s), leading to errors in the refined position and mean-square fluctuation.

The thermal expansion of myoglobin is studied by comparing the structure of the protein at 80K and at room temperature. Relatively large errors in the low temperature data complicate the comparison. Nevertheless, it is shown that interatomic distances expand by about 1%. Some of the expansion is correlated with changes in the unit cell of the crystal between the two temperatures.

The X-ray structure of CO-myoglobin has been refined at a resolution of 1.5Å. The CO binds to the iron in more than one conformation. The binding of CO causes larger changes in the structure of the protein than the binding of oxygen or water. The iron in CO-myoglobin is in the plane of the heme, having moved by 0.4Å relative to its out of plane position in deoxy-myoglobin. This motion of the iron is followed by the proximal histidine, resulting in large

TABLE OF CONTENTS

To my parents and my aunt,

Kuriyan John Kuriyan,
Anna Kuriyan,
and
Anna Mani.

Chapter 1

Protein Refinement and Empirical Energy Functions: An Introduction to the Thesis

Abstract

The first, and major, part of this chapter consists of an introduction to the refinement of protein structures against X-ray data. Various real-space and reciprocal-space refinement methods are described and compared. The problems remaining in the field are identified. In the second part of the chapter the use of empirical energy functions to simulate the structure and dynamics of proteins is outlined. Finally, the four subsequent chapters of the thesis are summarized.

Proteins, which are linear polymers of amino acids, serve as the messengers, controllers, catalysts, storehouses, transporters and structural elements of the living cell. A typical bacterial cell contains about ten thousand different kinds of proteins, ranging in size from small messenger hormones of twenty or thirty amino acids to complexes of catalytic enzymes or structural proteins made up of several thousand amino acids (Zubay, 1983, Alberts et al., 1983). Each kind of protein consists of a unique sequence of amino acids; this sequence is coded for by a corresponding sequence of nucleic acids in the DNA of the gene (Dickerson and Geis, 1969, 1983, Watson, 1976).

Most proteins spontaneously fold into well defined three-dimensional structures under physiological conditions; the ability of proteins to function is usually related to their ability to form these specific structures (Dickerson and Geis, 1969, 1983). There is a beautiful economy to this plan, which allows for the generation of these diverse and functionally efficient three-dimensional structures by specifying only a simple one-dimensional code. Two challenges in molecular biophysics today are to understand how the linear amino acid sequence determines the specific three-dimensional structure and how this structure is related to the specific function of a protein. Knowledge of the three-dimensional structures of proteins is, of course, essential to solving both problems.

The most intensely studied proteins are the globular proteins, which fold up to form compact structures. Many globular proteins can be induced to crystallize (McPherson, 1982), despite the fact that these molecules are large and irregularly shaped, and rarely have any internal symmetry at all. Protein single-crystals are used to obtain X-ray diffraction data, from which the three-dimensional structure of the molecule can be inferred (North and Phillips, 1969); the first such structure determination was for the oxygen storage protein, myoglobin, in 1957 (Kendrew et al., 1958,1960). Since then the structures of about two hundred globular proteins have been determined and these include various hormones, enzymes, storage proteins, transport proteins and DNA binding proteins.

It is becoming clear that understanding the relationship between the structure of a protein and its function requires that the static X-ray structure be complemented by information about the internal mobility and forces in proteins (Huber, 1979, Gurd and Rothgeb, 1979, Karplus and McCammon, 1981, 1983, Levitt, 1982, Debrunner and Frauenfelder, 1982, Clementi and Sharma (eds.), 1983, Ciba Foundation, 1983, Petsko and Ringe, 1984, Karplus, 1985, Hermans (ed.), 1985, Stuart and Phillips, 1985). This information can be obtained from simulation techniques such as molecular dynamics, Monte Carlo or normal mode calculations (all of which use empirical energy functions) and from various experimental methods such as NMR (Lipari et al.,1982), Raman and IR spectroscopy (Hilinski and Rentzepis, 1983, Friedman, 1985), fluorescence depolarization measurements (Lakowicz et al., 1983), inelastic neutron scattering (Smith et al., 1986), hydrogen exchange (Woodward and Hilton, 1979) and Mossbauer spectroscopy (Keller and Debrunner, 1980, Knapp et al., 1983).

This thesis is concerned with the problem of extracting such information from X-ray diffraction data. Since the motion of atoms in the protein crystal affects the scattering of X-rays, it is possible to obtain information about atomic dynamics from diffraction data (Willis and Pryor, 1975, Petsko and Ringe, 1984, Stuart and Phillips, 1985). Some information about the forces between atoms in the protein can also be obtained by studying the response of the structure to small perturbations such as ligand binding and changes in temperature. As described below, this requires a finer level of analysis of the X-ray data than is required to obtain the first, approximate, picture of the structure of the molecule.

Protein structure determination by X-ray crystallography can be thought of as proceeding in two stages. The phases of the measured reflections are estimated and a low-resolution structure is obtained in the first stage. More precise information about the structure is derived in the second stage by refining the parameters of a molecular model against the X-ray data (Jensen, 1985).

The first stage in the structure determination is relatively free from bias in that the results obtained do not depend on any assumptions one might make about the nature of protein structures (Richardson, 1981, Richardson and Richardson, 1985). The phases of the structure factors are estimated by methods such as the multiple isomorphous replacement technique (North and Phillips, 1969, Blundell and Johnson, 1976, McPherson, 1982, Watenpaugh, 1985); these enable the calculation of an electron density map. At this stage the secondary structural elements of the protein such as alpha helices, beta sheets, turns and loops can often be clearly discerned in the electron density map, and the architecture of the protein in terms of these elements can usually be unambiguously described (Richardson, 1981). However, the assumptions made in deriving the phases break down at high resolution and so one cannot determine very much more than the position and approximate conformation of each residue in the protein.

If high resolution data (3Å or better) are available, least-squares refinement of molecular models against the X-ray data significantly improves the accuracy of the model coordinates. For refinement against data to 2.0-1.5Å, the positions of individual (non-hydrogen) atoms in the protein can be determined with accuracies typically ranging from 0.1Å to 0.25Å for the better determined regions (Chambers and Stroud, 1979). Individual atomic temperature factors (B-factors) can also be refined and information about the dynamics of atoms in the protein can be obtained in this way (Willis and Pryor, 1975).

Refinement methods are essential for determining precise structures, but the results obtained are not free from bias because simplifying assumptions have to be made in the molecular model used in the refinement. The problem is that, unlike small molecule crystals, protein crystals do not diffract X-rays to very high resolution. Proteins are large flexible molecules which crystallize in unit cells with very high solvent content[1]. The consequent lack of rigidity in protein crystals leads to the disappearance of measurable diffraction intensity at resolutions higher than 2.0Å to 1.5Å, for most proteins. This severely lim-

[1] The composition of the solvent depends on the mother liquour used for crystallization; approximately half the solvent is disordered and cannot be located in electron density maps (Blake et al., 1983). Matthews (1968) surveyed a large number of protein crystals and reported that the solvent content varied from 25% to 60%, by volume. Recently Wilson et al. (1981) reported that hemagglutinin crystals contain 80% solvent, by weight.

its the complexity of models that can be used to describe proteins in refinement procedures.

It is the flexibility of protein molecules that reduces the number of measurable data and forces the use of simplified models in the refinement. However, the approximations which are introduced work best when the atoms being refined are not too mobile. The extent to which the results of protein refinement are dependent on the model is therefore of interest. This thesis addresses this problem by using the results of molecular dynamics simulations of myoglobin to evaluate the performance of the X-ray refinement method (Chapter 2). The rest of the thesis presents the results of X-ray diffraction studies on the structural effects of low temperature (Chapter 3) and ligand binding (Chapter 4) in myoglobin. The errors in the X-ray structure and B-factors are also discussed (Chapter 5).

The aim of this chapter is to serve as an introduction to the thesis and also to summarize the conclusions of the work reported here. In Section II a fairly detailed description of refinement methods and applications is given. Section III is a summary of the techniques which use empirical energy functions to simulate the structure and dynamics of proteins. This section is brief since several reviews of the field have appeared recently (see references in Section III), including the proceedings of a conference devoted to molecular dynamics and protein structure (Hermans, 1985). Section IV is an outline of the thesis and summarizes the four chapters that follow.

Section II: X-ray Diffraction Theory and Refinement Methods

The basic idea in refinement is to vary the parameters of the molecular model so as to minimize the deviations between the observed and calculated data. This requires a theory that relates the measured X-ray intensities to the parameters of the molecular model for the protein. This section describes the theory of X-ray scattering and then proceeds to discuss various refinement methods commonly used to improve protein models. Some remaining problems in the refinement of protein structures are also mentioned[2].

II (i) Bragg Scattering and Thermal Diffuse Scattering

There are two components to the scattering of X-rays from a crystal: Bragg scattering and diffuse scattering (Amoros and Amoros, 1968, Willis and Pryor, 1975, Stewart and Feil, 1980, Appendix to Chapter 2 of this thesis). The former is very sharply peaked around the reciprocal lattice points (defined by Bragg's law) and contains information about the average molecular structure and atomic distribution functions. These distribution functions are a sum of dynamic contributions (due to the motion of atoms in individual molecules) and static contributions (due to non-interconverting differences in the structure of different molecules). The diffuse scattering is so named because it is not restricted to the reciprocal lattice points; it too can be separated into a

[2] The most recent collection of papers on protein refinement are to be found in Volume 115 of "Methods in Enzymology", edited by Wyckoff, Hirs and Timasheff (1985). This volume is a survey of current refinement methods and it also offers a great deal of practical advice.

dynamic component and one that depends on static disorder (Amoros and Amoros, 1968). The dynamic component is called "Thermal Diffuse Scatter" (TDS) and arises because of correlations between the motions of atoms in different unit cells (Appendix to Chapter 2).

In theoretical treatments of X-ray scattering the effects of static disorder are usually ignored and the emphasis is on the calculation of the dynamic properties of the molecular or crystal system. These properties are then used to calculate the Bragg scattering or the TDS (Willis and Pryor, 1975).

The theory that is used to calculate the intensity of the Bragg scatter is well established (James, 1948, Woolfson, 1970, Willis and Pryor, 1975, Stewart and Feil, 1980). If the electron density in a unit cell (averaged over time and all the unit cells in the crystal) is known, the Bragg scatter can be calculated from the square of the Fourier transform of the average density (Stewart and Feil, 1980). The calculation of the TDS is much more difficult because it requires knowledge of the dynamics of the molecular lattice and not just that of individual molecules. In small molecules TDS has been shown to account for as much as 35% of the measured intensities (Stevenson and Harada, 1983) and attempts are made to correct for this by using harmonic lattice dynamics to estimate its magnitude (Stevenson and Harada, 1983, Gramaccioli and Filippini, 1983).

In protein crystallography, the effects of TDS have been noted (Wilson et al., 1983), but so far no corrections have been made for them. The analysis of TDS in proteins is a promising field of research that might yield information about correlated modes and intermolecular forces in protein crystals (Phillips et al., 1980), but it is not treated in the work described in this thesis (except for a brief discussion in the Appendix to Chapter 2).

II (ii) The Structure Factor

In relating the parameters of a molecular model to the intensities of X-rays scattered by a crystal, the function of interest is the structure factor of the molecule, which is defined as follows. Define a scattering vector, $\underset{\sim}{Q}$:

$$\underset{\sim}{Q} = 2\pi \frac{\underset{\sim}{e} - \underset{\sim}{e}_o}{\lambda} \tag{1}$$

where $\underset{\sim}{e}$ and $\underset{\sim}{e}_o$ are unit vectors along the wave vectors of the scattered and incident radiation, respectively, and λ is the wavelength of the radiation. The intensity of scattered radiation (Bragg Scatter), $I(\underset{\sim}{Q})$, is given by:

$$I(\underset{\sim}{Q}) = K \left|F(\underset{\sim}{Q})\right|^2 \tag{2}$$

where $F(\underset{\sim}{Q})$ is the structure factor and K is a constant. The structure factor is the Fourier transform of the average electron density, $\langle\rho(\underset{\sim}{r})\rangle$, in a unit cell of the crystal (Willis and Pryor, 1975):

$$F(\underset{\sim}{Q}) = \int d\underset{\sim}{r} \ \langle\rho(\underset{\sim}{r})\rangle \ e^{i\underset{\sim}{Q}\cdot\underset{\sim}{r}} \tag{3}$$

For the calculation of X-ray diffraction intensities, the average molecular electron density can, to a very good approximation, be represented as a superpositioning of average atomic electron densities (Ten Eyck, 1973, 1977). The average atomic density function for the i^{th} atom, $\langle\rho_i(\underset{\sim}{r})\rangle$, is the convolution of the electron density of the atom at rest, $\rho_{oi}(\underset{\sim}{r})$, and the distribution function for the atom, $P_i(\underset{\sim}{r})$ (Willis and Pryor, 1975):

$$\langle\rho_i(\underset{\sim}{r})\rangle = \rho_{oi}(\underset{\sim}{r}) * P_i(\underset{\sim}{r}) \tag{4}$$

$\rho_{oi}(\underset{\sim}{r})$ is assumed to be completely defined by the position of the atom (Ten Eyck, 1977, see also Chapter 2 of this thesis). The distribution function $P_i(\underset{\sim}{r})$ depends on the complexity of the atomic motion. For example, if the motion is harmonic $P_i(\underset{\sim}{r})$ is a Gaussian distribution.

The Fourier transform of a convolution of two functions is the product of the Fourier transforms of the individual functions. Hence, the structure factor is given by:

$$F(\underset{\sim}{Q}) = \sum_{i=1}^{N} FT\langle\rho_i(\underset{\sim}{r})\rangle = \sum_{i=1}^{N} FT[\rho_{io}(\underset{\sim}{r})] \, . \, FT[P_i(\underset{\sim}{r})] \tag{5}$$

where FT stands for the Fourier transform. Numerical values for the Fourier transform of the atomic electron density (at rest) are tabulated for a large number of atoms and ions in the International Tables for X-ray Crystallography (1974)[3].

For protein refinements, so far, the assumption has always been made that $P_i(\underset{\sim}{r})$ is a Gaussian distribution. The Fourier transform of this is also a Gaussian function, and is called the atomic Debye-Waller factor: $e^{W_i(\underset{\sim}{Q})}$. In the isotropic case $W_i(\underset{\sim}{Q})$ is given by

$$W_i(\underset{\sim}{Q}) = -\frac{1}{6}\langle\Delta\underset{\sim}{r}_i^2\rangle|\underset{\sim}{Q}|^2 = -\frac{8}{3}\pi^2\langle\Delta\underset{\sim}{r}_i^2\rangle s^2 \tag{6}$$

where $\langle\Delta\underset{\sim}{r}_i^2\rangle$ is the mean-square fluctuation of the i-th atom and $s = \frac{|\underset{\sim}{Q}|}{4\pi}$.

[3] Eqn. 5 is commonly expressed as a sum over atomic "scattering factors", "temperature factors" and "phase factors". See Chapter 2 for a detailed discussion of this and the calculation of structure factors.

The term $\frac{8}{3}\pi^2\langle\Delta r_j^2\rangle$ is referred to as B_i, the atomic B-factor or temperature factor (Willis and Pryor, 1975).

Eqns. 5 and 6 allow the calculation of structure factors from a molecular model which has four variable parameters per atom (three coordinates and a temperature factor). The progress of a refinement is usually monitored by calculating the R-factor, R:

$$R = \frac{\sum_{h} \left| \, |F_o| - |F_c| \, \right|}{\sum_{h} |F_o|} \qquad (7)$$

where the sum runs over all the $\underset{\sim}{h} = (h,k,l)$ indices of the measured reflections. F_o and F_c are the observed and the calculated structure factors, respectively.

The R-factor for a non-centrosymmetric structure with a completely random distribution of atoms is 59% (Wilson, 1949). Protein models obtained by manual fitting of experimentally phased electron density maps have R-factors around 35% or higher (Jensen, 1985). Refinement usually reduces R to around 20%. In cases where good data are available, and a great deal of effort is put into correctly modelling the protein and the solvent, R can be lowered to around 12% (for example, see Watenpaugh et al., 1980, James and Sielecki, 1983). This is in contrast to the the R-factors of 5% or less which are commonly obtained for small molecules (Dunitz, 1979).

In most cases, the resolution of X-ray data available for proteins is not enough to allow the use of more complicated models for atomic motion (such as anisotropic Gaussian distributions or anharmonic corrections to the distribution). However, a few proteins such as crambin

(Teeter and Hendrickson, 1979), bovine pancreatic trypsin inhibitor (Wlodawer et al., 1984) and avian pancreatic polypeptide (Glover et al., 1983), diffract to better than 1.0Å resolution and anisotropic temperature factors have been refined. The only treatment for anharmonicity that has been used so far has been to allow bi-modal distributions (i.e. discrete disorder) for sidechains (for example: Honzatko et al., 1985, Haneef et al., 1985a, and Chapters 2 and 4 or this thesis). Anharmonic corrections involving the third and higher moments of single peaked distributions are used in the refinement of small molecules (Zucker and Schulz, 1982), but simulation results indicate that in proteins disorder (i.e. multiple peaks in the distribution) is the dominant cause of anharmonicity (Ichiye and Karplus, 1986).

(iii) Real-Space and Reciprocal-Space Refinement

In "real-space" refinement the residual to be minimized is the difference between the "observed" and the calculated electron density (see Chapter 2 for a detailed discussion of the calculation of electron density from a model). For example, in the method of Diamond (1971, 1985), the residual, Δ, is of the form:

$$\Delta = \int_V |\rho_o - \rho_c|^2 dv \qquad (8)$$

where ρ_c is the electron density calculated from the model and ρ_o is the "observed" electron density calculated by combining the experimental structure factor magnitudes with experimental or calculated phases. A problem with this method is that ρ_o is not a true observable (if it were, protein crystallography would proceed at a much faster pace). Experimentally estimated phases (e.g. from multiple isomorphous replacement) can be used to calculate the electron density, but these are often subject to large errors (Watenpaugh, 1985).

In "reciprocal-space" refinement (often referred to in the literature simply as least-squares refinement), the residual that is minimized is usually of the form:

$$\Delta = \sum_{\underset{\sim}{h}} w(\underset{\sim}{h}) \quad [|F_o(\underset{\sim}{h})| - |F_c(\underset{\sim}{h})|]^2 \tag{8}$$

where the sum runs over all the $\underset{\sim}{h} = (h,k,l)$ indices of the observed reflections; F_o and F_c are the observed and calculated structure factors and $w(\underset{\sim}{h})$ is the weight assigned to each measurement (Rollett, 1970, 1982). Here the quantity being minimized is the sum of squared deviations from the actual measurements and is free from bias toward the initial phases. The drawback is that for most proteins the number of observed data are not enough to over-determine the problem significantly and so unconstrained minimizations are not possible. Also, the large number of atoms in proteins makes a full matrix treatment (see below) of the minimization computationally intractable at present.

The first attempts, in the 1960s, to refine a protein structure (that of myoglobin) were not very successful because the refinement methods and the computers then available proved to be inadequate (Branden et al., 1963, Watson et al., 1963). Attempts to improve the quality of the hand-built protein structures included difference-Fourier (also called ΔF) techniques where atoms are moved along the gradient in difference electron density maps (Watenpaugh et al., 1973, Freer, 1985). These methods are essentially peak-search algorithms and have a smaller radius of convergence than real-space refinements in which Eqn. 7 is minimized (Diamond, 1985).

ΔF methods do improve the agreement between calculated and observed structure factors, but they suffer from the disadvantage that the refinement is done without reference to the stereochemistry of the residues. The low resolution and large errors in the density maps lead to the refined structures often having unrealistic stereochemistry, which then has to be improved by regularization programs (Hermans, 1985).

A natural development therefore was the introduction of constrained and restrained refinements[4]. A constrained refinement is one in which certain parameters (such as the bond lengths) are kept fixed and not allowed to vary. In a restrained refinement all parameters are allowed to vary, but they are "restrained" to be near specified reference values. The use of constraints or restraints are essential for refinement of almost all proteins because the least-squares problem is not over-determined in the absence of very high resolution data (see Chapter 2 for a discussion of this point). It is interesting, however, that the first reciprocal-space refinement of a protein (rubredoxin) was done by a completely unrestrained least-squares method at 1.5Å resolution (Watenpaugh et al., 1973). This will be discussed in more detail below.

R. Diamond developed a constrained real-space refinement method (Diamond, 1971), which was successfully used to refine the structures of many proteins, including that of myoglobin (Diamond, 1971, Takano,

[4] Webster's Unabridged Dictionary (1961) considers "restraint" and "constraint" to be synonyms. Crystallographers, however, have specific and different meanings for these terms. See, for example, Sussman (1985).

1977). Diamond's method had several advantages over the unconstrained reciprocal-space methods that were available then. The protein is treated as a "flexible chain" in that bond lengths and most angles are kept fixed and flexibility is usually only allowed around certain torsional angles. Because real-space refinement is local in nature, the number of terms to be calculated in each cycle increases only linearly with the number of atoms, unlike the n^2 dependence of the matrix elements in full-matrix reciprocal lattice methods.

Diamond's method allows for the refinement of small parts of the protein independently, which was an essential feature at a time when computer memories and discs were not very large. In this method the parameters for only ten residues ("the molten zone") were refined at a time, allowing very large molecules to be treated piece by piece (Diamond, 1971). The disadvantages of the method are the ones common to all real-space refinement; the refinement is not done against the true observables, and there is no way to weight the different observations relative to one another (Diamond, 1971). The use of constraints introduces two further problems. First, this method is not suitable for high resolution refinement since the constraints limit the structural information that can be obtained. Second, since flexibility is only allowed at a few "joints", errors in the structure tend to accumulate at these points.

Deisenhofer et al. (1985) have compared Diamond's real-space method with reciprocal-space programs. They conclude that real-space refinement is useful only in the initial stages of the refinement. A list of proteins refined by this method is given.

In the last ten years high speed computers with large memories have become available. Along with this, reciprocal-space refinement programs have been developed which overcome the under-determinancy problem by including stereochemical or energy restraints (PROLSQ: Konnert, 1976, Konnert and Hendrickson, 1980, Hendrickson, 1985; CORELS: Sussman et al., 1977, Sussman, 1985; Agarwal: Agarwal, 1978, modified to include restraints by Dodson (1980); Jack-Levitt: Jack and Levitt, 1978; RESTRAIN: Moss and Morffew, 1982, Haneef et al., 1985a). The n^2 dependence of the number of operations on the number of atoms has been reduced by using diagonal or block-diagonal approximations in the normal matrix (Dodson, 1980). In many cases, fast Fourier transform (FFT) algorithms are used to speed up the calculation of structure factors and their derivatives (Agarwal, 1978, Jack and Levitt, 1978, Dodson, 1980). Due to advances such as these, reciprocal-space methods have largely superceded real-space refinement.

Real-space refinement is, however, becoming increasingly important as a tool for building initial models for very large molecules. At present this is usually done manually, by using a graphics software package such as BUILDER (Diamond, 1966), GRINCH (Brooks and Pique, 1985) or FRODO (Jones, 1982, 1985) to fit the molecular model to the density. For the large protein structures being solved today, with many hundreds of residues, this can be an extremely time consuming and error prone task. Real-space refinement can be used to fit residues into local regions of density (interactively, on a graphics system) and the completed model can then be improved by the more powerful method of reciprocal-space refinement. Such a system, for example, has been developed by A. Jones and coworkers (Jones and Liljas, 1984, Jones, 1985). An automated procedure to build atoms into electron density maps has been described by Greer (1985).

Finally, before proceeding to describe the details of least-squares refinement, it is instructive to consider how far crystallographic refinement has come in the last few decades. The first structure to be refined by least-squares, to my knowledge, is that of melamine (Hughes, 1941). The 18X18 normal matrix took two days to set up using an I.B.M. "tabulator" and punched cards. The normal equations were solved in four hours. Today, the largest asymmetric unit being refined by least-squares is that of the influenza virus coat protein, hemagglutinin, at 2.9Å resolution. The refinement is being done by D.C. Wiley and co-workers, using the method of Konnert and Hendrickson as modified by Lewis and Rees (Knossow et al., 1986, Konnert and Hendrickson, 1980). 49,525 independent variables are being refined for 12,381 atoms and one cycle takes less than an hour on a CRAY-1S supercomputer (Knossow et al., 1986), even though fast Fourier transform algorithms are not used. FFT algorithms optimized for use on parallel processor machines have been described by Raftery et al. (1985).

II (iv) Least-Squares Refinement: The Normal Equations

Descriptions of the principles of least-squares refinement have been given by Rollett (1970, 1982) and Sparks (1985). The details given here follow the recent article of Rollett (1982). The problem is to minimize a function, M:

$$M = \sum_{h} w(\underset{\sim}{h})[|F_o(\underset{\sim}{h})| - |F_c(\underset{\sim}{h})|]^2 \qquad (10)$$

with respect to s parameters $(p_1, p_2, \ldots, p_s)$. For a model with isotropic Gaussian temperature factors for each atoms, $s = 4N$, where N is the number of atoms. $|F_o(\underset{\sim}{h})|$ is the modulus of the observed structure factor at a reflection with indices $\underset{\sim}{h}=(h,k,l)$, and $|F_c(\underset{\sim}{h})|$ is that of the calculated structure factor (the $\underset{\sim}{h}$ dependence will not be explicitly given in the equations which follow).

We wish to find parameters $\underset{\sim}{p}$ such that:

$$\frac{\partial M}{\partial p_i} = 0 \tag{11}$$

Expanding M in a Taylor series around $\underset{\sim}{p}$ we get:

$$M(\underset{\sim}{p}+\Delta\underset{\sim}{p}) = M(\underset{\sim}{p}) + \sum_{i=1}^{s} \frac{\partial M(\underset{\sim}{p})}{\partial p_i}\Delta p_i + \ldots \tag{12}$$

$$\Rightarrow \quad \frac{\partial M(\underset{\sim}{p}+\Delta\underset{\sim}{p})}{\partial p_j} = \frac{\partial M(\underset{\sim}{p})}{\partial p_j} + \sum_{i=1}^{s} \frac{\partial^2 M(\underset{\sim}{p})}{\partial p_i \partial p_j}\Delta p_i + \ldots \tag{13}$$

$\Delta\underset{\sim}{p}$ is the shift vector to be added to the parameters to reach the minimum. In the "Newton approximation" (Rollett, 1982), terms higher than the second derivative are neglected in Eqn. 13. Setting the derivative on the left hand side of Eqn. 13 to be zero for a minimum, we get:

$$\sum_{i=1}^{s} \frac{\partial^2 M(\underset{\sim}{p})}{\partial p_i \partial p_j}\Delta p_i = -\frac{\partial M}{\partial p_j} \tag{14}$$

Let $\Delta\underset{\sim}{h} = |F_o|-|F_c|$, so that:

$$M = \sum_{h} w(\underset{\sim}{h})(\Delta\underset{\sim}{h})^2 \tag{15}$$

Then,

$$\frac{\partial M}{\partial p_j} = -2\sum_{h} w(\underset{\sim}{h})\ (\Delta\underset{\sim}{h})\frac{\partial|F_c|}{\partial p_j} \tag{16}$$

and,

$$\frac{\partial^2 M}{\partial p_i \partial p_j} = -2 \sum_h w(\underset{\sim}{h})(\Delta \underset{\sim}{h}) \frac{\partial^2 |F_c|}{\partial p_i \partial p_j} +2 \sum_h w(\underset{\sim}{h}) \frac{\partial |F_c|}{\partial p_j} \frac{\partial |F_c|}{\partial p_i} \qquad (17)$$

In the "Gauss-Newton" approximation, the second derivatives of $|F_c|$ are neglected, yielding the "normal equations" (Rollett,1982):

$$\sum_{i=1}^{s} \sum_h w(\underset{\sim}{h}) \frac{\partial |F_c|}{\partial p_i} \frac{\partial |F_c|}{\partial p_j} \Delta p_i = - \sum_h w(\underset{\sim}{h})(|F_o|-|F_c|)\frac{\partial |F_c|}{\partial p_j} \qquad (18)$$

$$j = 1,2,\ldots,s$$

The normal equations are readily put into matrix form. Let $\underset{\sim}{A}$ be a s X s square matrix with elements:

$$A_{ij} = \sum_h w(\underset{\sim}{h}) \frac{\partial |F_c|}{\partial p_i} \frac{\partial |F_c|}{\partial p_j} \qquad (19)$$

Let $\underset{\sim}{P}$ and $\underset{\sim}{B}$ be column vectors of length s with elements:

$$P_i = \Delta p_i \qquad (20)$$

$$B_i = - \sum_h w(\underset{\sim}{h})(|F_o|-|F_c|) \frac{\partial |F_c|}{\partial p_j} \qquad (21)$$

Thus, Eqn. 15 can be written as:

$$\underset{\sim}{A}\,\underset{\sim}{P} = \underset{\sim}{B} \qquad (22)$$

The problem of minimizing the crystallographic residual has been reduced to solving the linear matrix equation (Eqn. 22). This method involves only first derivatives and is the form commonly used for structure refinement. The convergence could be improved, in principle, by including second derivatives (as in Newton-Raphson minimization; Sparks, 1985), but this is too expensive for proteins.

The matrix $\underset{\sim}{A}$ is known as the normal matrix; it contains s^2 elements where s is the number of parameters to be refined. Each element of the normal matrix involves a sum of N_{ref} terms, where N_{ref} is the number of reflections measured. Typically, for the proteins being refined today, s is on the order of (10^3) and N_{ref} is on the order of (10^4). Computing the entire normal matrix therefore involves on the order of (10^{10}) operations, each involving the addition and multiplication of derivatives of F_c. This is prohibitively expensive, even on supercomputers, so a common approximation is to neglect all the off-diagonal terms except, perhaps, those that connect parameters of the same atom, or atoms connected by restraint terms (Watenpaugh et al., 1973, Konnert, 1976, Agarwal, 1978, Jack and Levitt, 1978, Dodson, 1980, Sparks, 1985).

Protein refinement programs do not usually solve Eqn. 22 by inverting the normal matrix. Instead, methods such as conjugate gradients (Konnert, 1976) or the Gauss-Seidel algorithm (Haneef et al., 1985a) are used. The rate limiting step in a refinement cycle is the calculation of the elements of the normal matrix; even with the diagonal approximation this takes up 80% of the computer time (Haneef et al., 1985a). Consequently, the exact method used to solve the normal equations is not very critical (Haneef et al.,1985a). The shifts obtained by solution of Eqn. 19 are inaccurate because of the linear approximation made in deriving the equation. Refinement must therefore be continued iteratively until convergence is obtained; usually this involves at least 10 to 15 least-squares cycles. Also, the shifts obtained from Eqn. 22 are usually multiplied by a damping factor to obtain the maximum decrease in the R-factor (Agarwal, 1978, Hendrickson and Konnert, 1980, Haneef et al., 1985a).

It is common to talk about the "radius of convergence" of a refinement; this is the maximum positional error that can be corrected by the least-squares minimization. The radius of convergence varies in different parts of the structure but it is approximately 0.5Å to 1.5Å for least-squares refinement at 1.5Å resolution (the radius of convergence increases with decreasing resolution). The positions of atoms which are in error by more than this amount in the initial model are unlikely to be improved by least-squares refinement at high resolution.

Difference Fourier maps have to be periodically examined during the refinement to correct such errors manually. This process also leads to the identification and placement of new solvent molecules and, sometimes, to the discovery of alternate conformations for sidechains.[5] The model is then modified to include these corrections and new features and least-squares refinement is continued. Hence, apart from the iteration of a number of least-squares cycles, protein refinement also involves alternating stages of least-squares refinement and examination of difference electron density maps (see, for example, Chapter 4 of this thesis and Honzatko et al., 1985).

A brief discussion of the results of the first least-squares refinement of a protein model is now given. This is followed by a description of the salient features of several commonly used least-

[5] A few years ago difference maps were examined by plotting them out section by section and looking for peaks. Today the molecular model and the electron density are superimposed, in three dimensions, on a powerful interactive graphics device such as the Evans and Sutherland Picture System 300. Because of the complexity of protein structures and the relatively high noise levels in difference maps, the use of graphics systems makes a a very significant difference in the extent to which a model can be improved (Jones, 1985).

squares refinement programs.

II (v) Refinement of Rubredoxin

In 1973 Watenpaugh, Sieker, Herriott and Jensen reported the first refinement of a protein model by conventional least-squares (Watenpaugh et al., 1973). An important feature of this work is that the authors point out, with great clarity, several features of protein refinement which are now known to be quite general.

Rubredoxin is a small bacterial iron-sulphur protein with 54 amino acids. The initial refinement (Watenpaugh et al., 1973) was against X-ray data to 1.54Å resolution, collected on a diffractometer. The Kendrew skeletal model built from the initial electron density map was improved by shifting coordinates along the gradient in a ΔF synthesis (an electron density map calculated using coefficients $(F_o-F_c)e^{i\alpha_c}$ where α_c is the calculated phase). After four cycles of ΔF refinement, three positional coordinates and one isotropic temperature factor were refined for each atom by unrestrained least-squares using a block-diagonal normal matrix. The normal matrix was inverted to solve the normal equations and the standard deviations of the refined parameters were estimated in this way. In the refinement programs commonly used today this is not possible because the normal matrix is not inverted; also, the use of restraints reduces the number of free parameters in an undetermined way (Hendrickson, 1980).

The following points are mentioned by the authors (Watenpaugh et al., 1973) as being clarified by their work. They have established that protein structures can be improved by conventional least-squares techniques, despite the limited resolution of the data available. They show that refinement leads to the emergence of structural features not apparent in the original electron density map and that the neglect of the disordered solvent in the refinement affects only a relatively small number of reflections, with d<10Å. Hydrogen atoms were shown to have a detectable effect on the R-factor and it is pointed out that neglecting them would lead to some error in the positions of the non-hydrogen atoms. The modelling of surface sidechains was found to be difficult because these are often extremely flexible or disordered.

Watenpaugh et al. (1973) also note that some atoms (usually those with high temperature factors) refined to give unreasonable bond lengths and angles. They suggest that this might be due to high thermal motion or disorder and suggest that the temperature factors and standard deviations for these atoms are not realistic. This anticipates the results of the molecular dynamics "experiment" described in Chapter 2 of this thesis.

Because of the lack of restraints, bond lengths vary considerably throughout the structure (Watenpaugh et al., 1973). The standard deviation of C_α-C_β bond lengths is 0.20Å, in agreement with the value derived from the estimated standard deviations in the coordinates. In restrained refinements (see below) the standard deviations of bond lengths are usually between 0.02Å and 0.03Å (Hendrickson, 1985). The standard deviations of bond lengths in the unrestrained refinement could be larger not only due to errors in the structure but also due to systematic deviations from ideal values due to motion. The errors in the atomic position seem to be the dominant factor, at least at this stage of the refinement.

The resolution of the X-ray data for rubredoxin was later extended to 1.2Å, which is beyond that measurable for most proteins. A new data set was collected from ∞ to 1.2Å; refinement against these data improved the precision of the model (standard deviations of C_α-C_β bonds decreased to 0.1Å). Detailed reports of the refinement and the modelling of the water structure and the use of anisotropic temperature factors have been published (Watenpaugh et al., 1978, 1980).

II (vi) Various Least-Squares Refinement Programs

In this section various commonly used refinement programs are described. These programs differ in the way they handle restraints and whether FFT algorithms are used to calculate structure factors and their derivatives. An early paper on restraints was that by Waser (1963), who suggested that including stereochemical information as additional observations was a more flexible method than using Lagrange multipliers to simultaneously improve stereochemistry and decrease the crystallographic residual. Another method is to include an explicit energy function in the minimization. As mentioned earlier, Volume 115 of Methods in Enzymology (Wyckoff, Hirs and Timasheff, eds., 1985) contains several useful papers on refinement.

II (vi) A: Agarwal's Fast Fourier Transform Method

The use of FFT techniques (Ten Eyck, 1973,1977,1985) to calculate structure factors and their derivatives speeds up the calculation by a factor of 10 or more, but involves a considerable programming investment. R.C. Agarwal introduced the first least-squares refinement in which FFT techniques were used at all possible stages. No restraint terms were included in the normal matrix (Agarwal, 1978, Isaacs and Agarwal, 1978); the stereochemistry of the model was regularized using a separate program (Dodson et al., 1976). The first application of the program was to the refinement of insulin at a resolution of 1.5Å (Isaacs and Agarwal, 1978). Later, Dodson introduced restraint terms into the program (Dodson, 1980) and, for example, the structure of actinidin (a plant protease) was refined in this way (Baker, 1980).

II (vi) B: Jack-Levitt

The requirement that the structure being refined should have good stereochemistry and non-bonded contacts while satisfying the X-ray data is best met by simultaneously minimizing the internal energy of the molecule and the crystallographic residual. A. Jack and M. Levitt (1978) combined Agarwal's (1978) method for the fast calculation of the crystallographic residual and its derivatives with Levitt's (1974) energy minimization program. The function minimized is E + kΔ where E is the internal energy of the protein (Levitt, 1974) and $\Delta = \sum_{h} w(\underset{\sim}{h})(|F_o|-|F_c|)^2$. k is a scale factor which controls the relative weighting of the energy terms and the crystallographic residual.

The Jack-Levitt procedure has been used, for example, to refine the structure of two forms of citrate synthase at resolutions of 1.7Å and 2.7Å respectively (Remington et al., 1982). The authors point out that refinement with an incomplete sequence dramatically improved the Fourier map to the point that major revisions of chain connectivity were possible. Phillips (1980) has used the Jack-Levitt method to refine the structure of oxy-myoglobin at 1.8Å resolution to a final R-factor of 15.9%. He has shown that, even at this resolution, the neglect of hydrogens in the atomic model leads to a noticeable increase in the apparent bond-lengths and the R-factor (this was initially noted by Watenpaugh et al., 1973). The empirical energy function used in the refinement was also used to rationalize the observed ligand geometry and the existence of discrete disorder in several sidechains. Phillips showed that the use of empirical models for the disordered solvent regions leads to significant reductions in the R-factor (Phillips, 1980).

II (vi) C: PROLSQ, Konnert and Hendrickson

The most widely used refinement program for proteins today seems to be PROLSQ (PROtein Least-SQuares), developed by J. H. Konnert and W. A. Hendrickson (Konnert, 1976, Konnert and Hendrickson, 1980, Hendrickson and Konnert, 1980, 1981, Hendrickson, 1980, 1985). Konnert (1976) incorporated the ideas of Waser (1963) regarding the use of subsidiary conditions (such as stereochemical restraints) in refinement, into a least-squares program for large molecules. Hendrickson and Konnert (1980) extended the program to include the refinement of isotropic temperature factors with restraints (Yu et al., 1985, see also Chapter 2 of this thesis) and anisotropic temperature factors for which the orientation of the thermal ellipsoid is determined by the local bonding (simulation studies, however, indicate that this assumption is unlikely to be correct for proteins (Yu et al., 1985)).

The function minimized in PROLSQ is similar in principle to that used in the Jack-Levitt program. Instead of an explicit energy function, there are terms involving deviations from ideal stereochemistry for bonds, bond angles, torsions and planar groups. Atoms are prevented from approaching each other too closely by the inclusion of repulsion terms at short interatomic distances. The dictionary of ideal values for the stereochemical parameters is derived from crystal structures of small molecules (Sielecki et al., 1979).

PROLSQ is the refinement program used in all the work reported in this thesis; there are discussions of its use in all the chapters that follow. A large number of structures have been refined using PROLSQ, and the papers describing the refinements are a useful source of information on strategies (for example, myoglobin (Frauenfelder et al., 1979), lysozyme (Artymiuk et al., 1979, Artymiuk and Blake, 1981), arabinose binding protein (Quicho and Vyas, 1984),chymotrypsin (Tsukada and Blow, 1985), hemerythrins (Sheriff et al., 1985), lamprey hemoglobin (Honzatko et al., 1985), α-lytic protease (Fujinaga et al., 1985), pepsinogen (James and Sielecki, 1985) and hemagglutinin (Knossow et al., 1986)). PROLSQ has been modified for use with nucleic acids by G. Quigley at M.I.T. (see, for example, Westhof et al., 1985).

Two excellent descriptions of refinements using PROLSQ are included in the reports on penicillopepsin (James and Sielecki, 1983) and lamprey hemoglobin (Honzatko et al., 1985). In particular, the treatment of discrete disorder of sidechains and the weighting of restraint terms relative to the X-ray data are discussed by Honzatko et al. (1985). James and Sielecki (1983) present a number of criteria for judging the convergence of a refinement and the quality of the refined structure.

These include searching the structure for unreasonable stereochemical parameters and examination of difference Fourier maps for unexplained peaks.

II (vi) D: RESTRAIN, Moss and coworkers.

This program, developed by D. Moss and his colleagues (Moss and Morffew, 1982, Haneef et al., 1985a), is very similar to PROLSQ in the way restraints are incorporated. Some differences and improvements are as follows. RESTRAIN allows the use of experimentally determined phases as additional observables, which is a useful feature for refinement at very low resolution (<3Å). The treatment of planar groups, such as phenyl or histidyl rings, has been improved. Haneef et al. (1985a) claim that the planar restraints used in PROLSQ (minimization of deviations from the current least-squares plane) tend to damp the rotation of planes, as a whole, toward new positions. They introduce a product-moment method which avoids this problem (Haneef et al., 1985a).

RESTRAIN allows for the refinement of anisotropic temperature factors for groups which can be approximated as rigid bodies. Rigid-body displacements are introduced, allowing the refinement of parameters which describe overall translational and librational motion (Haneef et al., 1985a). Such rigid body refinements, for groups of four atoms or more, require fewer parameters than refinements in which the motion of each atom is described by a six-parameter temperature factor tensor (Schomaker and Trueblood, 1968, Haneef et al., 1985a, Holbrook and Kim, 1984). Rigid body refinements have proved very useful in nucleic-acid studies where they have been shown to yield meaningful parameters and reduce the noise-level of difference Fourier maps by 20% (Holbrook and Kim, 1984,1985).

If sufficient data are available, RESTRAIN allows the refinement of unrestrained anisotropic temperature factors for each atom. Such a refinement has been reported for the small hormone, avian pancreatic polypeptide hormone (36 residues) for which data to 0.98Å are available (Glover et al., 1983, Haneef et al., 1985b).

II (vi) E: CORELS, Sussman and Coworkers

CORELS (COnstrained REstrained Least-Squares) was developed to combine the best features of constrained and restrained refinements into one reciprocal-space program (Sussman et al., 1977, Sussman, 1985). At very low resolution it is preferable to treat large parts of the structure, such as entire domains or alpha helices, as rigid bodies. As the resolution of the data increases, smaller parts of the structure can be constrained. CORELS allows the separation of the refinement model into rigidly constrained parts that are joined to each other by flexible (but restrained) parameters. The refinement is done by a least-squares sparse matrix method, and has been applied to molecules such as t-RNA and concanavalin-A (Sussman, 1985). The ability to refine anisotropic temperature factors for rigid-body motion has been incorporated recently (Holbrook and Kim, 1984,1985).

II (vii) Remaining Problems in X-ray Refinements of Proteins

To complete this section I shall briefly list some areas in which I believe more work is required.

(i) Treatment of atomic mobility. Two improvements over the isotropic model are to include anharmonicity, in the form of multiply-peaked Gaussians, and anisotropy (Ichiye and Karplus, 1986). Since most proteins do not diffract beyond 1.5Å, completely unrestrained refinements of parameters for these models may not be possible. The most appropriate approximations to introduce should be identified; simulation methods might aid in this process (Yu et al., 1985, Haneef et al., 1985b, Kuriyan, Chapter 2 of this thesis).

(ii) Estimation of errors in the structure and temperature factors. It is very difficult to relate the (known) errors in the structure factors to real-space parameters (Luzzati and Taupin, 1984). As mentioned earlier, the inverse of the least-squares normal matrix cannot be used in most cases. The "perturbation/refinement" approach suggested in Chapter 5 might be a useful method, especially since the availability of increased computing power will allow a large number of re-refinements to be done from perturbed structures.

(iii) Appropriate values for stereochemical restraints. Motional averaging can cause the time-averaged structure to exhibit apparently deviant stereochemistry. If tight restraints are applied, then the true average structure will not be obtained. It would be interesting to refine, without restraints, models for proteins for which diffraction data to better than 1.0Å resolution are available. The results should then be compared with those obtained from restrained refinements and from simulations to see if any systematic trends are observable.

(iv) The modelling of the solvent. Some water and solvent molecules are very tightly bound to the protein and are easily modelled (Watenpaugh et al., 1973, Blake et al., 1983, North and Smith, 1985). At the other extreme, the disordered solvent continuum can be modelled by a constant electron density or some other empirical function (Phillips, 1980, Blake et al., 1983). The problem lies in treating the large number of water molecules which are partially disordered.

The conformations of surface sidechains need to be checked to see if the solvent structure has been misinterpreted and partial occupancies might need to be refined for the waters. In myoglobin, for example, some of the surface histidines have two water molecules within H-bonding distance on opposite sides of the ring. The imidazole ring apparently flips between states in which the $N_{\delta 1}$ atom is H-bonded to one or the other water molecule (Kuriyan, unpublished). Simulation methods might prove useful in deciding what models to use for the loosely bound waters (Hermans et al., 1984, Moult and James, 1985).

(v) Simulation methods. Empirical energy functions have been used successfully to improve structures, from the early days of protein refinement (Levitt, 1974). However, attempts to directly use the results of molecular dynamics or Monte Carlo simulations in protein refinement have not been very successful (van Gunsteren et al., 1983). What is needed is a synthesis of the two methods where the simulations are used to provide limiting or approximate forms for the atomic distribution functions, which can then be refined against the X-ray data.

Section III: Simulation of Protein Structure and Dynamics

Empirical energy functions have proved to be very useful in the analysis of protein structure as well as in the calculation of the statistical mechanical properties of these molecules. Only a very brief introduction to this field will be given here since many comprehensive review articles have recently been published. For a recent survey of the literature in this area see Barlow et al. (1985).

The energy of the protein molecule, in isolation or surrounded by solvent or a crystal environment, is assumed to be given by an empirical energy function of the form:

$$
\begin{aligned}
E = & \sum_{\text{bonds}} \frac{1}{2} K_b \, (b-b_o)^2 + \sum_{\text{angles}} \frac{1}{2} K_\theta (\theta-\theta_o)^2 \\
& + \sum_{\text{impropers}} \frac{1}{2} K_\omega (\omega-\omega_o)^2 + \sum_{\text{dihedrals}} K_\phi \, (1 + \cos(n\phi+\delta)) \\
& + \sum_{\text{pairs}} \left[\frac{C_{12}}{r_{ij}^{12}} - \frac{C_6}{r_{ij}^{6}} + \frac{q_i \, q_j}{4\pi \varepsilon_0 r_{ij}}\right]
\end{aligned}
\tag{23}
$$

(Burkert and Allinger, 1982, Brooks et al., 1983, Weiner et al., 1984, van Gunsteren and Berendsen, 1985). The first two terms are harmonic bond and angle stretching terms. The third term ("improper torsions") is also harmonic and is used for chiral centers when one of the atoms of the center is not explicitly treated, and planar groups (Brooks et al.,1983). The fourth term is a torsional potential with multiple minima. Atomic interaction are assumed to be pairwise additive. The last term is a sum of van der Waals and electrostatic interactions between atoms, and includes the effects of hydrogen bonding (Reiher, 1985).

The program CHARMM (CHemistry at HARvard Macromolecular Mechanics, Brooks et al., 1983) was used for all the simulations described in this thesis. Eqn. 23 is the current form of the energy function used in CHARMM; the exact form varies from program to program. The parameters depend on the form that is used; for example, partial atomic charges will be different in models that have explicit H-bonding terms. The use of empirical energy functions to study the properties of molecules is justified and explained by Burkert and Allinger (1982). The development of functions to study biomolecules has recently been reviewed by van Gunsteren and Berendsen (1985) and by Pettitt and Karplus (1986). Detailed accounts of the derivation of commonly used biomolecular force-fields are given by Lifson et al. (1979a,b) and Weiner et al. (1984).

Given a force field, such as in Eqn. 23, and an X-ray structure, the structure and dynamics of the protein can be studied in various ways. Low energy structures can be obtained, and the results of perturbations examined, by energy minimization (Gelin et al., 1983, Novotny et al., 1984). Monte Carlo simulations generate ensembles of structures at a particular temperature; these can be used to obtain average values of molecular properties of interest (Northrup and McCammon, 1980). Molecular Dynamics (MD) simulations also generate ensembles of structures, but these are connected in time since they are points along a phase space trajectory obtained by solving classical equations of motion (McCammon et al., 1977).

MD simulations of proteins are capable of giving information about the system in ultimate detail, but they are currently restricted to time-scales of 100-1000 picoseconds (ps). Information about slower processes can sometimes be obtained by other methods such as stochastic dynamics (van Gunsteren et al., 1981) and activated dynamics (Northrup et al., 1982). A complete description of the dynamics in the harmonic approximation can be obtained by calculating all the normal modes of the protein (Go et al., 1983, Brooks and Karplus, 1983, 1985, Levitt et al., 1985). Though more approximate than molecular dynamics or Monte Carlo simulations, normal modes allow the estimation of changes in entropy in a relatively simple manner (Karplus and Kushick, 1981).

The use of simulation methods to study proteins has been reviewed by Karplus, (1981,1984,1985), Karplus and McCammon (1981,1983), Levitt (1982), van Gunsteren and Berendsen (1982, 1985), McCammon and Karplus (1983), McCammon (1984) and Kollman (1985). Though these simulations give a very detailed picture of protein dynamics, they are approximate. Energy minimization results in shifts of 0.25Å to 1.0Å from the X-ray crystal structure, depending on the minimization algorithm used. The shifts are smallest for the backbone atoms. Average structures from molecular dynamics simulations deviate from the X-ray crystal structures by 1.0Å to 3.0Å. A large part of the shift occurs in the first few picoseconds of the simulations, while the structure is being equilibrated; thus, the structure moves away from the X-ray structure to a new, stable, "vacuum" or solution structure. The backbone shifts are again smaller than those for the sidechain, usually by a factor of two. The average mean-square fluctuations calculated from the simulations are sometimes smaller and sometimes larger than the average X-ray values. The latter contain contributions due to static disorder (see Chapter 5), making comparison of the absolute magnitudes a little difficult. On setting the average values to be equal, the simulation results show much wider variation from residue to residue than the X-ray results, especially for the sidechains, though the overall patterns are often well reproduced (see references above).

Comparison of simulation results with X-ray structures and temperature factors are complicated by the neglect of the crystal environment in the simulation and the bias introduced by refinement (Chapter 2 of this thesis describes an attempt to get around this). The most serious shortcomings in the simulations, at present, include errors in the force-field, the neglect of solvent in many simulations and the short time scale explored by trajectory calculations. Supercomputers are making longer simulations possible, with explicit treatment of waters (van Gunsteren, 1985). The force-fields are gradually being improved by comparison with experiment and accurate quantum-mechanical calculations (Weiner et al., 1984, Reiher 1985).

The comparison of the results of simulations with experimentally derived parameters such as X-ray temperature factors, NMR order parameters and fluorescence depolarization rates is important for validating the simulations. On the other hand, the exteremely detailed information available from the simulation can lead to a better understanding of the experimental techniques (for example, see Chapter 2 of this thesis). Reviews of experimental studies of protein flexibility and dynamics are given by Huber (1979), Gurd and Rothgeb (1979), Debrunner and Frauenfelder (1982), Petsko and Ringe (1984) and Stuart and Phillips (1985). The Proceedings of the Ciba Foundation Symposium on Mobility and Function in Proteins and Nucleic Acids (1983) is also a useful reference.

SECTION IV: The Thesis.

Each of the chapters in this thesis is independent and includes a fairly detailed introduction to the specific topic. All of the work reported was done on the small protein, myoglobin, which has about 150 amino acids (depending on the species), and is an oxygen storage protein similar in conformation to one of the subunits of hemoglogin.

An excellant introduction to the structure and function of myoglobin and hemoglobin is the book by Dickerson and Geis (1983). The protein envelopes a protoporphyrin IX (heme) group which contains an iron. The iron is coordinated to four nitrogens in the heme group and to the nitrogen of a histidine, known as the proximal histidine. This is the only strong linkage to the protein; the sixth coordination site is either unoccupied (in deoxy Fe II myoglobin) or taken up by a water molecule (in the stable, but physiologically inactive, form at normal ph: Fe III (met) myoglobin), or an oxygen molecule (in oxy Fe II myoglobin). Several other ligands, such as carbon monoxide, azide and cyanate also can bind to the heme iron in the sixth coordination position (Antonini and Brunori, 1971).

The most commonly studied myoglobin is that from sperm whales; these diving mammals have large amounts of the oxygen storage protein in their tissues. Sperm whale myoglobin has 153 amino acids; two important residues are the proximal histidine (93), which is bound to the iron, and the distal histidine (64). His 64 is close to the binding site of the sixth ligand (on the opposite, or distal, face of the heme group from the proximal histidine) and is implicated in controlling binding of the ligand to the heme. Myoglobin is almost entirely helical; except for a few short stretches of random coil (called loops), all the amino acids are distributed among eight alpha helices, labelled A,B,...,H. A common numbering system that is adopted is to label each residue by its position in a helix (eg., A3, B4) or loop (eg. CD5).

Though myoglobin does not exhibit the allosteric properties that have made hemoglobin a fascinating system to study (Perutz, 1978), it does show an interesting, rather machine-like, behaviour on ligand binding. The iron atom is out of the plane of the heme in deoxy myoglobin. On ligand binding it moves towards the heme plane. Strong ligands, such as CO, cause the iron to be completely in the heme plane. This motion of the iron towards the heme is tracked by the proximal histidine, and consequently by the F helix, to which the histidine is rather rigidly attached. Thus, ligand binding results in a small, but global, change in the structure of the protein and makes the system a prototype of hemoglobin, where the motion of the iron initiates much larger changes (Dickerson and Geis, 1983).

Another reason for interest in myoglobin is that the ligand binding pocket (the sixth coordination site) is buried in the protein. Consequently, migration of the ligand from the solvent and into the binding pocket is a process occurring over protein mediated barriers (Frauenfelder and Wolynes, 1985). A large number of studies have focussed on breaking the iron-ligand bond by flash photolysis and examining the time dependance of the subsequent rebinding of the ligand or the structure of the photo-product immediately after dissociation. Interpretation of the results of these studies is still an area of active research, but it is one that is leading to an increased understanding of protein energetics and dynamics (for example, see Ansari et al., 1985, Henry et al., 1983, Fiamingo and Alben, 1985, Dasgupta et al., 1985).

Finally, myoglobin is a good system for crystallographic studies because it crystallizes readily and diffracts well. The maximum resolution of the data used for work reported here is 1.5Å, but data to 1.2Å have been collected on met-myoglobin (Petsko and Kuriyan, 1985, unpublished) and it is hoped that some form of anisotropic refinement will be possible against these data.

Summaries of the Chapters

Chapter (2). Molecular Dynamics and Protein Refinement: In this chapter an attempt is made to understand how motion affects the results of protein refinement. Molecular dynamics trajectories are used to calculate time-averaged diffraction data for myoglobin. Three positional coordinates and one isotropic temperature factor are then refined against these simulated data; the refined parameters are compared with the exact results obtained directly from the simulation and the sources of error are analysed. It is found that the more mobile atoms in the protein have multiple peaks in their distribution functions and that the refinement program invariably fits only part of the distribution, resulting in errors in the refined positions and under-estimation of the temperature factors. The use of stereochemical and temperature factor restraints is also examined. A surprising result is that positional errors for the less mobile atoms are reduced if tight stereochemical restraints are used. The restraints on temperature factors are found to be too restrictive, in keeping with the conclusions of Yu, Hendrickson and Karplus (Yu et al., 1985).

Chapter (3). The Thermal Expansion of Myoglobin: This study is based on 80K data collected by Parak and co-workers (Hartmann et al., 1982) and on room temperature data collected by Petsko (Frauenfelder et al., 1979) and Kuriyan (this thesis). Despite large errors in the 80K data it is shown that a small, but systematic, expansion in the structure of the protein occurs, corresponding to an average increase of 0.20Å in C_α-C_α distances. One component of the expansion is simply an overall increase in interatomic distances in the protein. Another component is more local and is probably correlated with changes in the unit cell between the two temperatures: the changes are such that the distribution of intermolecular contacts is approximately preserved.

Chapter (4). The Structure and Refinement of CO-Myoglobin: The structure of CO-myoglobin has been refined at a resolution of 1.5Å. This chapter discusses the structure of the CO-ligand, and the local and global changes in the protein structure induced by by the ligand. The ligand binds to the protein in more than one conformation and is distorted from the linear conformation seen in model compounds; part of the strain is taken up by the protein and there are a number of local changes around the ligand that lead to a larger binding cavity than in deoxy-myoglobin. The iron is in the plane of the heme in CO-myoglobin. This motion of 0.3-0.4Å relative to its position in deoxy-myoglobin results in a compaction of the proximal side of the protein. Finally, a possible pathway for ligand entry into the binding cavity is indicated by the alternate conformations found for an arginine sidechain near the distal histidine.

Chapter (5). Estimation of Errors in the Coordinates and Temperature Factors: This chapter deals with two problems. The first is the estimation of errors in the refined parameters for a single crystal structure and the second is the variation of temperature factors from crystal to crystal due to changes in the static disorder. The errors in a particular crystal structure are estimated by perturbing a refined structure by energy minimization (without reference to the X-ray data) and then continuing the X-ray refinement. Comparison of the final structure with the initial one allows an estimation of the errors in the structure. This procedure also results in a structure which fits the X-ray data as well as the intial structure, but has a lower internal energy. The effect of static disorder is evaluated by comparing temperature factors obtained by refinement against data from different crystals, including one which was used to collect both a met- and a CO-myoglobin data set. It is found that static disorder can cause temperature factors to vary by 35-40% and that a simple translational disorder model (i.e., a constant offset in the temperature factors) is inadequate to account for the observed effects.

Acknowledgements

My understanding of refinement methods has greatly benefitted from discussions with Bill Weis, Bob Tilton, Toshiko Ichiye, Steve Burley, Bob Davenport and, of course, Greg Petsko and Martin Karplus.

REFERENCES

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

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D. (1983) "The Molecular Biology of the Cell", Garland Publishing, New York, 1146 pages.

Ansari, A., Berendzen, J., Bowne, S.F., Frauenfelder, H.F., Iben, I.E.T., Sauke, T.E., Shyamsunder, E. and Young, R.D. (1985) Proc. Natl. Acad. Sci. (U.S.A.).

Amoros, J.L., and Amoros, M. (1968) "Molecular Crystals: Their Transforms and Diffuse Scattering", John Wiley and Sons, New York.

Antonini, E. and Brunori, M. (1971) "Hemoglobin and Myoglobin in Their Reaction with Ligands", North-Holland Publishing Co., Amsterdam.

Artymiuk, P.J., Blake C.C.F., Grace D.E.P., Oatley S.T., Phillips, D.C., and Sternberg, M.J.E. (1979) Nature, 280, 563-568.

Artymiuk, P.J. and Blake, C.C.F. (1981) J. Mol. Biol. 152, 737-762.

Baker, E.N. (1980) J. Mol. Biol. 141,441-484.

Barlow, D.J., Baum, J.O., Drummond, M.L.J., Finney, J.L., Smith, J.C. and Thornton, J.M. (1985) in "Amino Acids, Peptides and Proteins", Vol. 16, J.H. Jones (ed.). Specialist Periodical Report of the Royal Society of Chemistry (London), pages 189-203.

Blake, C.C.F., Pulford, W.C.A. and Artymiuk, P.J. (1983) J. Mol. Biol. 167, 693-723.

Blundell, T. and Johnson, L.N. (1976) "Protein Crystallography", Academic Press, New York.

Branden, C.-I, Holmes, K.C. and Kendrew, J.C. (1963) Acta Cryst. A16, 175.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M. (1983) J. Comp. Chem., 4, 187-217.

Brooks, B.R. and Karplus, M. (1983) Proc. Natl. Acad. Sci. (U.S.A.) 80, 6571.

Brooks, B.R. and Karplus, M. (1985) Proc. Natl. Acad. Sci. (U.S.A.) 4995-4999.

Brooks, F.P. and Pique, M. (1985) in "Molecular Dynamics and Protein Structure", J. Hermans, ed. page 109 (see complete reference under Hermans, J).

Burkert, U. and Allinger, N.L. (1982) "Molecular Mechanics", A.C.S. Monograph 177, American Chemical Society, Washington D.C., 319 pages.

Ciba Foundation Symposium 93 (1983) "Mobility and Function in Proteins and Nucleic Acids", Pitman, London, 345 pages.

Chambers, J.L. and Stroud R. M. (1979) Acta Cryst. B35 1861-74.

Clementi, E. and Sharma, R. (eds.) (1983) "Structure and Dynamics: Nucleic Acids and Proteins", Adenine Press, New York.

Dasgupta, S., Spiro, T.G., Johnson, C.K., Dalickas, G.A. and Hochstrasser, R.M. (1985) Biochemistry 24, 5295-5297.

Debrunner, P.G. and Frauenfelder, H. (1982) Ann. Rev. Phys. Chem. 33, 283.

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

Diamond, R. (1971) Acta Cryst. A27, 436-452.

Diamond, R. (1985) Meth. Enzymol. 115 (b), 237-251.

Dickerson, R.E. and Geis, I. (1969) "The Structure and Action of Proteins", Benjamin/Cummings, Menlo Park, CA. 118 pages.

Dickerson, R.E. and Geis, I. (1983) "Hemoglobin: Structure, Function, Evolution and Pathology", Benjamin/Cummings, Menlo Park, CA.

Deisenhofer, J., Remington, S.J. and Steigemann, W. (1985) Meth. Enzymol., 115 (b), 303-323.

Dodson, E.J. (1980) in "Refinement of Protein Structures; Proceedings of the Daresbury Study Weekend" (ed. Machin, P.A. and Elder, M.) Science and Engineering Research Council, Daresbury Laboratory, U.K., pages 29-39.

Dodson, E.J., Isaacs, N.W. and Rollett, J.S. (1976) Acta Cryst. 32, 311-315.

Dunitz, J. (1979) "X-ray Analysis and the Structure of Organic Molecules", Cornell University Press, Ithaca, 514 pages.

Fiamingo, F.G. and Alben, J.O. (1985) Biochemistry, 24, 7964-7970.

Frauenfelder, H., Petsko, G.A., and Tsernoglou, D. (1979) Nature, 280, 558-563.

Frauenfelder, H. and Wolynes, P.G. (1985) Science (Washington, D.C.) 229, 337-345.

Freer, S.T. (1985) Meth. Enzymol., 115 (b), 235-237.

Friedman, J.M. (1985) Science (Washington, D.C.), 228, 1273-1280.

Fujinaga, M., Delbaere, L.T.J., Brayer, G.D. and James, M.N.G. (1985) J. Mol. Biol. 183, 479-502.

Gelin, B.R., Lee, A. W.-M. and Karplus, M. (1983) J. Mol. Biol. 171, 489-559.

Glover, I.D., Haneef, I., Pitts, J.E., Wood, S.P., Moss, D.S., Tickle, I.J. and Blundell, T.L. (1983) Biopolymers 22, 293.

Go, N., Noguti, T. and Nishikawa (1983) Proc. Natl. Acad. Sci. (U.S.A.) 80, 3696-3700.

Gramaccioli, C.M. and Filippini, G. (1983) Acta Cryst. A39, 784-791.

Greer, J. (1985) Meth. Enzymol., 115 (b), 206-224.

van Gunsteren, W.F., Berendsen, H.J.C. and Rullmann J.A.C., (1981) Mol. Phys. 44, 69-95.

van Gunsteren, W.F. and Berendsen, H.J.C. (1982) Biochem. Soc. Trans. 10, 301.

van Gunsteren, W.F. and Berendsen, H.J.C. (1985) in "Molecular Dynamics and Protein Structure" ed. Hermans, J. page 1 (see complete reference under Hermans, J.).

van Gunsteren, W.F., Berendsen, H.J.C., Hermans, J., Hol, W.G.J. and Postma, J.P.M. (1983) Proc. Natl. Acad. Sci. (U.S.A.) 80, 4315.

Gurd, F.R.N. and Rothgeb, T.M. (1979) Adv. Prot. Chem. 34,166-339.

Haneef, I., Moss, D.S., Stanford, D.S. and Borkakoti, N. (1985a) Acta Cryst. A40, 426-433.

Haneef, I., Glover, I.D., Tickle, I.J., Moss, D.S., Pitts, J.E., Wood, S.P., Blundell, T.L., Hermans, J. and van Gunsteren, W.F. (1985b) in "Molecular Dynamics and Protein Structure", Hermans, J. (ed.) (see complete reference under Hermans, J.) pages 85-91.

Hartmann, H., Parak, F., Steigemann, W., Petsko, G.A., Ponzi, D.R., and Frauenfelder, H. (1982) Proc. Natl. Acad. Sci. (USA) 79, 4967-4971.

Hendrickson, W.A. (1980) in "Refinement of Protein Structures; Proceedings of the Daresbury Study Weekend" (ed. Machin, P.A. and Elder, M.) Science and Engineering Research Council, Daresbury Laboratory, U.K.

Hendrickson, W.A. (1985) Meth. Enzymol., 115 (b), 252-270.

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

Hendrickson W.A. and Konnert, J.H. (1981) in "Biomolecular Structure, Conformation, Function and Evolution", Vol. 1, R. Srinivasan (ed.) Pergamon Press, Oxford, pages 43-57.

Henry, E.R., Sommer, J.R., Hofrichter, J. and Eaton, W.A. (1983) J. Mol. Biol. 166, 443-451.

Hermans, J. (ed.) (1985a) "Molecular Dynamics and Protein Structure" Proceedings of a Workshop held 13-18 May, 1984 at the University of North Carolina, 194 pages. Copies can be orderd from Polycrystal Book Service, P.O. Box 27, Western Springs, IL 60558.

Hermans, J. (1985b) Meth. Enzymol., 115 (b), 171-189.

Hermans, J. Berendsen, H.J.C., van Gunsteren, W.F. and Postma, J.P.M. (1984) Biopolymers, 23, 1513-1518.

Hilinski, E.F. and Rentzepis, P.M. (1983) Nature (London) 302, 481-487.

Holbrook, S.R. and Kim, S.H. (1984) J. Mol. Biol. 173, 361-388.

Holbrook, S.R. and Kim, S.H. (1985) in "Molecular Dynamics and Protein Structure" Hermans, J. (ed.) (see complete reference under Hermans, J.) pages 83-85.

Honzatko, R.B., Hendrickson, W.A. and Love, W.A. (1985) J. Mol. Biol. 184, 147-64.

Huber, R. (1979) Trends in Biochem. Sci. 4, 271-276.

Hughes, E.W. (1941) J. Am. Chem. Soc. 63, 1737.

Ichiye, T. and Karplus, M. (1986) Biopolymers, submitted.

International Tables for X-Ray Crystallography, (1974) Vol. IV, (ed. Ibers, J. and Hamilton, W.C.), International Union of Crystallography, The Kynoch Press, Birmingham, U.K.

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

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

James, M.N.G. and Sielecki, A. (1983) J. Mol. Biol. 163, 299-361.

James, R.W. (1948) The Optical Priciples of X-Ray Diffraction, Reissued 1982, Ox Bow Press, Woodbridge, CT.

Jensen, L.H. (1985) Meth. Enzymol., 115 (b), 227-234.

Jones, T.A. (1982) in "Computational Crystallography" (ed. Sayre, D.) Clarendon, Oxford, 3303-3317.

Jones, T.A. (1985) Meth. Enzymol., 115 (b), 157-171.

Jones, T.A. and Liljas, L. (1984) Acta Cryst. A40, 50-57.

Karplus, M. (1981) Ann. of the N.Y. Acad. Sci. 367, 407-18.

Karplus, M. (1984) Adv. in Biophys. 18, 165.

Karplus, M. (1985) in "Molecular Dynamics and Protein Structure" ed. Hermans, J. pages 1-2 (see complete reference under Hermans, J.).

Karplus, M. and McCammon, J.A., (1981) C.R.C. Critical Reviews in Biochemistry, 9, 293-349.

Karplus, M. and McCammon, J.A. (1983) Ann. Rev. Biochem. 53, 263-300.

Karplus, M. and Kushick, J.N. (1981) Macromolecules 14, 325.

Keller, H. and Debrunner, P.G. (1980) Phys. Rev. Lett. 45, 68-71.

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G. and Wyckoff, H. (1958) Nature (London), 181, 662.

Kendrew, J.C., Dickerson, R.E., Strandberg, B.E., Hart, R.G., Davies, D.R., Phillips, D.C., and Shore, V.C. (1960) Nature (London), 185, 422-427.

Knapp, E.W., Fischer, S.F. and Parak, F. (1983) J. Chem. Phys. 78, 4701-4711.

Knossow, M., Lewis, M., Rees, D. and Wiley, D.C. (1986) Acta Cryst., to be submitted.

Kollman, P. (1985) Acc. Chem. Res. 18, 105-111.

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

Konnert, J.H. and Hendrickson, W.A. (1980) Acta Cryst. A36, 344-349.

Lakowicz, J.R., Maliwal, B.P. Cherek, H. and Balter, A. (1983) Biochemistry, 22, 174.

Levitt, M. (1982) Ann. Rev. Biophys. Bioeng. 11, 251.

Levitt, M. (1974) J. Mol. Biol. 82, 393-420.

Levitt, M., Sander, C. and Stern, P.S. (1985) J. Mol. Biol. 181, 423-447.

Lifson, S., Hagler, A.T. and Dauber, P. (1979a) J. Am. Chem. Soc. 101, 5111-5121.

Lifson, S., Hagler, A.T. and Dauber, P. (1979b) J. Am. Chem. Soc. 101, 5122-5130.

Lipari, G., Szabo, A. and Levy, R. M. (1982) Nature (London) 300, 197.

Luzzati, P.V. and Taupin, D. (1984) J. Appl. Cryst. 17, 273-285.

Matthews, B.W. (1968) J. Mol. Biol. 33, 491-497.

McCammon, J.A., Gelin, B.R., and Karplus, M. (1977) Nature 267, 585-590.

McCammon, J.A. and Karplus, M. (1983) Acc. Chem. Res. 16, 187.

McCammon, J.A. (1984) Report Prog. Phys. 47, 1.

McPherson, A. (1982) "Preparation and Analysis of Protein Crystals", John Wiley and Sons, New York, 371 pages.

Moss, D.S. and Morffew, A.J. (1982) Comput. Chem. 6, 1-3.

Moult, J. and James, M.N.G. (1985) in "Molecular Dynamics and Protein Structure", Hermans, J. (ed.) (see complete reference under Hermans, J.) pages 81-82.

North, A.C.T. and Phillips, D.C. (1969) Prog. Biophys. Mol. Biol. Butler, J.A.V. and Noble, D. (eds.), Pergammon Press, Oxford, pages 1-132.

North, A.C.T. and Smith, J.C. (1985) Int. J. Biol. Macromoleoules, 7, 223-225.

Northrup, S.H. and McCammon, J.A. (1980) Biopolymers, 19, 1001-1016.

Northrup, S.H., Pear, M.R., Lee, C.Y., McCammon, J.A. and Karplus, M. (1982) Proc. Natl. Acad. Sci. (U.S.A.) 79, 4035.

Novotny, J., Bruccoleri, R.E. and Karplus, M. (1984) J. Mol. Biol. 177, 787-818.

Perutz, M.F. (1978) (December) Scientific American, 92-123.

Petsko, G.A. and Ringe, D. (1984) Ann. Rev. Biophys. Bioeng. 13, 331-71.

Pettitt, B.M. and Karplus, M. (1986) to be published.

Phillips, G. N., Fillers, J.P. and Cohen, C. (1980) Biophys. J., 32, 484-502.

Phillips, S.E.V. (1980) J. Mol. Biol. 142 531-54.

Quicho, F.A. and Vyas, N.K. (1984) Nature (London) 310, 381-386.

Raftery, J., Sawyer, L. and Pawley, G.S. (1985) J. Appl. Cryst. 18, 424-429.

Reiher, W.E. (1985) "Theoretical Studies of Hydrogen Bonding", Ph.D. Thesis, Department of Chemistry, Harvard University, Cambridge MA.

Remington, S., Wiegand, G. and Huber, R. (1982) J. Mol. Biol. 158, 111-152.

Richardson, J.S. (1981) Adv. Prot. Chem. 34, 166-339.

Richardson, J.S. and Richardson, D.C. (1985) Methods in Enzymology, 115 (b), 189-206.

Rollett, J.S. (1970) "Crytallographic Computing", Munksgaard, Copenhagen.

Rollett, J.S. (1982) in "Computational Crystallography", Sayre, D. (ed.) Clarendon Press, Oxford, pages 338-353.

Schomaker, V. and Trueblood, K.N. (1968) Acta Cryst. B24, 63-76.

Sheriff, S., Hendrickson, W.A., Stenkamp, R.E., Sieker, L.C. and Jensen, L.H. (1985) Proc. Natl. Acad. Sci. (USA) 82, 1104-07.

Sielecki, A.R., Hendrickson, W.A., Broughton, C.G., Delbaere, L.T.J., Brayer, G.D. and James, M.N.G. (1979) J. Mol. Biol. 134, 781-804.

Smith, J.C., Brooks, B.R., Cusack, S., Finney, J.L., Pezzeca, U. and Karplus, M. (1986) J. Chem. Phys., submitted.

Sparks, R.A. (1985) Meth. Enzymol., 115 (b), 23-41.

Stevenson, A.W. and Harada, J. (1983) Acta Cryst. 39, 202-207.

Stewart, R.F. and Feil, D. (1980) Acta Cryst. A36 503-509.

Stuart, D.I. and Phillips, D.C. (1985) Meth. Enzymol., 115 (b), 117-142.

Sussman, J.L. (1985) Meth. Enzymol., 115 (b), 271-303.

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

Takano, T. (1977) J. Mol. Biol. 110, 537-568

Ten Eyck, L.F. (1973) Acta Cryst., A29, 183-191

Ten Eyck, L.F. (1977) Acta Cryst., A33, 486-492

Ten Eyck, L.F. (1985) Meth. Enzymol., 115 (b), 324-337.

Teeter, M.M. and Hendrickson, W.A. (1979) J. Mol. Biol., 127, 219-233.

Tsukada, H. and Blow, D.M. (1985) J. Mol. Biol. 184, 703-711.

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

Watenpaugh, K.D. (1985) Meth. Enzymol., 115 (b), 3-15.

Watenpaugh, K.D., Sieker, L.C., Herriott, J.R. and Jensen, L.H. (1973) Acta Cryst. B29, 943-956.

Watenpaugh, K.D., Margulis, T.N., Sieker, L.C. and Jensen, L.H. (1978) J. Mol. Biol. 122, 175-190.

Watenpaugh, K.D. Sieker,L.C., and Jensen L.H. (1980) J. Mol. Biol. 138, 615-633.

Watson, H.C., Kendrew, J.C., Coulter, C.L., Branden, C.-I., Phillips, D.C. and Blake, C.C.F. (1963) Acta Cryst. A16, 81.

Watson, J.D. (1976) "Molecular Biology of the Gene" (3rd. Edition) Benjamin/Cummings, Menlo Park, CA.

Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S. and Weiner, P. (1984) J. Am. Chem. Soc. 106, 765.

Westof, E., Dumas, P. and Moras, D. (1985) J. Mol. Biol. 184, 119-145.

Willis, B.T.M. and Pryor, W. (1975) "Thermal Vibrations in Crystallography", Cambridge Univ. Press, London.

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

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

Wilson, K.S., Stura, E.A., Wild, D.L., Todd, R.J., Stuart, D.I., Babu, Y.S., Jenkins, J.A., Standing, T.S., Johnson, L.N., Fourme, R., Kahn, R., Gadet, A., Bartels, K.S. and Bartunik, H.D. (1983) J. Appl. Cryst. 16, 28-41.

Wlodawer, A., Walter, J., Huber, R. and Sjolin, L. (1984) J. Mol. Biol. 180 301-329.

Woodward, C.K. and Hilton, B.D. (1979) Ann. Rev. Biophys. Bioeng. 8, 99-127.

Woolfson, M.M. (1970) "An Introduction to X-Ray Crystallography" Cambridge University Press, Cambridge, U.K., 380 pages.

Wyckoff, H.W., Hirs, C.H.W. and Timasheff, S.N. (editors) (1985) "Diffraction Methods for Biological Macromolecules", Methods in Enzymology, Volume 115, Academic Press, Orlando, FL.

Yu, H. Karplus, M. and Hendrickson, W. (1985) Acta Cryst., B41, 191-201.

Zubay, G. (Coordinating Author) (1983) "Biochemistry", Addison Wesley, Reading, MA. 1268 pages.

Zucker, U.H. and Schulz, H. (1982) Acta Cryst., A38, 563-568.

Chapter 2

The Effect of Anisotropy and Anharmonicity on Protein Crystallographic Refinement: An Evaluation by Molecular Dynamics

Abstract

Molecular dynamics simulations are used to evaluate the errors introduced by anharmonicity and anisotropy in the structure and temperature factors of proteins obtained by refinement of X-ray diffraction data. 25 ps and 300 ps simulations of met-myoglobin are used to generate time-averaged diffraction data at 1.5Å resolution. The crystallographic restrained-parameter least-squares refinement program (PROLSQ, Konnert and Hendrickson,1980) is used to refine models against these simulated data. The resulting atomic positions and isotropic temperature factors are compared with the average structure and fluctuations calculated directly from the simulations. It is found that significant errors in the atomic positions and fluctuations are introduced by the refinement and that the errors increase with the magnitude of the atomic motions. Of particular interest is the fact that the refinement generally underestimates the atomic motions. Moreover, while the actual motions go up to a mean-square value of about 5Å^2, the X-ray results never go above approximately 2Å^2. This systematic deviation in the motional parameters appears to be due to the use of a single-site isotropic model for the atomic fluctuations. Many atoms have multiple peaks in their probability distribution functions. For some atoms the multiple peaks are seen in difference electron density maps and it is possible to include these in the refinement as disordered residues. However, for most atoms the

refinement fits only one peak and neglects the rest, leading to the observed errors in position and temperature factor. The use of strict stereochemical restraints is inconsistent with the average dynamical structure; nevertheless, refinement with tight restraints results in structures that are comparable to those obtained with loose restraints and better than those obtained with no restraints. The results support the use of tight stereochemical restraints, but indicate that restraints on the variation of temperature factors are too restrictive.

I. Introduction

It is now recognized from a variety of experimental and theoretical studies that significant atomic motions occur in macromolecules of biological interest. Information concerning both the magnitudes and the time scales of the motions are available (Karplus and McCammon, 1981,1983). At room temperature thermal atomic displacements are in the range of 0.2 to 2.0 Å and vary significantly for different regions of the protein; their time scale is from 0.1 to 50 picoseconds (ps) (with the longer time scales generally associated with the large amplitudes). The Debye-Waller (temperature) factors evaluated in X-ray crystallographic refinements of protein structures are an important source of experimental data concerning the magnitudes of the fluctuations (Petsko and Ringe, 1984); this is based on the identification of the temperature factors with the mean-square fluctuations of individual atoms. With the assumption of isotropic and harmonic motion, temperature factors for all the non-hydrogen atoms have been determined for many proteins and some examples are given by Artymiuk et al. (1979), Frauenfelder et al. (1979), Watenpaugh et al. (1980), Takano and Dickerson (1981) and Sheriff et al. (1985).

It is clear, however, from molecular dynamics simulations that the atomic motions are highly anisotropic (Karplus and McCammon, (1981,1983), Northrup et al., (1981), van Gunsteren and Karplus, (1982a), Ichiye and Karplus,(1985a,b)), and, at least for some of the atoms, strongly anharmonic (Levy et al. (1985), Mao et al. (1982), van Gunsteren and Karplus, (1982,a,b), Ichiye and Karplus,(1985a,b)). Since neither of these deviations from the simple model are included in most

X-ray determinations of the structures of macromolecules, it is of interest to evaluate the errors introduced in the refinement process by their neglect. Such errors can involve the positions of atoms as well as their motional properties.

An evaluation of the errors is particulary important now that improved data sets can be obtained for macromolecules and more attention is being focused on deducing the motional properties by refinement of this data. With the advent of new techniques such as the use of area-detectors (Xuong et al., 1978), low-temperature crystallography (Hartmann et al., 1982) and intense X-ray sources available from synchrotrons (Wilson et al., 1984), significant improvements in the quality of protein diffraction data are expected. For several small proteins, such as bovine pancreatic trypsin inhibitor (BPTI), rubredoxin and crambin, the reflections have already been measured out to 1.2 to 1.0 Å (Wlodawer, et al. 1985, Watenpaugh et al. 1980, Teeter and Hendrickson, 1979) and it is also possible to collect high resolution data (i.e. 1.2-1.0 Å) for larger proteins such as ribonuclease, lysozyme and myoglobin at low temperature (R.F.Tilton and J.Dewan, personal communication). This will make it possible to probe more deeply into the nature of protein motions and their effects on the X-ray refinement procedure. For a few proteins, anisotropic harmonic temperature factors have been introduced, resulting in six thermal parameters per atom (for example, see Watenpaugh et al., 1980). Anharmonic corrections have not been used for proteins, although they have been employed in small molecule refinements (Zucker and Schulz, 1983).

A direct experimental estimate of the errors resulting from the assumption of isotropic, harmonic temperature factors is difficult because sufficient data are not yet available for protein crystals. Moreover, any data set includes other errors which would obscure the analysis, and the specific correlation of temperature factors and motion is complicated by the need to account for static disorder in the crystal. As an alternative to an experimental analysis of the errors in refinement procedures for proteins, we describe here a theoretical approach. The basic idea is to generate X-ray data from the results of a molecular dynamics simulation of a protein and to use these data to obtain a refined structure by standard methods. The error in the analysis is determined by comparing the results obtained from the refinement procedure with the known average structure and the mean square fluctuations of the original simulation. This type of comparison, in which no real experimental results are used, avoids problems due to inaccuracies in the measured data (exact calculated intensities are used), crystal disorder (there is none in the model), and due to approximations in the simulation (the simulation gives exact results for this case). The only question about such a comparison is whether the atomic motions found in the simulation are a meaningful representation of those occurring in proteins. A variety of comparisons (Karplus and McCammon, 1981,1983, Levy and Keepers, 1985) suggest that molecular dynamics simulations provide a reasonable picture of the motions inspite of the errors in the potentials, neglect of the crystal environment and the finite time classical trajectories used to obtain the results. However, as already stated, these inaccuracies do not affect the exactness of the computer "experiments" and their interpretation given in this paper. This strategy for analysing a crystallographic refinement model is similar to that used previously to analyse N.M.R. relaxation models for proteins (Levy et al., 1981).

A 25 picosecond (ps) molecular dynamics trajectory for myoglobin is used to carry out the test of the refinement procedure outlined above; the tests were also done using a 300ps trajectory of myoglobin, but the results of the shorter simulation will be the focus of most of the discussion. The average structure and the mean square fluctuations from that structure are calculated directly from the trajectory. To obtain the average electron density, appropriate atomic electron distributions are used for the individual atoms in each coordinate set in the trajectory and averaged. Given the symmetry, unit cell dimensions and position of the myoglobin molecule in the cell, average structure factors are calculated as the Fourier transform of the averaged electron densities. The resulting intensities at the Bragg reciprocal lattice points are used as input data for the widely applied crystallographic program, PROLSQ (Konnert and Hendrickson, 1980). The time-averaged atomic positions obtained from the simulation and a uniform temperature factor provide the initial model structure. The positions and an isotropic, harmonic temperature factor for each atom are then refined iteratively against the computer generated intensities in the standard way. PROLSQ is a restrained-parameter least-squares refinement program, and the refinements are done with tight, loose and no restraints on the parameters.

Differences between the refined results for the average atomic positions and their mean square fluctuations and those obtained from the molecular dynamics trajectory are due to errors introduced by the refinement procedure. Since these differences turn out to be significant and systematic, the simulation results concerning the magnitude, anisotropy, and anharmonicity of the motions are used to examine the source of the errors in the refinement.

Sect. II outlines the methods used in this study. The approach used to generate the X-ray intensities from the simulation results and the details of the procedures employed for refining the data are described. In Sect. III the results are presented and analysed. Emphasis is placed on the atomic positions, the stereochemistry of the structure, and the atomic motions. The conclusions are outlined in Sect. IV.

II METHODS

IIa The Calculation of Diffraction Intensities for a Static Structure

For a perfect crystal with no thermal motion, the intensity of scattered X-rays is proportional to the square of the Fourier transform of the electron density in a unit cell (Woolfson, 1970). The Fourier transform of the electron density is called the "structure factor", $F(\underset{\sim}{Q})$, where $\underset{\sim}{Q}$ is the scattering vector, defined by:

$$\underset{\sim}{Q} = \frac{2\pi(\underset{\sim}{e}-\underset{\sim}{e}_0)}{\lambda} \qquad (1a)$$

where $\underset{\sim}{e}_0$ and $\underset{\sim}{e}$ are unit vectors along the directions of the wave vectors of the incident and scattered radiation, respectively, and λ is the wavelength of the X-radiation. In terms of the reciprocal lattice vectors, $\underset{\sim}{a}^*$, $\underset{\sim}{b}^*$ and $\underset{\sim}{c}^*$, $\underset{\sim}{Q}$ is given by:

$$Q = 2\pi\ (\ h\underset{\sim}{a}^* + k\underset{\sim}{b}^* + \underset{\sim}{l}^*\ )\ \equiv 2\pi\ \underset{\sim}{H} \qquad (1b)$$

where h, k and l are not, in general, required to be integral and $\underset{\sim}{H}$ is the position vector in reciprocal space (Willis and Pryor, 1975). The structure factor, $F(\underset{\sim}{Q})$, is thus given by:

$$F(\underset{\sim}{Q}) = \int d\underset{\sim}{r}\rho(\underset{\sim}{r})e^{i\underset{\sim}{Q}\cdot\underset{\sim}{r}} \qquad (2)$$

where $\rho(\underset{\sim}{r})$ is the electron density at $\underset{\sim}{r}$ and the integral is over the unit cell. In most crystallographic applications the molecular electron density is approximated by a superposition of the electron densities of the individual atoms. The electron density at a point $\underset{\sim}{r}$ in the unit cell is then given by:

$$\rho(\underset{\sim}{r}) = \sum_{i=1}^{N} \rho_i(\underset{\sim}{r}-\underset{\sim}{r}_i) \qquad (3)$$

where the sum runs over the N atoms of the one or more molecules in the asymmetric unit of the unit cell, and $\rho_i(\underset{\sim}{r}-\underset{\sim}{r}_i)$ is the electron density at $\underset{\sim}{r}$ due to an atom at $\underset{\sim}{r}_i$. Substituting Eq. (3) in Eq. (2) above we obtain:

$$F(\underset{\sim}{Q}) = \sum_{i=1}^{N} \int d\underset{\sim}{r}\rho_i(\underset{\sim}{r}-\underset{\sim}{r}_i)e^{i\underset{\sim}{Q}\cdot\underset{\sim}{r}} \tag{4a}$$

$$= \sum_{i=1}^{N} f_i(\underset{\sim}{Q})e^{i\underset{\sim}{Q}\cdot\underset{\sim}{r}_i}$$

where $f_i(\underset{\sim}{Q})$ is the atomic scattering factor of the i^{th} atom:

$$f_i(\underset{\sim}{Q}) = \int d\underset{\sim}{r}\rho_i(\underset{\sim}{r})e^{i\underset{\sim}{Q}\cdot\underset{\sim}{r}} \tag{4b}$$

The atomic electron densities are obtained from ab-initio quantum mechanical calculations and are Fourier transformed to obtain atomic scattering factors. These atomic scattering factors are, in general, complicated anisotropic expressions and a further simplification is made by fitting a simple analytic, isotropic function to the ab-initio scattering factors. The most commonly used form is a sum of two to four gaussians plus a constant. Defining $s = \frac{|Q|}{4\pi}$, we have

$$f(s) = \sum_{\alpha=1}^{n} a_\alpha e^{-b_\alpha s^2} + c \tag{5}$$

The parameters a_α, b_α, and c are obtained by a least-squares fit to the ab-initio scattering factors. Parameters for a large number of atoms and ions are available for fits using two gaussians (Moore,1963) and four gaussians (International Tables for X-ray Crystallography, 1969). We note that the electron density corresponding to the scattering factor in Eq. 5 is a sum of gaussians plus a delta function centered at the position of the atom.

At the current resolution limits of crystallographic data on proteins, Eqns. (4) and (5) are of sufficient accuracy (F.H. Moore,1963) and so the structure factors can be readily computed as a sum of isotropic atomic scattering factors and phase factors (Eqn. 4). Such a calculation, referred to as a direct summation, is very expensive for large molecules (Ten Eyck, 1973, Agarwal, 1978); it takes about one half hour on a VAX 11/780 for a 1.5Å resolution calculation on myoglobin. We take an alternative approach, which is to use fast Fourier transform (FFT) algorithms to calculate structure factors. Programs to do this have been available since about 1973 (Ten Eyck 1973,1977, Agarwal, 1978). Such a calculation proceeds in two steps. The first is the calculation of the electron density in the asymmetric unit of the molecule from a superposition of atomic electron densities and the construction of an electron density grid. In the second step the structure factors are calculated by a finite discrete Fourier transform of the electron density grid using FFT algorithms. In the available programs (Ten Eyck,1977) the atomic electron densities used are obtained from the atomic scattering factors of Moore (1963) which are in the form of two gaussians plus a constant. Direct Fourier tranformation of Eqn. (5) to obtain an expression for the electron density would introduce a delta function into the expression. To avoid this, a psuedo-temperature factor, B_0, is added to each atom. This psuedo-temperature factor can be scaled out of the structure factors after the Fourier transformation, and its inclusion is also important for reducing aliasing errors due to the discrete sampling of the electron density (Ten Eyck, 1977). The resulting model for the atomic electron density is

$$\rho(r) = \sum_{\alpha=1}^{2} \frac{a_\alpha}{\sigma_\alpha^3} \exp\left[-\frac{\pi r^2}{\sigma_\alpha^2}\right] + \frac{c}{\sigma_3^3} \exp\left[-\frac{\pi r^2}{\sigma_3^2}\right] \qquad (6)$$

where a_α, b_α and c are the same as in Eq. 5 and $\sigma_\alpha^2 = \frac{b_\alpha+B_0}{4\pi}$ and $\sigma_3^2 = \frac{B_0}{4\pi}$. These methods are described in detail by Ten Eyck (1973,1977) and by Agarwal (1978), and hence need not be discussed further here.

IIb X-ray Intensities from a Molecular Dynamics Simulation

X-rays scattered from a crystal can be considered as having two components. One of them, usually referred to as the Bragg scatter, exists only at scattering angles that satisfy Bragg's Law and gives rise to the discrete spots observed in diffraction photographs (Willis and Pryor, 1975, see also Appendix 1). The intensity of the Bragg scatter is proportional to the square of the Fourier transform of the average electron density in a unit cell (Stewart and Feil, 1980). The other component is not restricted to the reciprocal lattice points and is referred to as "thermal diffuse scatter" (Willis and Pryor,1975, Amoros and Amoros, 1968), and to compute this one would need information about the correlations of the motions of atoms in one unit cell with those in another. Diffuse scatter also arises due to disorder in the crystal (Amoros and Amoros, 1968) and, though observed in protein crystals (Wilson et al., 1983), its effects are generally ignored in processing and analysing the data. The only way in which the effects of thermal motion on X-ray diffraction data are included for protein crystals is by assuming that the average electron density associated with a given atom is not that obtained if the atom were fixed in position (Eq. 6); instead, it is given by a convolution of the fixed density with a positional probability distribution function arising from the motion of the atom. These distribution functions, also called "thermal smearing functions" (Willis and Pryor, 1975), are the Fourier transforms of the atomic Debye-Waller factors.

In the computer "experiment" to be described here, we calculate average intensities by using a molecular dynamics simulation, which yields a trajectory that gives the position of every atom in the protein as a function of time (Karplus and McCammon, 1981,1983). The electron density for a given coordinate set is calculated by use of Eq. 6 for each atom and the resulting electron densities are averaged over the coordinate sets from the trajectory. It is important to note that no assumed model for the probability distribution functions of the atomic motions is used in this calculation. The averaging of the electron densities over the trajectory is equivalent to convoluting the static electron density with the probability distribution functions obtained from the simulation. This is equivalent to calculating structure factors from coordinate sets sampled from the simulation and averaging them:

$$
\begin{aligned}
I(\underset{\sim}{Q}) \;&\alpha\; \left|\int d\underset{\sim}{r}\langle\rho(\underset{\sim}{r})\rangle e^{i\underset{\sim}{Q}\cdot\underset{\sim}{r}}\right|^2 \\
&= \left|\langle F(\underset{\sim}{Q})\rangle\right|^2
\end{aligned}
\tag{7}
$$

Eqn. 7 is valid if it is assumed that there is no correlation in the motions of different protein molecules in the crystal (see Appendix 1). This assumption, which corresponds to neglecting thermal diffuse scattering, is the standard one made in crystal structure refinements (Willis and Pryor, 1975). An alternative limiting assumption would be that all protein motions are moving in phase, i.e. that at every instant the molecules in the crystal are all identical and they all evolve identically in time (Appendix 1). The correlated motion assumption yields an intensity of the form:

$$I(\underset{\sim}{Q}) \quad \alpha \quad \langle |F(\underset{\sim}{Q})|^2 \rangle \tag{8}$$

We calculated structure factors between 10.0Å and 1.5Å using both Eqn. 7 and Eqn. 8. The crystallographic R-factor, R, is commonly used to indicate the quality of agreement between two sets of structure factors:

$$R = \frac{\sum\limits_{\underset{\sim}{Q}} \big| |F_o(\underset{\sim}{Q})| - |F_c(\underset{\sim}{Q})| \big|}{\sum\limits_{\underset{\sim}{Q}} |F_o(\underset{\sim}{Q})|} \tag{9}$$

The R-factor between structure factors calculated from a 25 ps simulation of myoglobin using Eqn. 7 and Eqn. 8 is 36% (Appendix 1). This is obviously an extreme case, but does suggest that correlations may sometimes be important. Unfortunately the intermediate case is difficult to treat for proteins, though it may be approached by analysis of the lattice modes of crystals. As stated earlier, we follow standard practise and neglect all correlations between unit cells in calculating structure factors; i.e., we use Eqn. 7 for the calculation of data described in this paper to approach most closely the procedure usually followed in protein X-ray refinements.

We decided to do our "experiment" at a resolution of 1.5Å as this is comparable to the resolution of the best X-ray data currently available for proteins the size of myoglobin (Kuriyan et al., 1985a, Phillips,1980). The myoglobin molecule was placed in a crystal lattice of monoclinic system with the symmetry of space group $P2_1$, and with one molecule in the asymmetric unit. The unit cell was assumed to have parameters a = 64.31Å, b = 30.85Å, c = 34.85Å, α = 90.0°,

β = 105.85°, and γ = 90.0°, corresponding to the experimental parameters for met-myoglobin at 300K (Hartmann et al.,1982). The structure factors also depend on the orientation and position of the molecule in the unit cell. To make the situation comparable to the experimental one, the average structure from the dynamics was superimposed, by least-squares, on the experimental structure at 300 K. A small translation of 0.352Å along $\underset{\sim}{a}$, -0.006Å along $\underset{\sim}{b}$, 0.117Å along $\underset{\sim}{c}$ was obtained and applied to all the coordinate sets sampled from the simulation.

Given this space group, unit cell and orientation of the molecule in the unit cell, there are about 22000 unique structure factors (not including Friedel pairs) between 10.0Å and 1.5Å. Calculation of these structure factors for one structure from the simulation using FFT's takes about 5 minutes, which is about 6 times faster than the direct summation. However, much greater savings in computer time can be achieved in calculating the averaged structure factors because the electron density calculation is fast (about 1.5 minutes per structure) and, instead of averaging structure factors, we can average electron density and then do just one Fourier transform at the end.

A program written by L.F.Ten Eyck was used (Ten Eyck,1977). It employs Eqn. 6 to calculate atomic electron densities; they are superimposed to get the molecular electron density, from which the electron density in the asymmetric unit of the unit cell is calculated. The electron density was sampled on a grid with 160 grid points along $\underset{\sim}{a}$, 88 along $\underset{\sim}{b}$, and 88 along $\underset{\sim}{c}$. In the $P2_1$ space group only half of the unit cell along $\underset{\sim}{b}$ needs to be included. The sampling intervals used are much finer than the recommended interval of one-third the minimal interplanar spacing of the data (0.5Å in this case). This, along with a pseudo-temperature factor of 20Å^2 added to each atom (see Eq. 6), serves to reduce errors due to finite sampling (Ten Eyck, 1977). Structure factors were calculated from the electron density grid by use of a set of FFT subroutines written by L. F. Ten Eyck and modified by G. Bricogne (Ten Eyck, 1973, 1977).

The refinement program used in this work, PROLSQ (Konnert and Hendrickson,1980), uses a 4-gaussian form for the atomic scattering factors rather than the 2-gaussian form built into the FFT calculations. The 2-gaussian form is preferable in the electron density calculation as it leads to significant enhancement in the speed of the calculation. The refinement program also computes structure factors by direct summation (Eqn. 4) as the derivatives are more easily obtainable this way. To estimate errors introduced by these discrepancies between the two programs, R-factors between structure factors calculated using 2 and 4 gaussians and by the FFT program and by direct summation were calculated. The R-factor between the FFT and direct summation structure factors using a 2-gaussian electron density model in both cases is 0.64% . With a 4-gaussian model in the direct summation and a 2-gaussian model in the FFT the R-factor is 0.79%. These errors are negligible. Thus it was concluded that no significant advantage would be gained by using the four-gaussian form in the electron density calculation.

IIc The Simulation used and the X-ray Data Sets Generated

Most of the work reported here is based on a 25 picosecond (ps) segment of a 50 ps simulation of myoglobin. Some results are also presented for refinement of data generated from a 300 ps simulation of myoglobin. Both simulations were calculated with a version of the CHARMM program (Brooks et al., 1984). They used identical initial structures and parameters and have been described previously (Levy et. al, 1985). The initial coordinates for the simulation were obtained from a refined crystal structure of met-myoglobin at 250 K (Frauenfelder et al., 1979). The model for the protein included hydrogens only for methyl groups and the total system simulated included 1423 atoms (1217 non-hydrogen protein atoms, 162 methyl hydrogens, 43 heme atoms and one water bound to the iron). No solvent molecules were included. The average temperature of the simulations was 298K (Levy et al., 1985). No hydrogens were included in the structure factor calculations.

The 25 ps simulation was sampled at intervals of both 0.25 ps and 0.05 ps. While sampling the simulation every 0.25 ps the structure factors were calculated and averaged in two different ways, as an internal check. In one case structure factors were calculated for all the 100 molecules sampled from the trajectory, i.e. 100 electron density calculations and Fourier transformations were done to obtain the modulus of the complex mean structure factors. In the other case, the structure factors were calculated by averaging the electron density and doing only one Fourier transform on the averaged electron density. The R-factor between structure factors calculated in the two different ways is 1.6% (at 1.5Å resolution). Each fast Fourier transform introduces a small error due to finite sampling of the electron density (Ten Eyck, 1977) and the second method, which involves only one Fourier transform instead of 100, is expected to be more accurate. However, the differences found here are small compared to the final R-factors of the refined structures. The structure factors calculated using 100 FFT's were used in two of the refinements that follow and will referred to as the $|\langle F\rangle|^{0.25}$set.

The 25 ps simulation was also sampled every 0.05 ps and the averaged electron density obtained from the 500 coordinate sets was used to generate structure factors. This set is referred to as the $|\langle F\rangle|^{0.05}$ data set. Comparison of the $|\langle F\rangle|^{0.25}$ and $|\langle F\rangle|^{0.05}$ sets allows us to check that that the results obtained using the $|\langle F\rangle|^{0.25}$ set were not biased by poor sampling of the trajectory. The $|\langle F\rangle|^{0.05}$ data set was used for four of the refinements reported below. The R-factor between data calculated using sampling intervals of 0.25 ps and 0.05 ps is 3.2% (at 1.5Å resolution). Thus, a small error is introduced by the coarser sampling, but again this is well below the final R-factors of any of the refinements. A 0.25ps sampling interval is adequate to calculate the structure factors. Finally the entire 300 ps of the longer trajectory was sampled every 0.25 ps and the average electron density was determined from 1200 coordinate sets sampled from the simulation. Structure factors were calculated by Fourier transformation of the average density and this data set, which is referred to as the $|\langle F\rangle|^{300}$ set, was used for one of the refinements reported below.

IId Modelling Thermal Motion in Crystallographic Refinement

As mentioned earlier, most refinements of protein structures made to date assume a harmonic, isotropic model for the probability distribution functions. This leads to the following expression for the time averaged structure factors (Willis and Pryor, 1975):

$$\langle F(\mathbf{Q})\rangle = \sum_{j=1}^{N} f_j(\mathbf{Q}) e^{i\mathbf{Q}\cdot\langle \mathbf{r}_j\rangle} e^{W_j(\mathbf{Q})} \tag{10}$$

where, as before, $f_j(\mathbf{Q})$ is the atomic scattering factor and $\langle \mathbf{r}_j\rangle$ is the average position of the jth atom. The term $e^{W_j(\mathbf{Q})}$ is the atomic Debye-Waller factor and in the isotropic case $W_j(\mathbf{Q})$ is given by

$$W_j(\mathbf{Q}) = -\frac{1}{6}\langle \Delta\mathbf{r}_j^2\rangle|\mathbf{Q}|^2 = -\frac{8}{3}\pi^2\langle\Delta\mathbf{r}_j^2\rangle s^2 \tag{11}$$

where $\langle\Delta\mathbf{r}_j^2\rangle$ is the mean-square fluctuation of the j-th atom and $s = \frac{|\mathbf{Q}|}{4\pi}$. The term $\frac{8}{3}\pi^2\langle\Delta r_j^2\rangle$ is referred to as B_j, the atomic B-factor or temperature factor (Willis and Pryor, 1975). Eqn. 10 is a very different type of model from the approach used in the dynamics in that instead of a full averaging of the atomic electron density, as in the dynamics, averages of the position, $\langle\mathbf{r}_j\rangle$, and the mean-square fluctuation, $\langle\Delta\mathbf{r}_j^2\rangle$ are introduced. The average intensity at a reciprocal lattice point, which is what is measured, is proportional to the absolute value squared of the structure factor (Eqn. 7):

$$I(\mathbf{Q}) \quad \alpha \quad \left|\langle F(\mathbf{Q})\rangle\right|^2 \tag{12}$$

IIe Least Squares Refinement

The standard crystallographic refinement process iteratively improves the agreement between the structure factors calculated from a model structure and those derived from the measured X-ray intensities. This is done by varying parameters in the model based on solutions to the linearized least squares formulation of the problem (Konnert,1976). The function Φ minimized is of the form:

$$\Phi = \sum_{\underset{\sim}{Q}} w(\underset{\sim}{Q})\,\big|\,|F_o(\underset{\sim}{Q})|-|F_c(\underset{\sim}{Q})|\,\big|^2 \qquad (13)$$

where $|F_o|$ is the experimental amplitude of the structure factor and $|F_c|$ is that calculated from the model (Eqn. 7). $w(\underset{\sim}{Q})$ is the weight assigned to the structure factor. As discussed below, Eqn. (13) is generally modified in protein refinements to allow for the introduction of restraints on the structure.

All the refinements reported in this work were done using the restrained-parameter least squares program PROLSQ of Konnert and Hendrickson(1980). Four parameters were refined for every non-hydrogen atom in met-myoglobin: three cartesian coordinates and one isotropic temperature factor. The neglect of hydrogens is not an approximation in this work, as no hydrogens were included in the calculation of structure factors.

The initial model used in all cases was the averaged structure from the molecular dynamics simulation. The average coordinates obtained by sampling at 0.05ps and 0.25ps were identical to within 0.01Å and so just one structure was used as the initial model for all the refinements of the 25ps data. The dynamical average structure obtained by sampling the simulation every 0.25 ps was used as the initial structure for refinement of the 300 ps data. A uniform temperature factor was assigned to all the atoms at the start of the refinements.

IIf The use of restraints in refinement

While refining a protein structure it is usually the case that the number of structure factors experimentally measured is not enough to ensure that the refinement will be well behaved at the desired resolution. This point has been discussed in detail by Konnert and Hendrickson (1980), who include stereochemical data as additional information available to the refinement program. Table 1 shows the ratio of parameters to observables for myoglobin at various minimal inter-planar spacings for various numbers of refinement parameters. From this table it can be seen that for isotropic B-refinement at 1.5Å resolution, the number of independant data points exceeds the number of variable parameters by a factor of 4.3. Thus it should be possible to refine coordinates and isotropic temperature factors for each atom without necessarily resorting to restraints.Though this is shown to be possible in this work, it is much more difficult with experimental data where errors in the measurements often limit the number of reliable data.

To incorporate stereochemical restraints, PROLSQ requires a dictionary of ideal amino-acid structures. The stereochemical restraints include 1-2 distance restraints for the bonds, 1-3 distance restraints for the angles, planarity restraints, torsional restraints and non-bonded contact restraints (Hendrickson, 1980, Hendrickson and Konnert 1980). In addition to these, the variation of the temperature factors of atoms that are bonded to each other or to the same third atom are restrained to lie within specified values (Konnert and Hendrickson, 1980). Finally, the program also restrains the calculated shifts in the parameters. The restraints are added to the observational equation as additional "observations" and so the function that is minimized is:

$$\Phi = \sum_{\underset{\sim}{Q}} w(\underset{\sim}{Q}) \left| |F_o(\underset{\sim}{Q})| - |F_c(\underset{\sim}{Q})| \right|^2 + \sum_d w_d \Delta^2 \tag{14}$$

where $w(\underset{\sim}{Q})$, the weight assigned to the structure factors, varies linearly with $|\underset{\sim}{Q}|$ so that low resolution structure factors are weighted more than high resolution ones (Eqn. 14, Hendrickson,1980):

$$w(\underset{\sim}{Q}) = \alpha - \beta \ (s - \gamma) \tag{15}$$

with $s = \frac{|Q|}{4\pi}$. The value of α controls the weight assigned to the structure factor data relative to the restraint terms as a whole, while β controls the weight assigned to the higher resolution data relative to the lower resolution data. γ is the value of s at which the line defined by Eqn. 15 is pivoted, and a value of $\frac{1}{6}\text{Å}^{-1}$ was used in all the refinements described here (Hendrickson and Konnert, 1980, Hendrickson, 1980). The values of $\alpha$and β are set by trial and error during the course of a refinement (Hendrickson, 1980). Δ is the deviation of a restrained parameter from its ideal value and w_d is the weight assigned to the restraint. The weights on restraints used in this work are listed in Table 2a, and these are similar to the values suggested by Hendrickson (1980) and used in previous refinements of myoglobin (Frauenfelder et al.,1979, Hartmann et al., 1982, Kuriyan et al., 1985a).

The weights can be thought of as the inverse of the expected variance of Δ and their values can be changed to "tighten" or "loosen" any particular stereochemical restraint. In practice these values are used as target values for the observed values of Δ and the overall weights on the restraints are varied so as to make the refined structure conform to these target values (Hendrickson, (1980), Kuriyan et al., (1985a)). This can be done by varying the relative weights of the structure factors and the restraints during the course of the refinement by changing the value of α in Eqn. 14, thus controlling how much the model is forced to adhere to the restraints. One indication of the tightness of the restraints is the deviation of 1-2, 1-3 and 1-4 distances from their ideal values. A 1-2, 1-3 or 1-4 distance for which $|\Delta|$ is more than $2w_d^{-1/2}$ shall be referred to as a "deviant distance".

The ideal values for the stereochemical parameters are derived from crystal structures of small molecules. The large amplitude motions observed in protein crystals, both in simulation results and in refined temperature factors (Karplus and McCammon, 1981,1983, Petsko and Ringe, 1984) might be expected to cause the average structure to exhibit deviant stereochemistry (Karplus, 1981). The bias introduced by restraining parameters is examined by doing refinements with loose restraints, tight restraints and no restraints on the coordinates and temperature factors. Another possible cause for deviations between the stereochemical parameters in the dynamics average structure and the ideal structure is that the parameters used in the molecular dynamics simulation, which determine the equilibrium values of the individual internal coordinates of the molecule, are different from those used in the refinement. This point is discussed further in the Results section.

IIg The Refinements

i) Refinement against $|\langle F\rangle|^{0.25}$ data with loose restraints.

The initial atomic positions used as input to the refinement program were the average positions from the simulation. Only coordinates were refined for the initial model at 2.0Å resolution with a constant overall temperature factor of 2.0 Å^2. The R-factor dropped from 37% to 23.7% in 7 cycles. At this stage the overall temperature factor was set to 13.5 Å^2 , which is the value at which the R-factor is at a minimum, and the refinement was continued at 2.0Å resolution. Merely increasing the temperature factors from 2.0 Å^2 to 13.5Å^2 lowered the R-factor from 23.7% to 19.2%. Then 8 further cycles of coordinate and individual temperature factor refinement lowered the R-factor to 13.0% at which stage the refinement had reached apparent convergence.

Initially, tight restraints were kept on the stereochemistry and the number of deviant distances dropped to 171 from 1093 (see Table 3). For the 1.5Å resolution refinement, where the ratio of observations to variable parameters was greater, the weights on the stereochemical restraints were relaxed and kept very loose (by increasing the value of α in Eqn. 15.) and the number of deviant distances increased at every cycle. In eight cycles the R-factor had dropped to 13.6% with 1372 deviant distances. At this point the R-factor did not change on continuing the refinement, and the coordinates and temperature factors were saved for analysis. The R-factors in various resolution shells for the different refinements are reported in Table 4. Final coordinates and temperature factors from this set will be referred to as the $\langle F\rangle^{0.25}_{\text{restr}}$ set.

ii) Refinement against $|\langle F \rangle|^{0.25}$ data with no restraints.

In this case the weights on all stereochemical and temperature-factor restraints were set to zero. The refinement was started by including all the data between 10.0Å and 1.5Å and the initial model was the dynamics structure with an overall temperature factor of 15.0 Å^2. The initial R-factor was 37.0% at 1.5Å. The number of deviant distances increased to about 2000 in a few steps and then stayed more or less constant throughout the refinement. In fourteen cycles the refinement reached apparent convergence at an R-factor of 13.6%, with 2004 deviant distances (see Table 3). Coordinates and temperature factors from this refinement will be referred to as the $\langle F \rangle^{0.25}_{\text{unrestr}}$ set.

iii) Refinement against $|\langle F \rangle|^{0.05}$ data with loose restraints:

The refinement was started with an overall temperature factor of 15Å^2 and initially only data to 2.0Å were included. 10 cycles of least-squares refinement reduced the R-factor to 10.5% with 1326 deviant distances. At this stage data to 1.5Å were included and 9 more cycles of refinement dropped the R-factor to 13.6%, with 1391 deviant distances. Two further cycles of refinement resulted in no change in the R-factor and the refinement was stopped. Final coordinates and temperature factors from this refinement will be referred to as the $\langle F \rangle^{0.05}_{\text{restr}}$ set.

iv) Refinement against $|\langle F \rangle|^{0.05}$ data with tight restraints:

This refinement was a continuation of the previous one with greater weights on the restraints. The R-factor increased from 13.6% to 16.6%

but the number of deviant distances dropped from 1391 to 157. This structure will be referred to as the $\langle F\rangle^{0.05}_{tight}$ structure.

v) Refinement against $|\langle F\rangle|^{0.05}$ data with no restraints:

Refinement was started at 1.5Å and 14 cycles dropped the R-factor to 12.9% from 37%. Three further cycles of refinement resulted in no change in the R-factor and the final structure, with 1865 deviant distances, will be referred to as the $\langle F\rangle^{0.05}_{unrestr}$ structure.

vi) Refinement against $|\langle F\rangle|^{0.05}$ data with alternate conformations

Difference electron density maps with coefficients $(2F_o-F_c)\exp(i\alpha_c)$ (Blundell and Johnson, 1976) were examined on an Evans and Sutherland PS300 graphics system using the software FRODO (Jones, 1982). The phases, α_c, and the model structure factors, F_c, were calculated from the $\langle F\rangle^{0.05}_{restr}$ structure. F_o is the amplitude of the observed structure factor, i.e. the "experimental" data obtained from the simulation. On the basis of the difference maps alone, ten residues showed clear indications of conformational disorder in their sidechains and these were modelled by two conformations for each of the residues. In all cases only two conformations were built and these differed only from the C_γ atom outwards.

Only one variable occupancy factor was refined for each residue with alternate conformations. All the atoms belonging to one conformations were constrained to have the same occupancy and the occupancies of the two alternate conformations were constrained to add up to 1.0. The refinement was started with equal weights and temperature factors assigned to all atoms with alternate conformations. All other atoms had the same parameters as at the end of the $\langle F\rangle^{0.05}_{restr}$ refinement. Stereochemical restraints were applied to all the atoms and refinement was started at 2.0Å to allow the atomic positions to adjust. 9 cycles of refinement lowered the R-factor from 11.5 to 8.0%, with 1452 deviant distances. Data to 1.5Å were included at this point and a further 8 cycles of refinement lowered the R-factor from 14.1% to 12.6%. The R-factor would not drop on continuing the refinement and the process was stopped. The coordinates and temperature factors from this refinement will be referred to as the $\langle F\rangle^{0.05}_{altconf}$ set.

vii) Refinement against 300 ps data with loose restraints:

The initial model structure was the average structure from the 300 ps trajectory with a uniform temperature factor of 15 Å² assigned to each atom. The first few cycles of refinement included data between 10.0Å and 2.0Å, this was later extended to include all the data between 10.0Å and 1.5Å. 21 cycles of refinement with very loose restraints finally reduced the R-factor to 21.5% at which point the R-factor would not drop further. The refinement was stopped and coordinates and temperature factors saved for analysis; they are referred to as the $\langle F\rangle^{300}$ set.

III RESULTS AND DISCUSSION

IIIa The R-factors of the refined structures

The R-factor (Eqn. 9) is the most commonly used indicator of the quality of a refined model and the final R-factors of the six refinements of the 25ps data are compared in Tables 3 and 4. The R-factors range from 12.6% for the refinement with loose restraints and ten alternate conformations, to 16.6% for the refinement with tight restraints. Refinements with both loose restraints and with no restraints lead to similar R-factors, while tight restraints increase the R-factor by about 3%. On comparing these values with the R-factor of 19.2% obtained using the average dynamics structure (with isotropic atomic temperature factors calculated from the exact mean-square fluctuations from the simulation, using eqns. 10 and 11) we see that the average dynamics structure and exact, isotropic, fluctuations do not yield the best fit to the structure factor data.

The refined R-factors are all higher than the experimental refined R-factors for small molecules, which are usually less than 5% (Dunitz,1979), but they are slightly lower than experimental refined R-factors found for most proteins. A recent X-ray refinement of CO-myoglobin, for example, resulted in R-factors of 16.5% for a structure with loose restraints and several alternate conformations and 18.7% for a structure with tight restraints (Kuriyan et al., 1985a). Comparisons of R-factors should be treated with caution, however, because the R-factor depends on the resolution of the data (usually increasing with higher resolution data) and the ratio of observables to parameters. Decreasing the ratio of observables to variable parameters, either by increasing the complexity of the refinement model or by not including all the unique structure factor data at a particular resolution (usually because of experimental uncertainty), generally results in a decrease in the R-factor (Hirshfeld and Rabinovich, 1973). The CO-myoglobin model (Kuriyan et al.,1985a) was refined against only 10,449 unique structure factors between 10.0Å and 1.5Å with an observations to parameters ratio of 1.87 whereas in the refinements reported here all the 21942 unique structure factors between the same resolution limits were included resulting in an observations to parameters ratio of 4.35. The R-factors from the two refinements are therefore not strictly comparable; the R-factors reported in this work are expected to be higher than those that would be obtained from refinements done on a smaller subset of the same structure factors (Hirshfeld and Rabinovich, 1973).

The high R-factors that are the converged limits of all the refinements appear to be due to the neglect of the anharmonic, anisotropic nature of the atomic fluctuations in the refinement program. That they are not due to any kind of problem with the least-squares algorithm itself is demonstrated both by the high R-factor of the average dynamics structure with exact B-factors from the simulation and also by test refinements carried out on structure factor data generated from structures with isotropic B-factors (Kuriyan, unpublished). Refinement against test data generated from single structures with isotropic and harmonic fluctuations always converge rapidly to a low R-factor (less than 1.0%) and yield refined structures virtually identical to the structures used to generate the data. The refined temperature factors are sometimes in error by a constant amount, but this is easily detected and can be corrected by calculating the R-factor as a function of overall shifts in the B-factor (Kuriyan, unpublished).

The final R-factor of 21.5 % for the refinement of the 300ps data is significantly higher than any of the R-factors for the 25ps data. The mean-square fluctuations of the atoms over the 300ps period are roughly twice as great as the fluctuations over a 25ps period (see Table 5) and the structure seems to undergo some global changes over the longer time period (Levy et al., 1985); thus, the higher R-factor is probably due to the isotropic, harmonic model being even less applicable to the simulation results for the long time-scale. We have not investigated whether the R-factor can be lowered by modelling disordered regions of the protein.

IIIb Restraints on Stereochemistry and Temperature Factors

The bond and angle terms are the most important of the stereochemical restraints. In PROLSQ they are both given as distance restraints, 1-2 bonded distances being restrained for the bonds and 1-3 angle distances being restrained for the angles. To assess the effect of dynamics on the ideality of the bonds and angles, the distances were calculated from the simulation in two different ways. In the first case the distances were calculated from the average dynamics structure, i.e:

$$\langle d_{ij} \rangle = \left| \langle \underset{\sim}{r}_i \rangle - \langle \underset{\sim}{r}_j \rangle \right| \tag{16}$$

and in the second case the distances were calculated from structures sampled every 0.05 ps from the simulation, and then averaged :

$$\langle d_{ij} \rangle = \langle \left| \underset{\sim}{r}_i - \underset{\sim}{r}_j \right| \rangle \tag{17}$$

The average deviations and root mean square deviations of the distances for various classes of atoms from their ideal values (as defined by the PROLSQ dictionary) are shown in Table 2. It is seen that the average and the root mean square of the deviations from ideality for all distances is about a factor of 10 larger for the average structure than for structures sampled from the simulation. The average deviations indicate that the bonds and angles are systematically smaller in the average structure than in the ideal dictionary. The deviations are smallest, in both cases, for the backbone atoms, and largest for atoms more than two atoms along the side chain. Comparing the root mean square (r.m.s.) deviations for sidechain distances with the weights used in refinement, r.m.s. deviations in the average structure are about a factor of 10 higher than the standard deviations implicit in the weights. The deviations from ideality for the two different averages were also calculated explicitly for the bond angles (instead of just the 1-3 distance) and these results are also given in Table 2.

Fig. 1a shows the deviations, for both kinds of averages, as a function of residue number for bonds between sidechain atoms. Except for one or two residues, both averages yield uniformly low deviations (less than about 0.04 Å) for backbone bonds. For sidechain bonds the deviations calculated from averages over the simulation are still uniformly low, but those calculated from the average structure have a much larger deviation, ranging from about 0.1Å to 0.8Å (see Fig 1a). The result for bond angles are very similar, with backbone angles deviating less than 3°-4° for both averages, but with sidechain angles deviating as much as 40°-50° in the average structure.

In all cases the r.m.s. deviations in structures sampled from the simulation are smaller than or equal to the target r.m.s. deviations from ideality (see Table 3) despite the fact that individual structures exhibit fluctuations in the internal coordinates. This indicates that the equilibrium values for bonds, angles, etc. in the CHARMM potential (Brooks et al., 1983) are not significantly different, as far as restraints in the refinement is concerned, from the ideal values in the PROLSQ dictionary. As expected, the regions of the protein with large deviations in geometry in the average structure correlate well with regions of the protein with high mobility in the simulation, (cf. Fig. 5).

It is clear from this analysis that if there are large scale motions ocurring in the protein it may be inappropriate to impose strict stereochemical restraints (Karplus, 1981, Yu et al., 1985). If the average dynamical structure is considered to be the correct structure, then refinement with large weights on the sterochemical restraints would clearly lead to deviations from the average structure. The r.m.s. values of Δ (see eqn. 14) for four of the most important classes of restraints are shown in Table 2 for the average dynamical structure as well as for the various refined structures. For 1-2, 1-3 and 1-4 distance restraints and for planarity restraints the r.m.s. values of Δ are 0.15 Å, 0.20Å, 0.16Å and 0.04Å respectively in the average structure, as compared to 0.09Å, 0.10Å, 0.10Å and 0.05Å in the loosely restrained structures and 0.17Å, 0.20Å, 0.18Å and 0.10Å in the unrestrained structures. For the tightly restrained structure the values of Δ are less than or equal to the target values of 0.03Å, 0.04Å, 0.052Å and 0.025Å.

The major difference between the structures obtained from refinements with loose restraints and no restraints is that very large deviations in geometry are absent in the former structure; for example, the r.m.s. deviation from ideality for sidechain angles is 9.9° in the loosely restrained structure and 20.15° in the unrestrained structure.

Apart from the restraints on stereochemistry, the refinement procedure also imposes restraints on the absolute differences between the temperature factors of atoms that are bonded together (1-2 pairs) or bonded to the same third atom (1-3 pairs) (Konnert and Hendrickson, 1980). Based on an analysis of a 30ps molecular dynamics simulation of bovine pancreatic trypsin inhibitor (BPTI), Yu et al. (1985) support the use of these restraints but indicate that they are about twice as restrictive as they should be.

Let $\Delta=\langle|B_a-B_b|\rangle$ where B_a and B_b are the temperature factors of the two atoms in a 1-2 or 1-3 pair. The average values and standard deviations of Δ for 1-2 and 1-3 pairs between backbone and sidechain atoms are shown in Table 6 for the exact results from the simulation as well as for the refined structures. The variation of temperature factors is lowest for backbone atoms, as expected, and the exact simulation results show a very large difference between backbone atoms and sidechain atoms. For backbone 1-2 pairs Δ is 2.87 Å^2 while for sidechain 1-2 pairs Δ is 8.58 Å^2. The restraint value of Δ is 1.0Å^2 for backbone and sidechain 1-2 pairs and 1.5Å^2 for backbone and sidechain 1-3 pairs (Hendrickson and Konnert, 1980, Hendrickson, 1980). This results in the restrained refinements having values of Δ that are much lower than the exact results (see Table 6). For the unrestrained refinements, the backbone values of Δ are relatively close to the exact results. However, for sidechain pairs the refined values of Δ are only half as much as the exact values even though no restraints were placed on them. As we show below, this is consistent with a general trend that is seen in the values of the fluctuations obtained from the refinements.

IIIc Errors in atomic positions

Root mean square deviations between all the refined structures and the dynamical average structure, as well as deviations between the refined structures themselves, are shown in Table 7. The overall r.m.s. error in atomic positions ranges from 0.24Å to 0.29Å in the various structures. The errors in backbone positions (0.10-0.20Å) are less than for sidechain atoms (0.28-0.33Å r.m.s.). These backbone errors, though small, are comparable to the r.m.s. deviation of 0.21Å between the positions of the backbone atoms in the refined experimental structures of Oxy- and CO- myoglobin (Phillips, 1980, Kuriyan et al., 1985a).

The structures from the two loosely restrained refinements are very similar to each other as are the two structures from the unrestrained refinements. For both backbone and sidechain atoms the unrestrained refinement results in larger positional errors than the restrained refinement, which is surprising because one might have expected that imposing stereochemical restraints would move the atoms further away from the dynamical average than the unrestrained refinement; the origin of this effect is discussed below. Some backbone atoms in the latter refinement are in error by as much as 0.5Å and some sidechain atoms in both refinements are in error by as much as 1.0Å.

The shifts in atomic positions introduced by tightening the restraints in the refinement (0.09Å r.m.s. for the backbone and 0.13Å r.m.s. for the sidechains) is seen to be smaller than the differences between any of the other refined structures and the average dynamical structure, and is comparable to the differences between the two loosely restrained structures. Most importantly, the refinement with tight restraints actually results in the lowest r.m.s. error in backbone positions (0.10Å). The errors in sidechain positions for this structure are comparable to those in the loosely restrained refinements (0.31Å) and slightly less than the errors in the unrestrainted refinements. Refinement with tight restraints results in a structure having, as expected, stereochemical parameters very much closer to their ideal values than for any other refined structure; yet this structure is as good as or better than structures obtained from refinements with loose restraints or no restraints. The correct average structure is not obtained in any of the refinements.

The positional errors are not uniform over the whole structure, These are plotted as a function of residue number for a loosely restrained refinement ($\langle F\rangle^{0.05}_{restr}$) and an unrestrained refinement ($\langle F\rangle^{0.05}_{unrestr}$) in Fig. 1. There is a strong correlation between positional error and the magnitude of the mean square fluctuation for an atom, with certain regions of the protein, such as loops and external sidechains, having greater errors in refined position. We define a local principal axis coordinate system for each atom, which is the coordinate frame in which the fluctuation second moment tensor is diagonal(Willis and Pryor, 1975). The principal X-axis is taken as the axis along which the fluctuations are the largest and the Z-axis is taken as that along which the fluctuations are the smallest. In Fig. 2(a) the dependance of the positional errors on mean-square fluctuation is shown. Fig. 2(b) gives the distribution of positional errors along the principal X-axis for the $\langle F\rangle^{0.05}_{restr}$ structure. The errors in position are largest along the principal X-axis, as expected, as this is along the direction of greatest motion.

The coordinates obtained by refinement against the 300 ps data have errors that are about twice as large as for the refinements against the 25 ps data, with backbone r.m.s. errors of 0.29Å and sidechain r.m.s. errors of 0.56Å. As mentioned above, the errors increase with mean-square fluctuation and the larger errors are consistent with the larger fluctuations in the 300ps simulation (see Table 5).

<u>IIId Errors in refined fluctuations</u>:

The refined mean-square fluctuations are systematically smaller than the fluctuations calculated directly from the simulation for all four refinements. Scatter plots of the fluctuations for all atoms (Fig. 3) show that fluctuations greater than $\approx 0.75\text{Å}^2$ (B = 20Å^2) are almost always underestimated by the refinement. Fluctuations less than 0.75Å^2 are still underestimated more often than they are overestimated in the restrained refinements, though this is less true for the unrestrained refinements. This figure makes very clear the fact that the B-factors (mean-square fluctuations) obtained from the refinement have an effective upper limit independent of the actual values calculated from the dynamics. Fig. 4 shows the distribution of errors in the mean-square fluctuations for restrained and unrestrained refinement and Table 8 gives the correlation coefficients, average absolute errors, average fractional errors and average errors between the four refined sets of temperature factors and the values calculated from the dynamics. The absolute errors averaged over all the atoms range from 5.5 to 7.0 B-factor units (0.21 - 0.27 Å^2) and the average errors range from 4.7 to 7.0 B-factor units (0.18 - 0.27 Å^2). That the average errors are so close to the average absolute errors indicates once again that the refined temperature factors are lower than the values calculated directly from the simulation. The magnitudes and variation of temperature factors along the backbone are very well reproduced by the refinement (Fig. 5 a) but the refined sidechain fluctuations are almost always too low (Fig.5 b,c). Regions of the protein that have high mobility also have large errors in refined position and temperature factor.

The average backbone, sidechain and overall B-factors for the various refined structures are compared with the exact simulation results and with experimental B-factors for various liganded forms of myoglobin in Table 5. In the case of backbone temperature factors, both loosely restrained and unrestrained refinements result in a slight lowering of the B-factors, from 12.4 Å^2 to 11.3 Å^2 and 11.7 Å^2 respectively. The damping of the B-factors is much more marked in the sidechain average, which drops from 26.8Å^2 (exact) to 16.5Å^2 and 17.6Å^2 (restrained and unrestrained, respectively). The effect of tightening the restraints is to decrease the variation in the B-factors (see Table 6). This results in the backbone average being slightly larger (12.5Å^2) than the exact result and the sidechain average being lowered even further than in the other refinements (14.5Å^2). The tight restraints also result in the largest errors in B-factors for any of the refinements, with an average fractional error of 36% and an average absolute error of 7.20 B-factor units (0.27Å^2). This is in contrast to the positional errors, where the tight restraints resulted in the lowest errors.

The systematic underestimation of the B-factors by the refinement increases as the fluctuations increase (see Fig. 3) and so one might expect that the B-factors in the 300ps simulation would be greatly reduced on refinement since the average fluctuations are twice as large over this time-scale than over a 25ps time-scale (Levy et al.,1985).This is indeed seen to be the case and Table 5 gives the average B-factors for the exact simulation results and the refinement. The backbone and sidechain B-factors drop from 25.6Å^2 and 48.6Å^2 to 16.8Å^2 and 21.1Å^2 respectively.

To check for a systematic error in the B-factor (a constant offset), R-factors were calculated as a function of a constant shift applied to the refined B-factors. The result, for refinement of the 300 ps data, is shown in Fig. 6, which shows that there is no constant offset to the B-factors that would improve the R-factor. This was also the case for all the 25ps refinements. The 300ps refinement had been started with an initial B-factor of 15.0Å^2 which is much lower than the exact average B-factor (Table 5). The effect of the initial B-factor on the final values of the refined B-factors was tested by increasing all the B-factors in the refined model by 13.5 B-factor units (0.5Å^2)and continuing the refinement. The final B-factors obtained this way were higher than the previous results, with backbone and sidechain averages of 22.2Å^2 and 26.5Å^2 respectively, an increase of about 3.0Å^2. Once again we checked for a whether a constant B-factor offset would improve the R-factors and the R-factor as a function of B-factor shift is shown in Fig. 6. This time it is seen that a B-factor shift of -3.0Å^2 reduces the R-factor by about 1.5%. This shift is in accord with the original B-factor values. This B-factor shift was applied to all the atomic temperature factors and in Fig. 3(d) a scatter plot of these final mean-square fluctuations against the exact results from the simulation is presented. The underestimation of the fluctuations is an extremely striking feature of this plot.

IIIe The anisotropy and anharmonicity of the atomic fluctuations:

IIIe i) The anisotropy

In this and the following sections some of the characteristics of the atomic probability distribution functions obtained from the simulation will be discussed. The probability distribution function, $p(\underset{\sim}{u})$, gives the probability density of finding the atom displaced $\underset{\sim}{u}$ from its mean position. Probability distribution functions are characterized by their mean, $\underset{\sim}{m}$, and their higher moments; the second moment, σ^2, is the most important one in crystallography since it is related to the temperature factor. The average electron density of an atom is the convolution of its electron density at rest with its probability distribution function. Since the atomic electron density can be considered to be time-independent, the probability distribution function can be used instead of the average electron density (Willis and Pryor, 1975). We shall use σ^2_{md} to mean the exact mean-square fluctuation calculated from the simulation and σ^2_{ref} to mean that obtained from refinement.

The anisotropy and anharmonicity of the distribution functions in molecular dynamics simulations of proteins have been studied in detail previously (Mao et. al.,1982; Ichiye and Karplus, 1985a,b). Rather than carrying out an extensive analysis of the probability distributions in the simulation used here, it will be shown that the magnitudes of the anisotropy and anharmonicity in myoglobin are very similar to those found in the 30ps simulation of lysozyme by Ichiye and Karplus (1985 a,b) and the approach of these authors will be used in the following analysis. Let U_x, U_y and U_z be the fluctuations from the mean position along the principal X, Y and Z axes and

$$\sigma_x^2 = \langle U_x^2 \rangle , \sigma_y^2 = \langle U_y^2 \rangle \text{ and } \sigma_z^2 = \langle U_z^2 \rangle \qquad (15)$$

with $\sigma_x \geq \sigma_y \geq \sigma_z$. We define one measure of the anisotropy by:

$$A_1 = \left[\frac{\sigma_x^2}{\frac{1}{2}(\sigma_y^2+\sigma_z^2)} \right]^{\frac{1}{2}} - 1.0 \qquad (16)$$

This measures the amount by which the ratio of the fluctuation in the principal X-direction to the average of that in the other two directions exceeds that of an isotropic distribution, for which A_1 is zero. We also define another measure of the anisotropy:

$$A_2 = \left[\frac{\sigma_y^2}{\frac{1}{2}(\sigma_y^2+\sigma_z^2)} \right]^{\frac{1}{2}} - 1.0 \qquad (17)$$

which measures how isotropic the motion is in the principal Y-Z plane. A_1 and A_2 have been calculated for various classes of atoms and the values are tabulated in Table 9, which also includes the results of Ichiye and Karplus (1985a) for the same classes of atoms in lysozyme. While the anisotropy defined either way is seen to be slightly lower in myoglobin, the general trends are the same in both molecules. The motions tend to be quite anisotropic; very few atoms (about 1.4 %) have A_1 less than 0.02. 61% of the atoms have A_1 greater than 0.5, and 31% have A_1 greater than 0.75. Atoms further out along the sidechain have higher values of A_1, but the value of A_2 remains uniformly low for all classes of atoms (at about 0.15). This indicates that the most significant contribution to the anisotropy is along the direction of largest motion and that the motion is rather more isotropic along the principal Y-Z plane.

Ichiye and Karplus (1985b) have studied the errors introduced by refining an isotropic gaussian model for the distribution function against both analytic probability distributions as well as distributions obtained from the simulation. The procedure they use is analogous to real space refinement of the electron density of an isolated atom; instead of refining a model for the electron density, they refine models for the distribution function. The function they minimize is of the form :

$$R = \int (p_c(\underset{\sim}{u}) - p_o(\underset{\sim}{u}))^2 d\underset{\sim}{u} \qquad (18)$$

where p_o and p_c are the actual and model distributions, respectively. Diamond (1971) has shown that this kind of refinement is equivalent to reciprocal space refinement where all the structure factors are weighted equally. In the refinement procedures we used in this work the structure factors are not weighted equally (Eqn. 14). Nevertheless, it is helpful to use the results for lysozyme to aid in the analysis of the reciprocal space refinements reported here.

To examine the effect of anisotropy separately from that of anharmonicity, Ichiye and Karplus (1985b) studied the case where p_o is a three-dimensional anisotropic gaussian and p_c is an isotropic gaussian. They found that the refined values of σ are close to the actual values only for small values of the anisotropy ($\sigma_x/\sigma_y<1.5$). For larger anisotropies the refined values of σ are always lower than the actual value. This suggests that there should be a correlation between the values of the anisotropy in the myoglobin simulation with the errors in the refined values of the temperature factor. Fig. 7 shows the dependence of the position and fluctuation errors on anisotropy for one of the refinements with loose restraints. The errors in position and temperature factor increase with anisotropy, but the effects of anisotropy cannot be seperated from those of anharmonicity. Also, if the pdf were an anisotropic gaussian there would be no predicted error in the refined position; however, the error in position also increases with anisotropy.

IIIe ii) The anharmonicity.

The third and fourth moments of the distribution can be used to characterize the anharmonicity (Mao et. al.,1982; Ichiye and Karplus, 1985a). The skewness, α_{3i}, where i is x,y or z, is defined by:

$$\alpha_{3i} = \frac{\langle U_i^3 \rangle}{\langle U_i^2 \rangle^{3/2}} \qquad (19)$$

and the coefficient of excess kurtosis, α_{4i} , is given by:

$$\alpha_{4i} = \frac{\langle U_i^4 \rangle}{\langle U_i^2 \rangle^2} - 3.0 \qquad (20)$$

Both α_3 and α_4 are zero for a gaussian distribution. The average values of $|\alpha_{3i}|$ and $|\alpha_{4i}|$ for various classes of atoms have been calculated and compared with the values obtained for lysozyme (Ichiye and Karplus, 1985a). The values for the two proteins are strikingly similar (Table 10). From a detailed study of the moments of the atomic distributions Ichiye and Karplus conclude that most atoms with large anharmonicity have multiple peaks in their distribution functions, with each peak being close to harmonic. They suggest that the best description of anharmonicity for atoms with large fluctuations should be not based on pertubations to a local gaussian distribution, but rather should include contributions from separated gaussian distributions.

IIIf Probability distribution functions from dynamics and refinement

It is of interest to determine whether the errors in the refinement are localized to just a few residues of the protein and also to determine what causes the systematic underestimation of the fluctuations. In Table 11 all the residues that have at least one atom with a refined temperature factor 50% lower than the exact value are listed. The temperature factors used are from the refinement of the $\langle F\rangle^{0.05}$ data with loose restraints. There are 77 such atoms distributed over 45 residues, and Table 11 also includes the secondary structure elements (helices or loops) and average solvent accessible area for these residues.

Most of the residues are in the helix regions, with only seven in the inter-helix loops. Among the helices, the B and F helices have the lowest number of such residues, with two each, and the A helix has the most, with ten. Most of the residues have charged sidechains and are on the surface, but eleven of them are partially or completely buried (with average sidechain solvent accessibility less than 3.0Å^2). These include a tryptophan, a valine, four leucines, an isoleucine, a glycine, a histidine, an arginine and an aspartic acid.

The probability distribution functions from the 25 ps simulation of met-myoglobin for about 30 such atoms have been studied in the following way. For each atom the second moment tensor was calculated and diagonalized to obtain the transformation to the local pricipal axis frame. The time-series for the 3 principal axis coordinates were calculated from the simulation and, from the time-series, the probability of fluctuations along the three principal axes was estimated by dividing the coordinate ranges into 25 bins and counting the number of times the trajectory was in each bin. The resulting distribution was normalized to have unit total probability. For each atom the gaussian distributions corresponding to the simulation average position and σ^2_{md} and the refined position and σ^2_{ref} were also calculated. For every such atom studied the simulation had two or more well separated regions of high probability and the refinement had fit only one of the regions, neglecting the rest. This explains both the lower refined fluctuation and the error in the refined position, as the refinement moves the atom from the true average position into one of the regions of high probability.

Fig. 8(a) and Fig. 8(b) show the molecular dynamics distribution function along the principal X-axis as well as the equivalent gaussian distribution and the refined (restrained and unrestrained) gaussian distributions for the $C_{\varepsilon 1}$ atom in Histidine 81. The exact distribution function has two major peaks and both restrained and unrestrained refinements fit only one of the two peaks. The positional error in the unrestrained refinement is larger because it has moved even further towards fitting just one peak. Two more examples of this kind of distribution are shown in Fig. 8. They are for the $C_{\delta 1}$ atom of Leucine 69 (Fig. 8(d)) and for the $C_{\delta 2}$ atom of Leucine 11 (Fig. 8(e)). The former has two well separated peaks and a long tail in the distribution and the refinement fits only the major peak. The distribution for Leucine 11 is interesting because it has three peaks and the refinement fits two of them but not the third.

These results are consistent with the findings of Ichiye and

Karplus (1985b). On refining molecular dynamics distributions with isotropic gaussians they found that the refinement would usually fit the major peak of a multi-peaked distribution and neglect the rest. They also studied the refinement of a double peaked gaussian distribution (two identical isotropic gaussians separated along the principal X-axis) by a single isotropic gaussian. The solution to this problem is obtained analytically (Ichiye and Karplus,1985b). For small values of the separation, δ, between the two peaks, the refinement will fit both gaussians. However, when $\delta > 3.74\sigma_0$ where σ_0^2 is the second moment of one of the two gaussians in the double peaked distribution, the refinement will fit only one of the gaussians with a refined σ only slightly larger than σ_0.

The distribution functions for atoms which have low fluctuations in the simulation but have high refined temperature factors have also been examined. Such atoms are usually close to atoms which have large mean-square fluctuations and multiple peaks in the distribution and the larger refined fluctuation was seen to be due to the refinement moving these atoms towards the extra density of the disordered atoms. This feature was most marked in the unrestrained refinement and the C_γ atom of Leucine 11 is an example. The terminal atoms of this residue, $C_{\delta 1}$ and $C_{\delta 2}$, are disordered (see above), and the exact temperature factors for the C_γ, $C_{\delta 1}$ and $C_{\delta 2}$ atoms are $19.46Å^2$, $63.5Å^2$ and $71.4Å^2$. The fluctuations obtained from unrestrained refinement are $31.9Å^2$, $26.9Å^2$, and $24.3Å^2$ respectively, with large errors in the refined position of all three atoms and with the temperature factor of the C_γ atom significantly overestimated. Examination of the distribution functions for this atom along with the distribution functions for the terminal atoms clearly showed that C_γ atom moves to fit the extra density due to the disordered terminal atoms.

Refinement with restraints prevents atoms from moving into the density of neighboring atoms, as this would lead to violations of stereochemistry. The result is a lowering of the errors in well defined atoms which are close to disordered atoms and this explains the fact that refinements with restraints yield structures closer to the true average.

The large separation between the peaks in some of the distributions examined immediately suggested that several residues probably needed to be modelled by including alternate conformations. The exact distributions were not used to decide which residues to model in this way because this information is never accessible in an experimental situation. Instead, difference electron density maps were calculated. As described in the Methods section, a $(2F_o-F_c)\exp(i\alpha_c)$ synthesis was used, where the phases, α_c, and the model structure factors, F_c, were calculated from the $\langle F\rangle^{0.05}_{restr}$ structure. F_o is the amplitude of the structure factor calculated from the simulation.

For ten residues the difference map clearly showed the existence of alternate sidechain conformations which were not accounted for in the refined model. These alternate conformations were modelled by changing the sidechain torsion angles and fitting the extra electron density. For other residues the situation was not so clear and building in alternate conformations would have required some judgement. This is partly because the errors in the refined model used to phase the data affect the quality of the difference map (Blundell and Johnson, 1976) and also because the limited resolution of the data (1.5Å) makes it difficult to identify the conformations that are separated by 0.5Å-1.5Å or less. Another difficulty in modelling the disorder is that simple sidechain torsional isomerization may be an inadequate model of the complicated dynamics actually taking place. Some examples are given below of residues which sample multiple conformations without undergoing torsional transitions.

The refinement of a model including alternate conformations for ten sidechains (these are listed in Table 11) is described in the Methods section. The refinement works very well for residues which had very large errors in the previous refinement. Fig. 6(c) shows the distribution functions from molecular dynamics and from the alternate conformations refinement for Histidine 81 $C_{\delta 1}$. The double peaked gaussian describes the actual molecular dynamics distribution rather well.If the average position and temperature factor for the atom are calculated from its two refined positions and occupancies, we see that the error in refined position is only 0.18Å and the error in refined mean-square fluctuation is only 0.25$Å^2$. Refinement with only one conformation led to errors of 2.0Å in position and 5.7$Å^2$ in mean-square fluctuation.

However, the overall agreement between the refined fluctuations and the dynamics fluctuations is not greatly improved by including alternate conformations for just ten residues. Fig 3(c) shows a scatter plot of all the fluctuations for the dynamics and the refinement with alternate conformations and the effect of this refinement is seen to be an improvement in the agreement only for atoms which previously had extremely high errors. The same is true of the positional errors and Table 6 includes the correlations and errors for this refinement, which are similar to those for all the other refinements. Fig 5(d) shows the fluctuations as a function of residue number for this refinement. It appears that there might be a few sidechains for which alternate conformations might still be built, but for the reasons mentioned above it is difficult to do much more with a 1.5Å resolution difference map.

Refining alternate conformations for 10 residues has not lowered the R-factor very much. The final R-factor for restrained refinement is 13.6% and for restrained refinement with alternate conformations it is 12.6%, indicating that the major inadequacies in the refinement model have not been removed by refining alternate conformations for just a few residues.

IIIg An examination of structural transitions

The well separated regions of density seen for residues with large errors imply the existence of structural transitions from one local potential minimum to another. An exhaustive analysis would require the calculation and examination of the positional time-series of all the atoms in question, which has not been attempted. Here two preliminary analyses are presented, one of dihedral transitions and another of larger scale helix deformations.

Dihedral transitions were monitored by following trajectories from one minimum in the torsional potential to another for all the dihedral angles in the protein. This analysis was done using the program CHARMM (Brooks et al.,1983). A transition is defined as a change in the dihedral angle from one well of the torsional potential to another, the wells in the potential being defined by the periodicity of the energy function for that torsion (Brooks et al.,1983). A transition is counted as such only if it involves crossing at least 30° beyond the maximum of the barrier. Torsional angles which underwent transitions in the 45 residues with large error are listed in Table 11.

The time-series for these torsion angles have not been examined to see if the transitions are merely transient jumps to another well or if they actually represent significant population of two or more conformations. Nevertheless they indicate the kinds of motion that might lead to disorder in the sidechain or backbone. Of the sidechain torsions, transitions in χ_1 or χ_2 lead to the largest shifts in sidechain position, and the ten residues that were modelled by alternate conformations in the refinement all have transitions in at least one of these dihedrals. Many residues also have transitions in the backbone ϕ and φ dihedrals and, for some, this results in the carbonyl oxygen of the backbone being disordered. Fig. 8(f) shows the distribution functions for the carbonyl oxygen of Aspartate 141, and once again it is seen that the refinement fits only the major peak of a multi-peaked distribution.

Four of the residues in Table 11 actually have no transitions in torsional angles, either for the sidechain or the backbone. However, they too have multi-peaked distributions; Fig. 8(g,h) show the distributions for two such residues, Histidine 12, in the A-helix and Threonine 51, in the D-helix. The dynamics of these helices was examined on a PS300 graphics system using the molecular graphics program HYDRA (Hubbard, 1985) and it was clear that deformations of the helices as a whole were leading to the large motions of these sidechains.

Examination of the trajectory on the graphics system indicated that the A-helix as a whole was twisting about the helix axis during the trajectory, leading to large fluctuations in the positions of many sidechains. Some of these sidechains undergo torsional transitions as well (eg. Leucine 11, see above), leading to very complicated overall dynamics. The dihedrals of Histidine 12, however, fluctuate very little during the trajectory and most of the disorder is due to the sidechain following the twist of the helix backbone.

Using the dynamical average structure a rigid body principal axis frame is defined for the A-helix backbone. This is the coordinate frame in which the moment of inertia tensor for the backbone atoms is diagonal. In this coordinate system the Z-coordinate of an atom is its position along the helix axis and the X and Y coordinates specify its distance from the axis and its rotation about the helix axis. A time-series for the rotation of the C_α atom of HIS 12 about the helix axis was calculated by rotating and translating every coordinate set from the simulation into the helix principal axis frame defined above and defining the twist angle as $\tan^{-1}\frac{y}{x}$. The twist about the helix axis is plotted as a function of time in Fig. 9 along with the time-series for the positional fluctuation of the $C_{\varepsilon 1}$ atom of the imidazole ring. It is seen that the motion of the ring atom follows the twist of the backbone about the helix axis.

CONCLUSIONS

Protein refinement procedures have been tested by means of data generated by molecular dynamics simulations. It has been shown that the use of single site isotropic models for atoms in refinement methods for proteins leads to errors in the determination of their temperature factors and average positions. These errors are smallest for atoms with low mobility, but can be very serious for atoms with temperature factors exceeding about 20Å^2 (mean-square fluctuations exceeding about 0.75Å^2). The magnitudes of the anharmonicity and anisotropy in myoglobin are very similar to those found for lysozyme (Ichiye and Karplus,1985 a,b), suggesting that these results may be of general significance.

The neglect of solvent in the simulation is an approximation that is likely to alter the details of the dynamics, especially for the surface residues. Even though this may affect the magnitudes of the fluctuations, it should not affect the conclusion of this paper that distribution functions with multiple peaks are the most important cause of anharmonicity and anisotropy in proteins and that this leads to temperature factors being underestimated and to the refined structure being differant from the dynamical averaged structure.

The multiple peaks in the distribution function are well seperated in some residues and alternate conformations can be picked out by difference Fourier techniques and these can be included in the refinement model. For most atoms, however, such conformations are difficult to resolve at 1.5Å resolution. Modelling disorder is further complicated by the fact that changes in sidechain torsional angles alone might not be enough to account for the changes in the conformation of a residue.

The use of stereochemical restraints in the refinements leads to better agreement with the average dynamical structure than the use of no restraints. The inadequacy of the single site model causes unrestrained atoms to move away from their true positions if they are close to disordered regions of the protein. The positional errors obtained on applying tight restraints on stereochemistry are not larger than those obtained using loose restraints, while the stereochemical parameters are greatly improved. Thus, refinement even at 1.5Å resolution benefits from the use of such restraints. The restraints on temperature factor differences, however, are too restrictive and result in a significant damping of the sidechain temperature factors.

Refinements of data generated from the simulation always converged to R-factors greater than 12.0% regardless of whether restraints were used or not. Higher resolution refinements would need to be done in order to determine the most appropriate models to be used for protein refinements.

Acknowledgements

I would like to thank Toshiko Ichiye for many stimulating discussions about protein refinement and Rod Hubbard for help with the display program HYDRA. The molecular dynamics simulations of myoglobin used were done by Ron Levy (Levy et al., 1985). This work has been accepted for publication (J. Kuriyan, G.A. Petsko, R.M. Levy and M. Karplus (1986) J. Mol. Biol., in press).

APPENDIX

To generate time averaged structure factors from a molecular dynamics trajectory, the central assumption is that the simulation of the dynamics of an isolated protein molecule yields an adequate description of the probability of atomic fluctuations in a crystal. Given no information about the crystal lattice dynamics, one of two assumptions can be made about the model crystal from which the structure factors will be derived.

The first assumption, which leads to Model 1, is that at every instant the crystal contains a large number of identical unit cells, each of which evolves identically in time according to the molecular dynamics trajectory. An alternative assumption is that while the molecular dynamics simulation describes the time evolution of any one unit cell, at any instant the crystal would have unit cells which, rather than being identical, represented different configurations along the molecular dynamics trajectory; in other words, it is assumed that the unit cells are all uncorrelated with each other.

The consequences of these assumptions can be analysed in terms of simple diffraction theory. The diffracted intensity, $I(\underset{\sim}{Q})$, at a point $\underset{\sim}{Q}$ in reciprocal space (see Eqn. 1 in the main text), is proportional to the square of the Fourier transform of the electron density in the crystal:

$$I(\underset{\sim}{Q}) \quad = \quad \text{const.}\left|\int d\underset{\sim}{X}\ P(\underset{\sim}{X})\ e^{i\underset{\sim}{Q}\cdot\underset{\sim}{X}}\right|^2 \qquad \text{(A1)}$$

where $P(\underset{\sim}{X})$ is the electron density at position $\underset{\sim}{X}$, the integral being over the whole crystal. A more convenient form is obtained by writing the integral as a sum over unit cells. Let $\rho_\alpha(\underset{\sim}{r})$ be the electron density within unit cell α . Then, ignoring the constant term in what follows,

$$I(\underset{\sim}{Q}) \quad = \quad \left| \sum_{\alpha=1}^{N} \int d\underset{\sim}{r}\ \rho_\alpha(r) e^{i\underset{\sim}{Q}.(\underset{\sim}{r}+\underset{\sim\alpha}{r})} \right|^2 \qquad (A2)$$

where $\underset{\sim}{X} \; = \; \underset{\sim}{r} + \underset{\sim\alpha}{r}$ and the integral is now only over a single unit cell and is summed over the N unit cells in the crystal.

If there is motion in the crystal then the density at any point is time dependent. Including the time, t, explicitly, and expanding the square, gives:

$$I(\underset{\sim}{Q},t) \quad = \quad \sum_{\alpha=1}^{N} \sum_{\alpha'=1}^{N} e^{i\underset{\sim}{Q}.(\underset{\sim\alpha}{r}-\underset{\sim\alpha'}{r})} \left[\int d\underset{\sim}{r} \int d\underset{\sim}{r}' \rho_\alpha(\underset{\sim}{r},t) \rho_{\alpha'}(\underset{\sim}{r}',t) e^{i\underset{\sim}{Q}.(\underset{\sim}{r}-\underset{\sim}{r}')} \right] (A3)$$

The time averaged intensity, which is the quantity of interest, is:

$$\langle I(\underset{\sim}{Q})\rangle \quad = \quad \sum_{\alpha=1}^{N} \sum_{\alpha'=1}^{N} e^{i\underset{\sim}{Q}.(\underset{\sim\alpha}{r}-\underset{\sim}{r}'_{\alpha'})} \int d\underset{\sim}{r} \int d\underset{\sim}{r}' \langle \rho_\alpha(\underset{\sim}{r},t) \rho_{\alpha'}(\underset{\sim}{r}',t)\rangle e^{i\underset{\sim}{Q}.(\underset{\sim}{r}-\underset{\sim}{r}'} (A4)$$

The two models will now be considered separately:

<u>Model 1</u>

Since the unit cells are identical at every instant:

$$\rho_\alpha(\underset{\sim}{r},t) \quad = \quad \rho_{\alpha'}(\underset{\sim}{r},t) \quad \text{for all } \alpha,\ \alpha' \text{ and } t \qquad (A5)$$

Therefore,

$$\langle I(\underset{\sim}{Q})\rangle \quad = \quad \sum_{\alpha=1}^{N} \sum_{\alpha'=1}^{N} e^{i\underset{\sim}{Q}.(\underset{\sim}{r}_\alpha-\underset{\sim}{r}_{\alpha'})} \int d\underset{\sim}{r} \int d\underset{\sim}{r}' \langle \rho(\underset{\sim}{r},t) \rho(\underset{\sim}{r}',t)\rangle e^{i\underset{\sim}{Q}.(\underset{\sim}{r}-\underset{\sim}{r}')} \quad (A6)$$

or,

$$=\rangle\langle I(\underset{\sim}{Q})\rangle \quad = \quad \langle \left| \sum_{\alpha=1}^{N} e^{i\underset{\sim}{Q}.\underset{\sim}{r}_\alpha} \int d\underset{\sim}{r} \rho(\underset{\sim}{r}) e^{i\underset{\sim}{Q}.\underset{\sim}{r}} \right|^2 \rangle \qquad (A7)$$

and

$$\Rightarrow \langle I(\mathbf{Q})\rangle = 2\pi^3 N \sum_{\mathbf{H}} \delta(\mathbf{Q}-2\pi\mathbf{H})\langle \left|F(\mathbf{Q},t)\right|^2\rangle \tag{A8}$$

where $F(\mathbf{Q},t)$ is the structure factor, i.e. the Fourier transform of the contents of one of the N identical unit cells, at time t. The steps in going from Eqn. A7 to A8 are explained in the section on Model 2. Thus model 1 predicts non-zero intensity only at the reciprocal lattice points , $\mathbf{H}$, and:

$$\langle I(\mathbf{Q})\rangle \propto \langle \left|F(\mathbf{Q},t)\right|^2\rangle \quad : \text{at } \mathbf{Q} = 2\pi\mathbf{H} \tag{A9}$$

This quantity can be evaluated readily from the molecular dynamics trajectory by Fourier transforming the electron density at each time step and averaging the squares of the calculated structure factors.

<u>Model 2</u>

In this case no two unit cells are identical, but the fact that different unit cells are uncorrelated can be used to simplify the expression. The instantaneous density can be written as:

$$\rho_\alpha(\mathbf{r},t) = \langle \rho_\alpha(\mathbf{r})\rangle + \Delta\rho_\alpha(\mathbf{r},t) \tag{A10}$$

where $\langle \rho_\alpha(\mathbf{r})\rangle$ is the time averaged electron density at $\mathbf{r}$ and $\Delta\rho_\alpha(\mathbf{r},t)$ is the instantaneous fluctuation in the density. The time-averaged density is constant from unit cell to unit cell assuming a homogeneous crystal, and the α dependence in the first term can be dropped. Then,

$$\begin{aligned} I(\mathbf{Q}) = {} & \sum_{\alpha=1}^{N}\sum_{\alpha'=1}^{N} e^{i\mathbf{Q}\cdot(\mathbf{r}_\alpha-\mathbf{r}_{\alpha'})}\int d\mathbf{r}\int d\mathbf{r}'\langle \rho(\mathbf{r})\rangle\langle \rho(\mathbf{r}')\rangle e^{i\mathbf{Q}\cdot(\mathbf{r}-\mathbf{r}')} \\ & + \sum_{\alpha=1}^{N}\sum_{\alpha'=1}^{N} e^{i\mathbf{Q}\cdot(\mathbf{r}_\alpha-\mathbf{r}_{\alpha'})}\int d\mathbf{r}\int d\mathbf{r}'\langle \Delta\rho_\alpha(\mathbf{r},t)\Delta\rho_{\alpha'}(\mathbf{r}',t)\rangle e^{i\mathbf{Q}\cdot(\mathbf{r}-\mathbf{r}')} \end{aligned} \tag{A11}$$

Naming the first term I_1 and the second I_2 , it can now be shown that for a homogeneous crystal I_1 is $O(N^2)$ while I_2 is $O(N)$.

We have :

$$I_1 \;=\; \sum_{\alpha=1}^{N}\sum_{\alpha'=1}^{N} e^{i\underset{\sim}{Q}.(\underset{\sim}{r}_\alpha-\underset{\sim}{r}_{\alpha'})}\int d\underset{\sim}{r}\int d\underset{\sim}{r}'\langle\rho(\underset{\sim}{r})\rangle\langle\rho(\underset{\sim}{r}')\rangle e^{(i\underset{\sim}{Q}.(\underset{\sim}{r}-\underset{\sim}{r}'))} \qquad (A12)$$

Or :

$$I_1 \;=\; \sum_{\alpha=1}^{N}\sum_{\alpha'=1}^{N} e^{i\underset{\sim}{Q}.(\underset{\sim}{r}_\alpha-\underset{\sim}{r}_{\alpha'})}\left|\langle F(\underset{\sim}{Q})\rangle\right|^2 \qquad (A13)$$

where $\langle F(\underset{\sim}{Q})\rangle$ is the Fourier transform of the time-averaged electron density in a unit cell. If $\underset{\sim}{a},\underset{\sim}{b}$, and $\underset{\sim}{c}$ are the direct unit cell vectors and $\underset{\sim}{a}^*,\underset{\sim}{b}^*$, and $\underset{\sim}{c}^*$ are the reciprocal unit cell vectors, then $\underset{\sim}{Q} = 2\pi(h\underset{\sim}{a}^* + k\underset{\sim}{b}^* + l\underset{\sim}{c}^*)$ (h,k and l are not, in general, integral; see Eqn.1 in the main text) and $\underset{\sim}{r}_\alpha = \beta\underset{\sim}{a} + \gamma\underset{\sim}{b} + \delta\underset{\sim}{c}$ where β,γ and δ are integers. Then, Eqn. A13 can be written as:

$$I_1(\underset{\sim}{Q}) \;=\; \left|\langle F(\underset{\sim}{Q})\rangle\right|^2 \left|\sum_{\beta=1}^{N_x} e^{2\pi ih\beta}\right|^2 \left|\sum_{\gamma=1}^{N_y} e^{2\pi ik\gamma}\right|^2 \left|\sum_{\delta=1}^{N_z} e^{2\pi il\delta}\right|^2 \qquad (A14)$$

where N_x, N_y, N_z are the numbers of unit cells along the three cell edges and $N_x.N_y.N_z = N$. By expanding the exponentials and simplifying, this leads to:

$$I_1(\underset{\sim}{Q}) \;=\; \frac{\sin^2(\pi hN_x)}{\sin^2(\pi h)}\;\frac{\sin^2(\pi kN_y)}{\sin^2(\pi k)}\;\frac{\sin^2(\pi lN_z)}{\sin^2(\pi l)}\left|\langle F(\underset{\sim}{Q})\rangle\right|^2 \qquad (A15)$$

The function

$$\frac{\sin^2(\pi hN_x)}{\sin^2(\pi h)}\;\frac{\sin^2(\pi kN_y)}{\sin^2(\pi k)}\;\frac{\sin^2(\pi lN_z)}{\sin^2(\pi l)}$$

is periodic in h, k, l and is called the Laue Interference Function. It is very sharply peaked around:

$$h = \text{integer},\ k = \text{integer},\ l = \text{integer} \tag{A16}$$

These are the Bragg conditions which define the reciprocal lattice points (h,k,l) and so the first conclusion is that I_1 corresponds to Bragg scattering in that it proportional to the square of the Fourier transform of the time-averaged electron density and also in that it is sharply peaked at the reciprocal lattice points. The Laue Interference Function must be evaluated in order to compare the relative magnitudes of I_1 and I_2 . If the Laue Interference Function is evaluated exactly at a reciprocal lattice point, its value is N^2. This is because

$$\lim_{h \to \text{integer}} \frac{\sin^2(\pi h N_x)}{\sin^2(\pi h)} = N_x^2 \tag{A17}$$

It is more common to write I_1 in terms of its behaviour under integration. This may be done by looking at its behaviour near one diffraction spot (h = 0, k = 0, l = 0) and recognizing that the Laue Interference Function is periodic. If we integrate I_1 over a volume small compared to the reciprocal lattice cell volume (i.e. integrate near a lattice point ,presumably as is done by a diffractometer) and treat $|F(\underset{\sim}{Q})|^2$ as constant, we get :

$$\int_{\underset{\sim}{Q}-\Delta\underset{\sim}{Q}}^{\underset{\sim}{Q}+\Delta\underset{\sim}{Q}} d\underset{\sim}{Q}\ I_1(\underset{\sim}{Q}) = \left|\langle F(\underset{\sim}{Q})\rangle\right|^2$$

$$\int_{-\Delta h}^{+\Delta h} \frac{\sin^2(\pi h N_x)}{\sin^2(\pi h)} 2\pi dh \int_{-\Delta k}^{+\Delta k} \frac{\sin^2(\pi k N_y)}{\sin^2(\pi k)} 2\pi dk \int_{-\Delta l}^{+\Delta l} \frac{\sin^2(\pi l N_z)}{\sin^2(\pi l)} 2\pi dl \tag{A18}$$

For large N each of the $\sin^2$ terms rapidly goes to zero as h, k, l deviate from zero, i.e., the terms go to zero for

$$|h| > \frac{1}{N_x},\quad |k| > \frac{1}{N_y} \text{ and } |l| > \frac{1}{N_z}.$$

So, in the limit of large N_x,

$$\int_{-\Delta h}^{+\Delta h} \frac{\sin^2(\pi h N_X)}{\sin^2(\pi h)} 2\pi dh \quad = \quad \int_{-\infty}^{+\infty} \frac{\sin^2(\pi h N_X)}{(\pi h)^2} 2\pi dh \quad = \quad 2\pi N_X \qquad \text{(A19)}$$

and similarly for the integrals over k and l. Thus, Eqn. A18 reduces to:

$$\int_{\mathbf{Q}-\Delta\mathbf{Q}}^{\mathbf{Q}+\Delta\mathbf{Q}} d\mathbf{Q}\ I_1(\mathbf{Q}) \quad = \quad (2\pi)^3 N \left| \langle F(\mathbf{Q}) \rangle \right|^2 \qquad \text{(A20)}$$

Eqn. A20 implies that $I_1(\mathbf{Q})$ can be written as:

$$I_1(\mathbf{Q}) \quad = \quad (2\pi)^3 N \sum_{\mathbf{H}} \delta(\mathbf{Q}-2\pi\mathbf{H}) \left| \langle F(\mathbf{Q}) \rangle \right|^2 \qquad \text{(A21)}$$

where $\delta(\mathbf{Q}-2\pi\mathbf{H})$ is the Dirac delta function and is non-zero only at the reciprocal lattive points.

To evaluate I_2 the fact that the dynamics of atoms in differant unit cells are uncorrelated is used to simplify the expression in the following way :

$$\begin{aligned} I_2(\mathbf{Q}) \quad &= \sum_{\alpha=1}^{N} \sum_{\alpha'=1}^{N} e^{i\mathbf{Q}\cdot(\mathbf{r}_\alpha-\mathbf{r}_{\alpha'})} \int d\mathbf{r} \int d\mathbf{r}' \langle \Delta\rho_\alpha(\mathbf{r},t)\Delta(\rho(\mathbf{r}',t)\rangle e^{i\mathbf{Q}\cdot(\mathbf{r}-\mathbf{r}')} \qquad \text{(A22)} \\ &= \sum_{\alpha=1}^{N} \int d\mathbf{r} \int d\mathbf{r}' \langle \Delta\rho_\alpha(\mathbf{r},t)\Delta\rho_\alpha(\mathbf{r}',t)\rangle e^{i\mathbf{Q}\cdot(\mathbf{r}-\mathbf{r}')} \\ &\quad + \sum_{\alpha=1}^{N} \sum_{\alpha'\neq\alpha}^{N} e^{i\mathbf{Q}\cdot(\mathbf{r}_\alpha-\mathbf{r}_{\alpha'})} \int d\mathbf{r} \int d\mathbf{r}' \langle \Delta\rho_\alpha(\mathbf{r},t)\Delta\rho_{\alpha'}(\mathbf{r}',t)\rangle e^{i\mathbf{Q}\cdot(\mathbf{r}-\mathbf{r}')} \end{aligned}$$

When different unit cells are uncorrelated, the term $\langle \Delta\rho_\alpha(\mathbf{r},t)\Delta\rho_{\alpha'}(\mathbf{r}',t)\rangle$ is zero for $\alpha\neq\alpha'$. Hence

$$I_2(\mathbf{Q}) \quad = \quad N\left\langle \left| \Delta F(\mathbf{Q}) \right|^2 \right\rangle \qquad \text{(A23)}$$

where $\Delta F(\mathbf{Q}) \quad = \quad \int d\mathbf{r}'\Delta\rho(\mathbf{r})e^{i\mathbf{Q}\cdot\mathbf{r}}$ So now we have :

$$I(\mathbf{Q}) \quad = \quad (2\pi)^3 N \sum_{\mathbf{H}} \delta(\mathbf{Q}-2\pi\mathbf{H}) \left| \langle F(\mathbf{Q}) \rangle \right|^2 \; + \; N\left\langle \left| \Delta F(\mathbf{Q}) \right|^2 \right\rangle \qquad \text{(A24)}$$

The presence of the δ-function ensures that the first term dominates the scattering at the reciprocal lattice points.

To summarize, there are two simple ways in which structure factors calculated from a simulation of a single molecule can be used to derive time-averaged diffraction intensities for crystals. In one case the structure factors are first squared and then averaged and in the other the structure factors are first averaged and then squared. The problem with the first method is that it cannot be directly related to the average electron density and so attempts to use conventional crystallographic methods to arrive at a model for the intensities calculated that way may not be very successful. When averaged intensities are calculated using both methods and compared, a crystallographic R-factor (Eqn. 9) of 36% is obtained. A refinement of a model structure against the $\langle|F|^2\rangle$ data (i.e. assuming perfectly correlated unit cells) using the average coordinates from the dynamics as the starting model was made. Three positional parameters and a temperature factor were refined for each atom, using data between 10.0Å and 2.0Å. The final R-factor did not drop below 24% which is actually worse than the final R-factor for many proteins at 2.0Å resolution (for well refined structures these are around 15% to 17%). The temperature factors obtained from this refinement were about a factor of 10 lower than those calculated directly from the simulation.

These results provide evidence for the fact that of the two limiting models, that assuming no correlations cells is much closer to the truth than one assuming perfect correlation of the molecular motion in different unit cells. However, some correlation probably does exist and it would be of considerable interest to examine this question in more detail.

REFERENCES

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

Amoros, J.L., and Amoros, M. (1968) "Molecular Crystals: Their Transforms and Diffuse Scattering", John Wiley and Sons, New York.

Artymiuk, P.J., Blake C.C.F., Grace D.E.P., Oatley S.T., Phillips, D.C., and Sternberg, M.J.E. (1979) Nature, 280, 563-68

Blundell,T. and Johnson, L.N. (1976) "Protein Crystallography", Academic Press, New York.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M. (1983) J. Comp. Chem., 4, 187-217

Diamond, R. (1971) Acta Cryst. A27 436-52.

Dunitz, J. (1979) "X-ray Analysis and the Structure of Organic Molecules", Cornell University Press, Ithaca, 514 pages.

Frauenfelder, H., Petsko, G.A., and Tsernoglou, D. (1979) Nature, 280, 558-63.

van Gunsteren, W.F. and Karplus M., (1982a) Biochemistry, 21, 2259-74.

van Gunsteren, W.F. and Karplus M., (1982b) Macromolecues, 15, 1528-44.

Hartmann, H., Parak, F., Steigmann, W., Petsko, G.A., Ponzi, D.R., and Frauenfelder, H. (1982) Proc. Natl. Acad. Sci. (USA) 79, 4967-71.

Hendrickson, W.A. (1980) in "Refinement of Protein Structures; Proceedings of the Daresbury Study Weekend" (ed. Machin, P.A. and Elder, M.) Science and Engineering Research Council, Daresbury Laboratory, U.K., pages 1-8.

Hendrickson, W.A. and Konnert, J. (1980) in "Computing in Crystallography", (ed. Diamond, R., Ramasheshan, S. and Venkatesan, K.) Indian Institute of Science, Bangalore, pages 13.01 - 13.23.

Hirshfeld,F.L. and Rabinovich, D., (1973) Acta Cryst A29, 510-13.

Hubbard, R.E. (1985) "Harvard-York Drawing Program: HYDRA", to be published.

Ichiye, T. and Karplus, M. (1985a) to be submitted.

Ichiye, T. and Karplus, M. (1985b) to be submitted.

International Tables for X-Ray Crystallography, (1974) Vol. IV, (ed. Ibers, J. and Hamilton, W.C.), International Union of Crystallography, The Kynoch Press, Birmingham.

James, R.W. (1948) The Optical Priciples of X-Ray Diffraction, Reissued 1982, Ox Bow Press, Woodbridge Ct.

Jones, T.A. (1982) in "Computational Crystallography" (ed. Sayre, D.) Clarendon, Oxford, 3303-17.

Karplus, M. (1981) Ann. of the N.Y. Acad. Sci. 367 407-18.

Karplus, M. and McCammon, J.A., (1981), C.R.C. Critical Reviews in Biochemistry, 9, 293-349.

Karplus, M. and McCammon, J.A. (1983) Ann. Rev. Biochem. 53, 263-300.

Konnert, J.H. (1976) Acta Cryst. A32, 614-17.

Konnert, J.H., Hendrickson, W.A., (1980) Acta Cryst. A36, 344-49.

Kuriyan, J., Wilz, W., Karplus, M. and Petsko, G.A. to be submitted.

Levy, R.M., Sheridan, R.P., Keepers, J., Dubey, G.S., Swaminathan, S. and Karplus, M. (1985) Biophys. J. (in press).

Levy, R.M. and Keepers, J. (1985) Comments on Molecular and Cellular Biophys., in press.

Levy, R.M., Karplus, M. and Wolynes, P. (1981) J. Am. Chem. Soc. 103 5998.

Mao, B., Pear, M.R., and McCammon, J.A. (1982) Biopolymers, 21, 1979-89.

Moore, F.H. (1963) Acta Cryst. 16, 1169-75.

McCammon, J.A., Gelin, B.R., and Karplus, M. (1977) Nature 267, 585-90.

Northrup, S.H., Pear, M.R., Morgan, J.D., McCammon, J.A., and Karplus, M. (1981), J. Mol. Biol., 153, 1087-90.

Petsko,G.A. and Ringe, D. (1984) Ann. Rev. Biophys. Bioeng. 13, 331-71.

Phillips, S.E.V. (1980) J. Mol. Biol. 142 531-54.

Sheriff, S., Hendrickson, W.A., Stenkamp, R.E., Sieker, L.C. and Jensen, L.H. (1985) Proc. Natl. Acad. Sci. (USA) 82 1104-07.

Stewart, R.F. and Feil, D. (1980) Acta Cryst. A36 503-9.

Takano, T. and Dickerson, R.E. (1981) in "Interaction between Iron and Protein in Oxygen and Electron Transport" (ed. Ho, C.) Elsevier/North Holland, New York.

Ten Eyck, L.F. (1973) Acta Cryst., A29, 183-91

Ten Eyck, L.F. (1977) Acta Cryst., A33, 486-92

Teeter, M.M. and Hendrickson, W.A. (1979) J. Mol. Biol., 127, 219-233

Watenpaugh, K.D. Sieker,L.C., and Jensen L.H. (1980) J. Mol. Biol. 138, 615-33.

Willis, B.T.M. and Pryor, W. (1975) "Thermal Vibrations in Crystallography", Cambridge Univ. Press, London.

Wilson, K.S., Stura, E.A., Wild, D.L., Todd, R.J., Stuart, D.I., Babu, Y.S., Jenkins, J.A., Standing, T.S., Johnson, L.N., Fourme, R., Kahn, R., Gadet, A., Bartels, K.S. and Bartunik, H.D. (1983) J. Appl. Cryst. 16 28-41.

Wlodawer, A., Walter, J., Huber, R. and Sjolin, L. (1984) J. Mol. Biol. 180 301-29.

Woolfson, M.M. (1970) "An Introduction to X-Ray Crystallography" Cambridge University Press, Cambridge, 380 pages.

Xuong, N.H., Freer, S.T., Hamlin, R., Nielsen, C., and Vernon, W. (1978) Acta Cryst. A34, 289-96.

Yu, H. Karplus, M. and Hendrickson, W. (1985) Acta Cryst., B41, 191-201.

Zucker, U.H. and Schulz, H. (1982) Acta Cryst., A38, 563-8.

TABLE 1

The number of independant reflections for myoglobin as a function of resolution.

RESOLUTION	NUMBER OF INDEPENDANT DATA POINTS	DIFFRACTION DATA PER VARIABLE PARAMETER		
		X,Y,Z THERMAL ELLIPSOID	X,Y,Z B	DIHEDRALS ONLY
3.0 A	~ 3000	-	0.6	5
2.0 A	~ 10000	-	2.0	15
1.5 A	~ 22000	1.9	4.3	
1.2 A	~ 40000	3.5	7.9	
0.86A	~125000	11.0	24.8	

Table 2(a) Weights on the various classes of restraints

Type of Restraint	Target standard deviation	Unit
1-2 Bond distances	0.030	Å
1-3 Angle distances	0.040	
1-4 Planar distances	0.052	
Planar groups	0.025	Å
Chiral groups	0.150	Å^3
Temperature factor restraints:		
Backbone bonded pairs	1.0	Å^2 (B-factor units)
Backbone angle pairs	1.5	
Sidechain bonded pairs	1.0	
Sidechain angle pairs	1.5	
Non-bonded contact restraints:		
Non-bonded contact pairs	0.5	Å
Torsional restraints:		
Torsion angles	10.0	degrees
Positional shift restraints:	0.3	Å
B-factor shift restraint:	3.0	Å^{2*}
Occupancy factor shift restraint:	0.05	(unitless)

TABLE 2(b)

DEVIATIONS FROM IDEAL PROLSQ STEREOCHEMISTRY
25 PICOSECOND SIMULATION

	STATISTICS FROM THE AVERAGE DYNAMICS STRUCTURE		STATISTICS DONE ALONG THE SIMULATION	
TYPE	AVERAGE DEVIATION (OVER THE MOLECULE)	R.M.S.	AVERAGE DEVIATION (OVER THE MOLECULE)	R.M.S.
Distances is Angstroms:				
1.Backbone Bonds	-0.02	0.049	0.0007	0.015
2.Carbonyl O Bonds	-0.06	0.100	-0.00827	0.009
3.Sidechain Bonds	-0.11	0.200	0.002	0.023
4.Backbone Angles (1-3 distances)	- 0.019	0.055	0.014	0.040
5.Carbonyl O Angles (1-3 distances)	-0.090	0.150	-0.020	0.020
6.Sidechain Angles (1.3 distances)	-0.140	0.270	-0.005	0.040
Angles in degrees:				
7.Backbone angles	-2.39	4.05	-1.234	2.12
8.Sidechain angles	-3.90	12.72	1.090	3.31

Table 3 Overall Statistics for the Seven Refinements

Description	R-factor	Deviant	r.m.s. delta of			
of structure	%	distances	bonds	angles	1-4 dist	planes
Average structure and fluctuations from 25 ps dynamics	19.2	1093	0.146	0.204	0.160	0.035
Restr1 : $\langle F\rangle^{0.25}$ with loose restraints.	13.6	1372	0.089	0.105	0.099	0.051
Restr2 : $\langle F\rangle^{0.05}$ with loose restraints.	13.6	1391	0.080	0.105	0.105	0.050
Trestr2: $\langle F\rangle^{0.05}$ with tight restraints.	16.6	157	0.026	0.042	0.043	0.018
Unrestr1: $\langle F\rangle^{0.25}$ with no restraints.	13.6	2035	0.175	0.210	0.186	0.095
Unrestr2: $\langle F\rangle^{0.05}$ with no restraints.	12.9	1865	0.171	0.211	0.172	0.112
Altconf: $\langle F\rangle^{0.05}$ alternate conformations.	12.6	1478	0.090	0.111	0.109	0.060
300ps with loose restraints.	21.5	1759	0.098	0.134	0.141	0.055
TARGET stereochemical standard deviations:			0.030	0.040	0.052	0.025

Table 4

The final refined R-factors between various resolution limits are given for refinements of the $\langle F\rangle^{0.05}$ data set from the 25ps simulation.

R-Factors (%) in Shells of Resolution

Resolution in Å

	5.0	4.00	3.20	2.50	2.00	1.75	1.50	OVERALL
1. Restrained refinement	9.6	8.0	8.0	9.9	13.2	18.0	20.4	13.6
2. Unrestrained refinement	7.3	6.0	6.6	9.4	12.9	17.7	21.7	12.9
3. Restrained refinement with alternate conformations.	7.0	6.2	6.4	8.5	11.9	17.4	21.0	12.6

TABLE 5

B-factors for various crystal structures of myoglobin compared with molecular dynamics values:

Source	Backbone Average Temperature Factor	Sidechain Average Temperature Factor
1. Met-myoglobin, Frauenfelder et al. (1979).	11.8	13.1
2. Oxy-myoglobin Phillips (1980).	11.5	21.1
3. 25 ps molecular dynamics	12.3	26.8
4. 25 ps molecular dynamics restrained refinement	11.3	16.5
5. 25 ps molecular dynamics refinement with tight restraints	12.5	14.5
6. 25 ps molecular dynamics unrestrained refinement	11.7	17.6
7. 300 ps molecular dynamics	25.6	48.6
8. 300 ps molecular dynamics restrained refinement	16.8	21.1
9. 300 ps molecular dynamics restrained refinement with higher initial B	19.2	23.5

Table 6 B-factor Variation

The refinement program restrains the variation of B-factors between 1-2 and 1-3 pairs of atoms in bonds and angles. This table gives the average and standard deviation (in paranthesis) of $|B_i-B_j|$ where i and j are atoms in a 1-2 (bond) pair or 1-3 (angle) pair. These values are to be compared with the target value, σ, for each class of restraint. All values are in B-factor units ($\mathring{A}^2$). The abbreviations used for the various structures are explained in Table 7. The B-factor variations were not restrained in the two unrestrained refinements (Unrestr1 and Unrestr2); they are given here for comparison only.

	Backbone bonds	Backbone angles	Sidechain bonds	Sidechain angles
	$\sigma = 1.0$	$\sigma = 1.5$	$\sigma = 1.0$	$\sigma = 1.5$
Restr1	1.92 (1.59)	2.74 (2.39)	2.73 (2.28)	4.01 (3.63)
Restr2	2.14 (1.81)	2.96 (2.65)	3.03 (2.53)	4.26 (3.91)
Trestr2	0.85 (0.66)	1.39 (1.10)	1.08 (0.86)	1.78 (1.55)
Unrestr1	2.82 (2.91)	3.53 (3.75)	4.46 (4.97)	5.46 (5.69)
Unrestr2	2.78 (2.64)	3.68 (3.52)	4.44 (4.90)	5.59 (5.99)
Altconf	2.03 (1.70)	2.85 (2.52)	2.62 (2.10)	3.83 (3.18)
M.D. (25)	2.07 (4.51)	4.10 (5.94)	8.58 (13.30)	11.11 (18.47)

Table 7 Positional Deviations Between Structures

The following table compares the positions of atoms in the 6 refined structures and the molecular dynamics average structure. The backbone atoms (C_α,C,N) of the structures being compared were superimposed by least-squares before the the deviations were calculated. There are three entries in each box: the first entry is the rms deviation in Å between all the atoms in the two structures, the second is for just the backbone (C_α,C,N) atoms and the last entry is for all atoms that are not backbone atoms. The following abbreviations for the structures are used:

M.D.(25): average structure from 25ps of the simulation
restr1 : refinement of $\langle F \rangle^{0.25}$ with loose restraints
restr2 : refinement of $\langle F \rangle^{0.05}$ with loose restraints
trestr2 : refinement of $\langle F \rangle^{0.05}$ with tight restraints
altconf : refinement of $\langle F \rangle^{0.05}$ with alternate conformations
unrestr1: refinement of $\langle F \rangle^{0.25}$ with no restraints
unrestr2: refinement of $\langle F \rangle^{0.05}$ with no restraints

	restr1	restr2	trestr2	altconf	unrestr1	unrestr2
M.D.(25)	0.242 0.132 0.287	0.260 0.137 0.308	0.258 0.103 0.314	0.238 0.145 0.277	0.286 0.181 0.332	0.285 0.200 0.320
restr1		0.116 0.063 0.140	0.122 0.086 0.139	0.149 0.064 0.180	0.197 0.114 0.232	0.222 0.151 0.254
restr2			0.119 0.087 0.133	0.167 0.061 0.204	0.214 0.119 0.253	0.221 0.150 0.253
trestr2				0.178 0.100 0.211	0.260 0.185 0.294	0.251 0.159 0.291
altconf					0.217 0.107 0.260	0.218 0.142 0.252
unrestr1						0.188 0.126 0.216

Table 8 Comparison of B-factors

The following table compares the B-factors of the six refined structures and the exact B-factors obtained from the 25ps simulation. The abbreviations used for the structures is the same as in the previous table. There are four entries in each box:

FIRST entry: The correlation coefficient between the B-factors for all atoms in two structures being compared

SECOND entry: The fractional error defined as:

$$\text{Fractional Error} = \frac{\Sigma |B_i - B_j|}{\Sigma B_i} \quad X \quad 100.0$$

where B_i refers to the i^{th} row and B_j to the j^{th} column.

THIRD ENTRY: The average absolute error $|B_i - B_j|$.

FOURTH ENTRY: The average error $B_i - B_j$.

	restr1	restr2	trestr2	altconf	unrestr1	unrestr2
M.D.(25)	0.82 30.7 6.16 5.71	0.82 30.4 6.09 5.61	0.69 35.9 7.20 6.27	0.90 25.1 5.04 4.53	0.80 35.1 7.03 6.58	0.80 29.2 5.84 4.72
restr1		0.99 5.2 0.75 -0.10	0.93 14.2 2.03 0.56	0.78 12.1 1.73 -1.18	0.94 14.1 2.02 0.87	0.95 12.5 1.78 -0.99
restr2			0.92 15.1 2.18 0.66	0.78 11.4 1.64 -1.08	0.94 14.1 2.03 0.97	0.96 10.7 1.54 -0.89
trestr3				0.66 23.7 3.26 -1.74	0.82 22.9 3.15 0.31	0.84 22.2 3.05 -1.55
altconf					0.76 19.3 2.99 2.05	0.76 16.2 2.51 0.19
unrestr1						0.96 17.4 2.35 -1.86

TABLE 9

STATISTICS ON ANISOTROPY (for 25 ps sampled 0.05ps)

Numbers are averages over all atoms for a particular class except that Prolines were excluded for outer sidechain averages. Numbers in parantheses are standard deviations. The results for lysozyme are taken from Ichiye and Karplus, (1985,a). The anisotropy given is A_1 as defined in the text. The value of A_2 is the same (appoximately 0.15) for all classes of atoms.

	MYOGLOBIN	LYSOZYME
ALL ATOMS	0.68(0.39)	0.85(0.55)
BACKBONE	0.57(0.28)	0.77(0.50)
SIDECHAIN	0.74(0.43)	0.93(0.59)
N	0.55(0.26)	0.68(0.30)
C	0.58(0.28)	0.76(0.45)
O	0.70(0.40)	0.93(0.60)
C_α	0.59(0.30)	0.73(0.47)
C_β	0.67(0.40)	0.74(0.45)
γ	0.72(0.46)	0.90(0.55)
δ	0.76(0.45)	0.95(0.56)
ε	0.85(0.47)	1.03(0.67)
ζ	0.77(0.42)	1.14(0.73)

TABLE 10

Statistics on skewness $|\alpha_3|$ by atom-type for myoglobin and lysozyme.

Numbers are averages are over all the atoms of a particular class except that outer sidechain averages were omitted for Prolines. Numbers in parantheses are standard deviations. The following results are for met-myoglobin (25ps).

$|\alpha_3|$ by atom type for MYOGLOBIN

	U_x	U_y	U_z
ALL ATOMS	0.38(0.32)	0.28(0.25)	0.21(0.21)
BACKBONE	0.36(0.28)	0.26(0.24)	0.21(0.17)
SIDECHAIN	0.40(0.34)	0.29(0.26)	0.22(0.24)
N	0.36(0.27)	0.25(0.22)	0.20(0.18)
C	0.37(0.28)	0.26(0.25)	0.22(0.17)
O	0.41(0.34)	0.26(0.20)	0.22(0.17)
C_α	0.35(0.28)	0.27(0.26)	0.21(0.17)
C_β	0.34(0.30)	0.33(0.28)	0.20(0.16)
γ	0.40(0.37)	0.30(0.25)	0.23(0.42)
δ	0.38(0.34)	0.26(0.25)	0.21(0.18)
ε	0.40(0.35)	0.33(0.31)	0.24(0.20)
ζ	0.43(0.34)	0.31(0.30)	0.22(0.16)

Table 10 contd.

The following data are taken from Ichiye and Karplus (1985a), and are from a 30ps simulation of lysozyme.

$|\alpha_3|$ by atom type for LYSOZYME

	U_x	U_y	U_z
All	0.38 (0.32)	0.25 (0.23)	0.18 (0.16)
Backbone	0.34 (0.28)	0.22 (0.20)	0.17 (0.14)
Sidechain	0.42 (0.36)	0.28 (0.26)	0.20 (0.18)
N	0.30 (0.25)	0.21 (0.16)	0.16 (0.11)
C	0.33 (0.27)	0.22 (0.21)	0.17 (0.13)
O	0.38 (0.33)	0.27 (0.25)	0.18 (0.15)
C_α	0.33 (0.25)	0.19 (0.17)	0.18 (0.15)
C_β	0.32 (0.24)	0.24 (0.22)	0.17 (0.14)
γ	0.40 (0.36)	0.25 (0.20)	0.18 (0.15)
δ	0.45 (0.38)	0.31 (0.30)	0.21 (0.22)
ε	0.53 (0.51)	0.32 (0.27)	0.22 (0.20)
ζ	0.47 (0.36)	0.30 (0.27)	0.21 (0.15)

Table 10 contd.

Kurtosis ($|\alpha_4|$) by atom-type for myoglobin and lysozyme.

Numbers in parantheses are standard deviations. The following data are for met-myoglobin and are calculated from the 25ps simulation.

$|\alpha_4|$ by atom-type for MYOGLOBIN

	U_x	U_y	U_z
ALL ATOMS	0.58(0.58)	0.45(0.46)	0.36(0.67)
BACKBONE	0.56(0.46)	0.43(0.36)	0.36(0.33)
SIDECHAIN	0.59(0.64)	0.46(0.51)	0.37(0.80)
N	0.53(0.51)	0.42(0.33)	0.36(0.44)
C	0.55(0.42)	0.45(0.38)	0.34(0.25)
O	0.56(0.67)	0.42(0.35)	0.31(0.29)
C_α	0.58(0.44)	0.44(0.36)	0.37(0.27)
C_β	0.49(0.46)	0.52(0.61)	0.36(0.47)
γ	0.60(0.75)	0.43(0.44)	0.48(1.75)
δ	0.63(0.56)	0.46(0.48)	0.35(0.28)
ε	0.68(0.71)	0.52(0.69)	0.34(0.27)
ζ	0.72(0.72)	0.53(0.53)	0.33(0.34)

Table 10 contd.

$|\alpha_4|$ by atom-type for LYSOZYME

	U_x	U_y	U_z
All	0.56 (0.52)	0.39 (0.49)	0.31 (0.36)
Backbone	0.50 (0.43)	0.33 (0.37)	0.27 (0.25)
Sidechain	0.61 (0.59)	0.46 (0.58)	0.35 (0.44)
N	0.46 (0.38)	0.30 (0.24)	0.24 (0.19)
C	0.48 (0.38)	0.33 (0.33)	0.26 (0.22)
O	0.57 (0.56)	0.37 (0.56)	0.27 (0.21)
C_α	0.50 (0.38)	0.32 (0.27)	0.31 (0.36)
C_β	0.48 (0.37)	0.36 (0.36)	0.28 (0.20)
γ	0.57 (0.48)	0.42 (0.39)	0.26 (0.25)
δ	0.65 (0.53)	0.54 (0.79)	0.45 (0.72)
ε	0.85 (1.07)	0.61 (0.72)	0.46 (0.51)
ζ	0.67 (0.52)	0.40 (0.52)	0.41 (0.24)

Table 11 Residues with atoms which have large errors in B-factor

The following table is a list of the residues which, in the 25 ps simulation of met-myoglobin, have atleast one atom with a refined B-factor less than half the exact value. The refined structure used is from the refinement of $\langle F\rangle^{0.05}$ with loose restraints. For each such residue the torsion angles which undergo transitions are listed. Since the transitions could be transient, this list does not imply that alternate conformations corresponding to different equilibrium values for these torsions are actually sampled to a significant extent in the simulation (see Text). The average accessible surface areas for the backbone and sidechain atoms in the residue are also listed. A water-sized spherical probe was used in the surface calculations. The residues which were modelled by two alternate conformations in the refinement are labelled with a '*'.

Helix	Residue	Residue		Torsional	Accessible Surface areas	
	number	type	Atom	transitions	backbone	sidechain
NA1	1	VAL	$C_{\gamma 1}$		11.3	17.2
A1	3	SER	C_{β}	none	5.7	18.3
A2	4*	GLU	$O_{\varepsilon 2}$	$\chi_{1,2,3}$	0.0	17.1
A5	7	TRP	$C_{\zeta 3}$	$\chi_{2,3}$	0.0	0.5
A6	8*	GLN	$N_{\varepsilon 2}$	$\chi_{2,3}$	0.7	14.7
A7	9	LEU	$C_{\delta 1}$	$\chi_{1,2}$	2.3	10.1
A8	10	VAL	$C_{\gamma 2}$	χ_{1}	0.0	0.0
A9	11	LEU	$C_{\delta 2}$	$\chi_{1,2}$	0.2	2.9
A10	12	HIS	$C_{\varepsilon 1}$	none	3.5	14.5
B11	30	ILE	$C_{\gamma 2}$	$\chi_{1,2}$	0.0	1.7

B15	34	LYS	N_{ζ}	$\psi,\chi_{1,2,3}$	2.81	13.93
C3	38	GLU	$O_{\varepsilon 1}$	$\chi_{2,3}$	2.3	9.5
CD6	48*	HIS	$N_{\varepsilon 2}$	φ,ψ,χ_{1}	2.6	14.1
D1	51	THR	$C_{\gamma 2}$	none	0.15	18.7
D2	52*	GLU	$O_{\varepsilon 2}$	$\chi_{1,2,3}$	0.3	10.8
D3	53	ALA	C_{β}	ψ	5.2	38.5
D4	54	GLU	$O_{\varepsilon 1}$	$\varphi,\psi,\chi_{1,2,3}$	1.9	8.0
D5	56	LYS	N_{ζ}	$\psi,\chi_{2,3,4}$	0.1	15.3
E2	59*	GLU	$O_{\varepsilon 2}$	$\chi_{1,2,3}$	01.	18.1
E5	62	LYS	N_{ζ}	$\chi_{1,2,3,4}$	0.0	7.9
E10	67	THR	$C_{\gamma 2}$	ψ	2.1	5.7
E12	69	LEU	$C_{\delta 1}$	ψ,χ_{2}	0.0	0.0
E16	73	GLY	O	ψ	0.43	-
EF4	81*	HIS	$N_{\varepsilon 2}$	$\chi_{1,2}$	0.1	13.8
EF5	82	HIS	$N_{\varepsilon 2}$	none	0.0	0.0
F4	89	LEU	$C_{\delta 2}$	$\chi_{1,2}$	0.0	2.4
F6	91	GLN	$N_{\varepsilon 2}$	$\chi_{1,2,3}$	0.1	10.5
FG2	96	LYS	N_{ζ}	χ_{1}	4.1	18.0
FG4	98*	LYS	N_{ζ}	$\chi_{2,3,4}$	0.0	14.1
G10	109*	GLU	$O_{\varepsilon 1}$	$\chi_{1,2,3}$	0.3	7.5
G14	113*	HIS	$C_{\varepsilon 1}$	$\chi_{1,2}$	2.5	9.5
G16	115	LEU	$C_{\delta 1}$	χ_{2}	0.0	0.0
G19	118	ARG	N_{H2}	$\chi_{1,2,3,4}$	2.4	8.9
H2	126	ASP	$O_{\delta 1}$	χ_{2}	6.2	16.0
H12	136	GLU	$O_{\varepsilon 1}$	$\chi_{1,2,3}$	0.0	7.7
H13	137	LEU	$C_{\delta 1}$	$\chi_{1,2}$	0.0	4.3
H15	139	ARG	N_{H2}	χ_{3}	0.3	2.7

H17	141	ASP	O	ϕ	0.0	0.0
H19	143	ALA	O	ϕ	0.0	11.2
H23	147	LYS	N_{ζ}	$\chi_{1,2,3,4}$	1.5	22.0
HC4	152	GLN	C_{γ}	φ, ϕ, χ_3	7.8	6.1

FIGURE LEGENDS

Figure 1

(a) Deviation of sidechain bonds from PROLSQ ideal values. The average bond length, d_{ij} , is calculated in two ways, sampling tne 25ps simulation every 0.05 ps. i) $d_{ij} = |\langle \underline{r}_i \rangle - \langle \underline{r}_j \rangle|$ (dotted line), ii) $d_{ij} = \langle |\underline{r}_i - \underline{r}_j| \rangle$ (solid line). (b,c) Positional error in refinement or the 25ps data: the deviations between atomic positions are calculated arter the molecular dynamics average structure and the refined structures are superimposed by least-squares. The deviations for backbone atoms (N, C_{α} and C, solid line) and sidechain atoms (dotted line) are averagea over residues. Deviations are shown for a restrained refinement ($\langle F \rangle^{0.05}_{restr}$) (Fig. 1b) and for an unrestrained refinement ($\langle F \rangle^{0.05}_{unrestr}$) (Fig. 1c).

Figure 2

(a) Average positional error as a function or $\sigma^2_{m.d.}$ for a restrained refinement of 25ps data, ($\langle F \rangle^{0.05}_{restr}$). The error bars represent +/- one standard deviation, but for points beyond 2.0Å^2 there are relatively few points per average. (b) Distribution of positional errors along the principal X axes for a restrained refinement or the 25ps data,

($\langle F\rangle^{0.05}_{restr}$).

Figure 3

Scatter plots or mean-square fluctuations calculated from the simulation and from the refinements. All the atoms are included in these plots. The exact mean-square fluctuations, $\langle\Delta r_j\rangle^2$, calculated directly from the simulations, are plotted along the Y-AXIS. The refined mean-square fluctuations, obtained from the refined temperature factors are plotted along the X-AXIS. (a) Results of a restrained refinement ($\langle F\rangle^{0.05}_{restr}$) of the 25ps data, (b) Results of an unrestrained refinement ($\langle F\rangle^{0.05)}_{unrestr}$ of the 25ps data, (c) Results of refinement of the 25ps data with ten residues modelled with disordered sidechains ($\langle F\rangle^{0.05}_{altconf}$). The refined mean-square fluctuations for atoms with more than one conformation were obtained by averaging over both conformations. (d) Refinement of the 300ps data with loose restraints. These results were obtained by re-refining the model obtained from the initial refinement after increasing all the B-factors by 13.5 $\mathring{A}^2$ and by scaling the B-factors to minimize the R-factor (see Fig. 6 and text).

Figure 4

Distribution of errors in mean square fluctuation for a restrained and an unrestrained refinement of the 25ps data. The error is defined as $\sigma^2_{m.d.} - \sigma^2_{ref}$. (a) distribution of errors in the restrained refinement results (b) distribution or errors in the unrestrained refinement results.

Figure 5

Residue averages of mean-square fluctuations from molecular dynamics (dotted line) and refinements (solid line). All plots are for tne results or refining the 25ps data. (a) Backbone (N, C and C_α) averages for the restrained $\langle F\rangle^{0.05}_{restr}$ refinement (b) sidechain averages for a restrained ($\langle F\rangle^{0.05}_{unrestr}$) refinement (c) sidechain averages for an unrestrained ($\langle F\rangle^{0.05}_{unrestr}$) refinement (d) sidechain averages for the refinement of the ($\langle F\rangle^{0.05}_{unrestr}$) data with alternate conformations for ten sidechains.

Figure 6

R-factor vs. ΔB. The R-factor as a function of a uniform shift in temperature factor, ΔB, is plotted for the initial refined structure (300ps data) (dotted line) and the structure obtained from increasing the B-factors by 13.5 B-factor units and re-refining (solid line). The R-factor is given by:

$$R = \frac{\Sigma|F_o - F'_c|}{\Sigma|F_o|}$$

where F'_c is given by:

$$F'_c(\underset{\sim}{Q}) = F_c(\underset{\sim}{Q})\, e^{-(\Delta B)\,|\underset{\sim}{Q}|^2}$$

F_c is the structure factor calculated from the refined model.

Figure 7

Errors in the positions and fluctuations vs. the anisotropy A_1 for restrained refinement of the 25ps data. The errors are averaged in bins and the vaiues are plotted with ±1 standard deviation bars. All errors are calculated from the refinement of $\langle F\rangle^{0.05}$ data with loose restraints. (a) Positional error. (b) Error in fluctuations, where error is given by $\sigma^2_{m.d.} - \sigma^2_{ref}$.

Figure 8

Probability distribution functions along the principal X-axis for the 25ps data. Solid line: exact distribution function calculated from the simulation. A spline was used to smooth the data. Large dots: The gaussian determined by the average molecular dynamics position and $\sigma^2_{m.d.}/3$. Small dots: The gaussian determined by tne refined position and $\sigma^2_{ref}/3$. Fig. 8 (a,b,c) are for the $C_{\varepsilon 1}$ atom of Histidine 81. (a) restrained refinement, (b) unrestrained refinement and (c) refinement with two conformations for the Histidine sidechain. The other distributions shown are for (d) Leucine 69 $C_{\delta 1}$, (e) Leucine 11 $C_{\delta 2}$, (f) Aspartate 141 O, (g) Threonine 51 C_γ and (h) Histidine 12 $C\varepsilon 1$

Figure 9

Correlation of the motion of the sidechain of Histidine 12 with twisting oı the A-helix. The twist angle (as defined in the text, solid line) and the positional fluctuation, in Å, of the $C_{\varepsilon 1}$ atom of the imidazole ring (dotted line) are plotted as a function of time. The time-series are taken from the 25ps simulation.

FIGURE 1

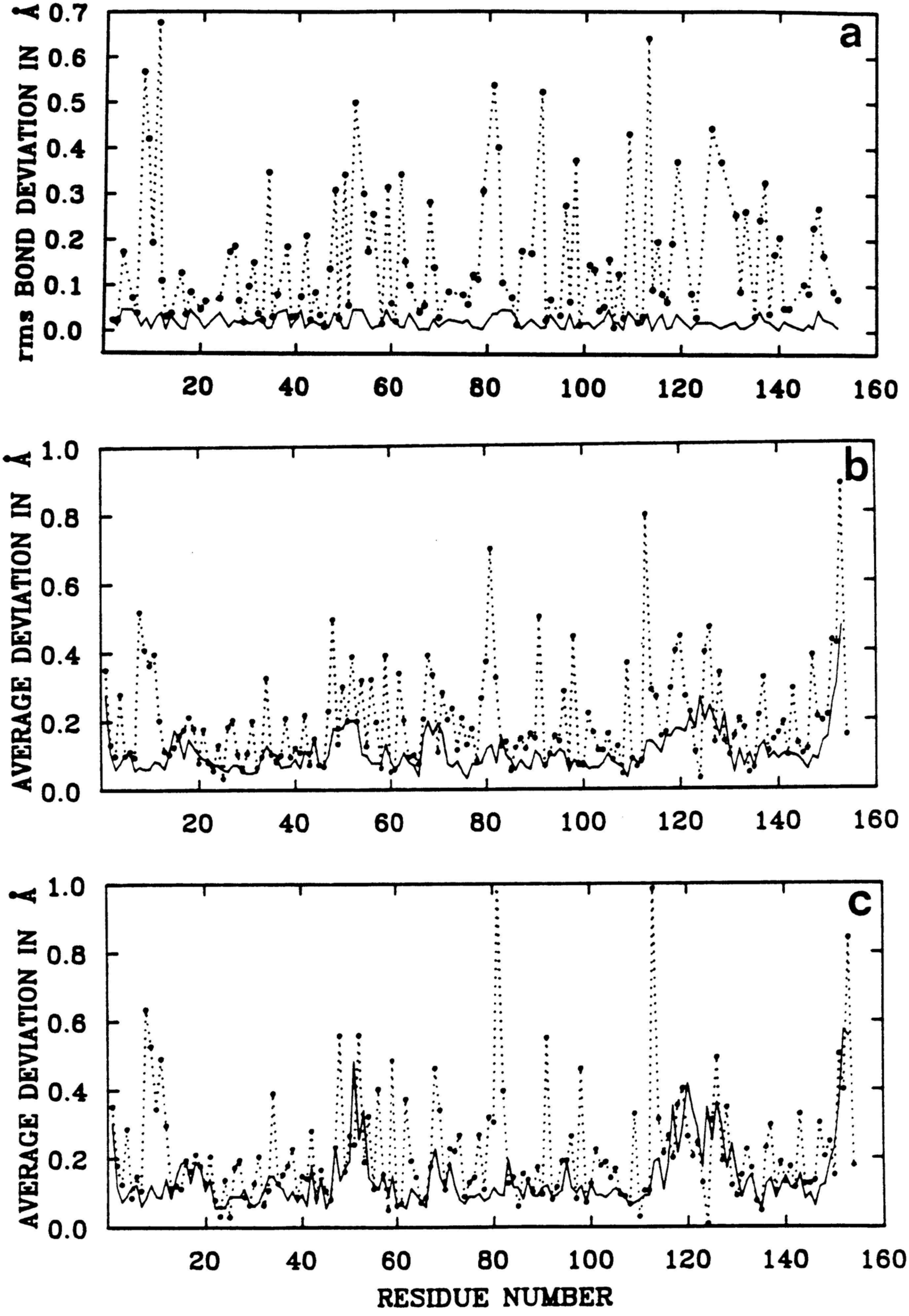

a
rms BOND DEVIATION IN Å
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
20
40
60
80
100
120
140
160
b
AVERAGE DEVIATION IN Å
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100
120
140
160
c
AVERAGE DEVIATION IN Å
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100
120
140
160
RESIDUE NUMBER

FIGURE 2

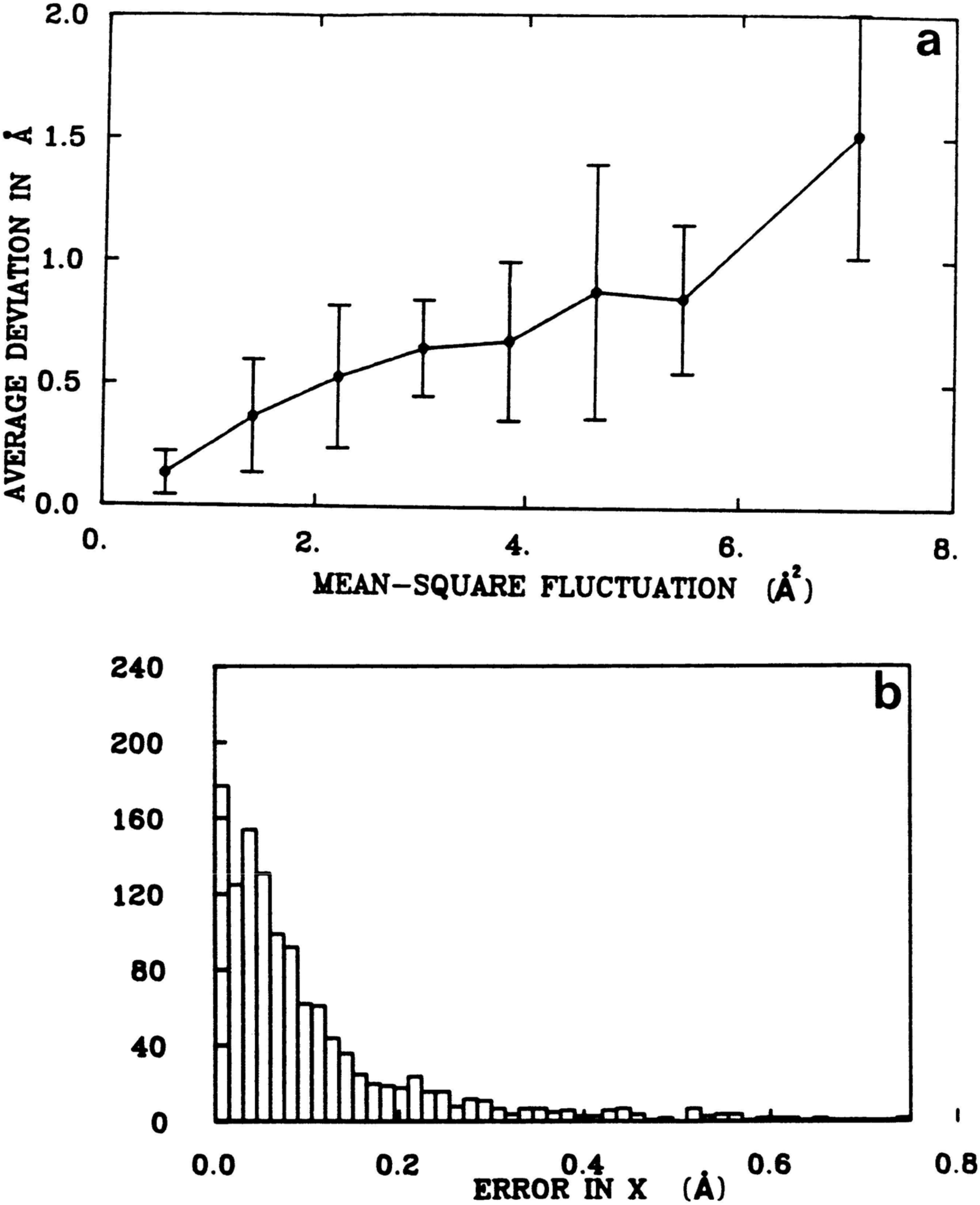

a
2.0
1.5
1.0
0.5
0.0
AVERAGE DEVIATION IN Å
0.
2.
4.
6.
8.
MEAN-SQUARE FLUCTUATION (Å^2)
b
240
200
160
120
80
40
0
0.0
0.2
0.4
0.6
0.8
ERROR IN X (Å)

FIGURE 3

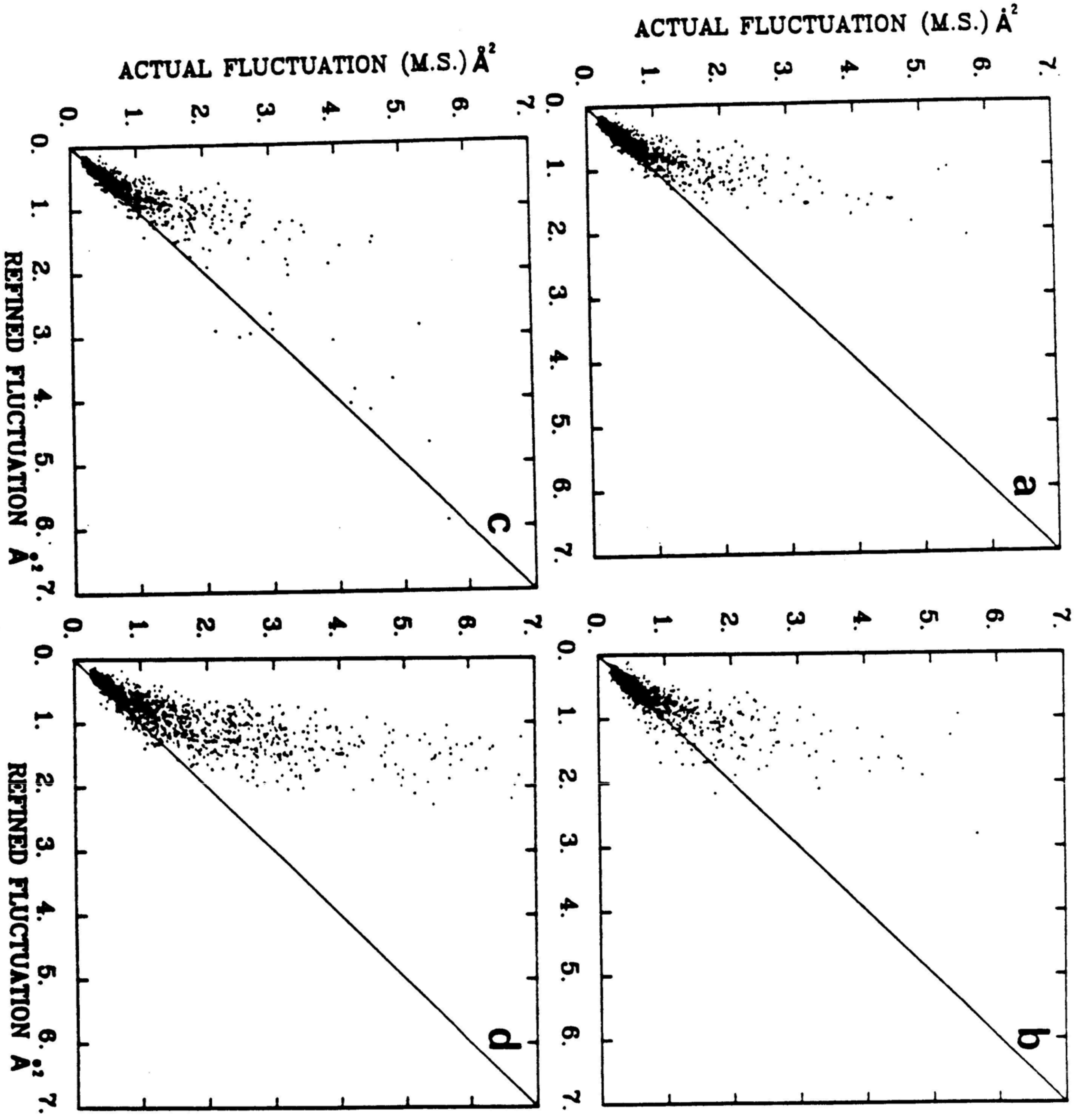

ACTUAL FLUCTUATION (M.S.) Å²
ACTUAL FLUCTUATION (M.S.) Å²
REFINED FLUCTUATION Å²
REFINED FLUCTUATION Å²
0.
1.
2.
3.
4.
5.
6.
7.
a
b
c
d

FIGURE 4

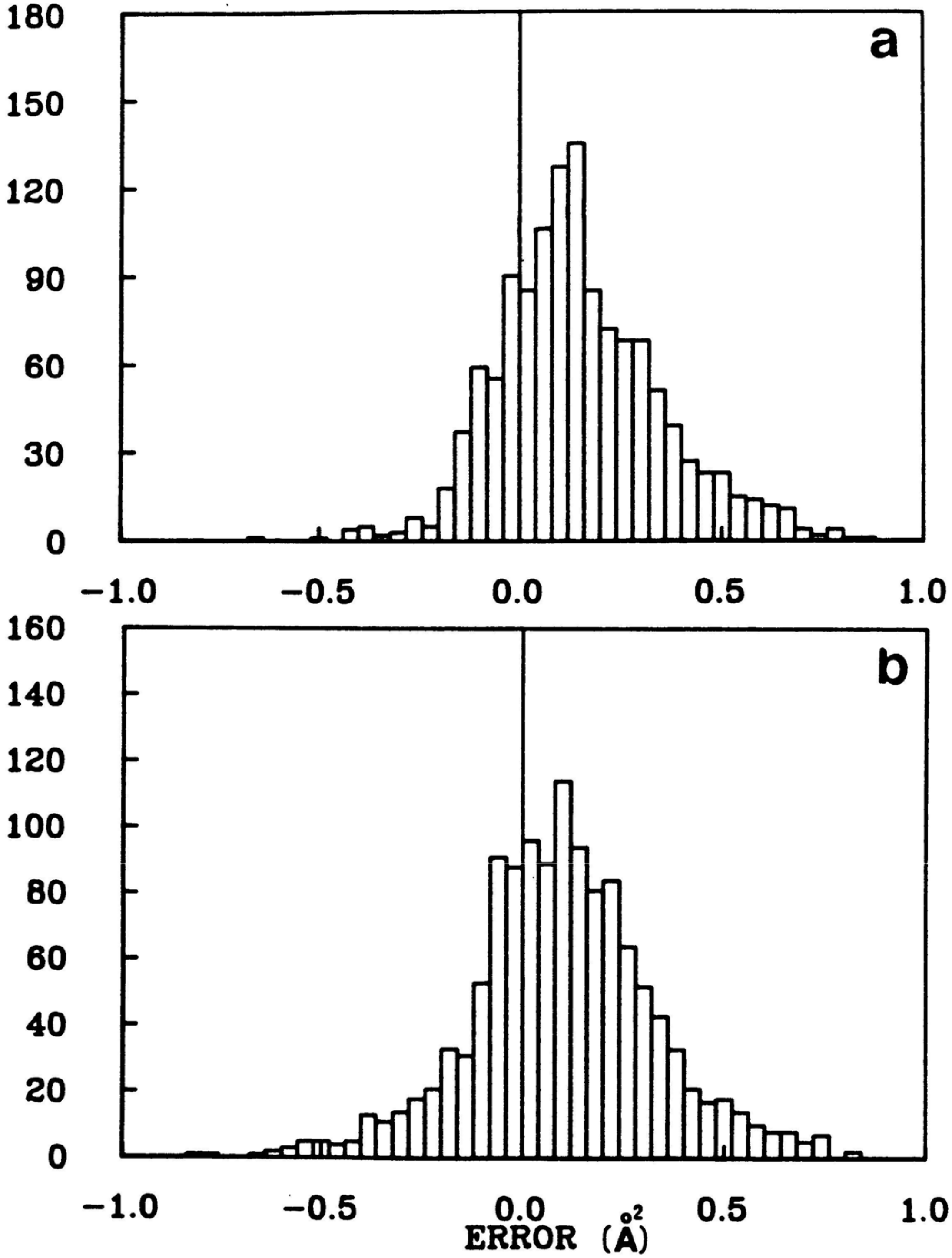

180
150
120
90
60
30
0
a
−1.0
−0.5
0.0
0.5
1.0
160
140
120
100
80
60
40
20
0
b
−1.0
−0.5
0.0
0.5
1.0
ERROR (Å^2)

FIGURE 5

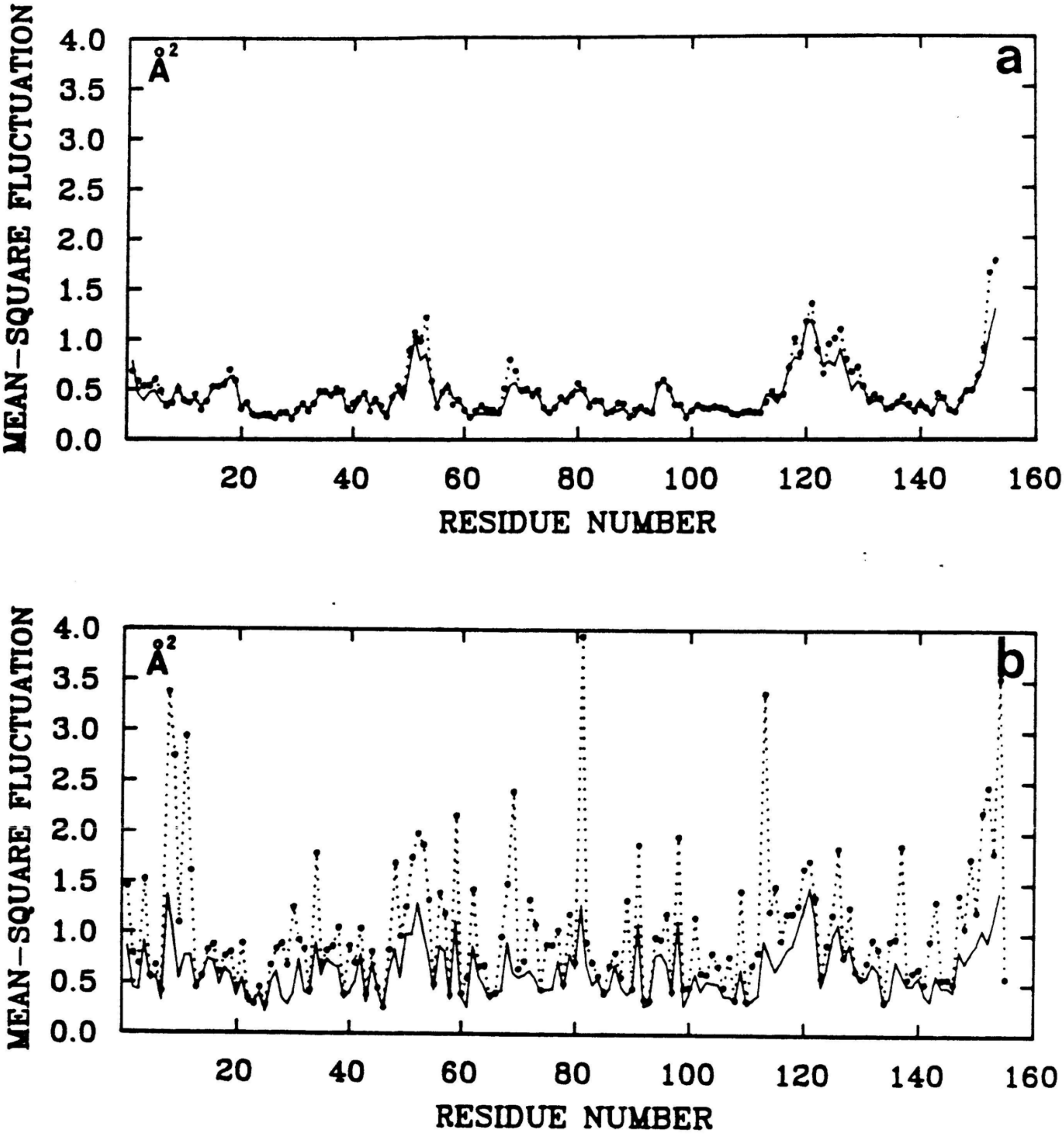

MEAN-SQUARE FLUCTUATION
Å²
a
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20
40
60
80
100
120
140
160
RESIDUE NUMBER
MEAN-SQUARE FLUCTUATION
Å²
b
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20
40
60
80
100
120
140
160
RESIDUE NUMBER

FIGURE 5, contd.

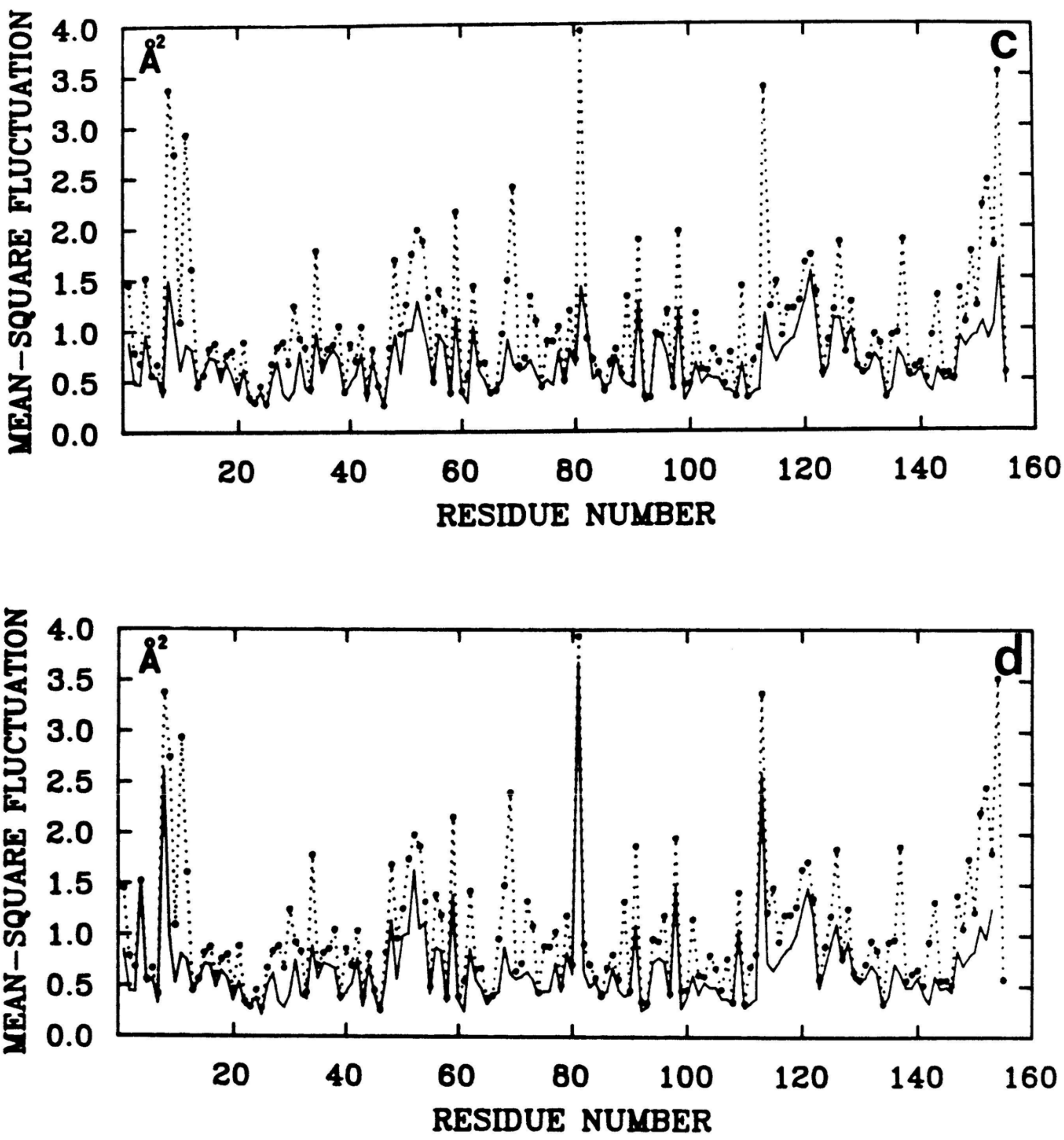


FIGURE 7
(Figure 6 is on page 146)

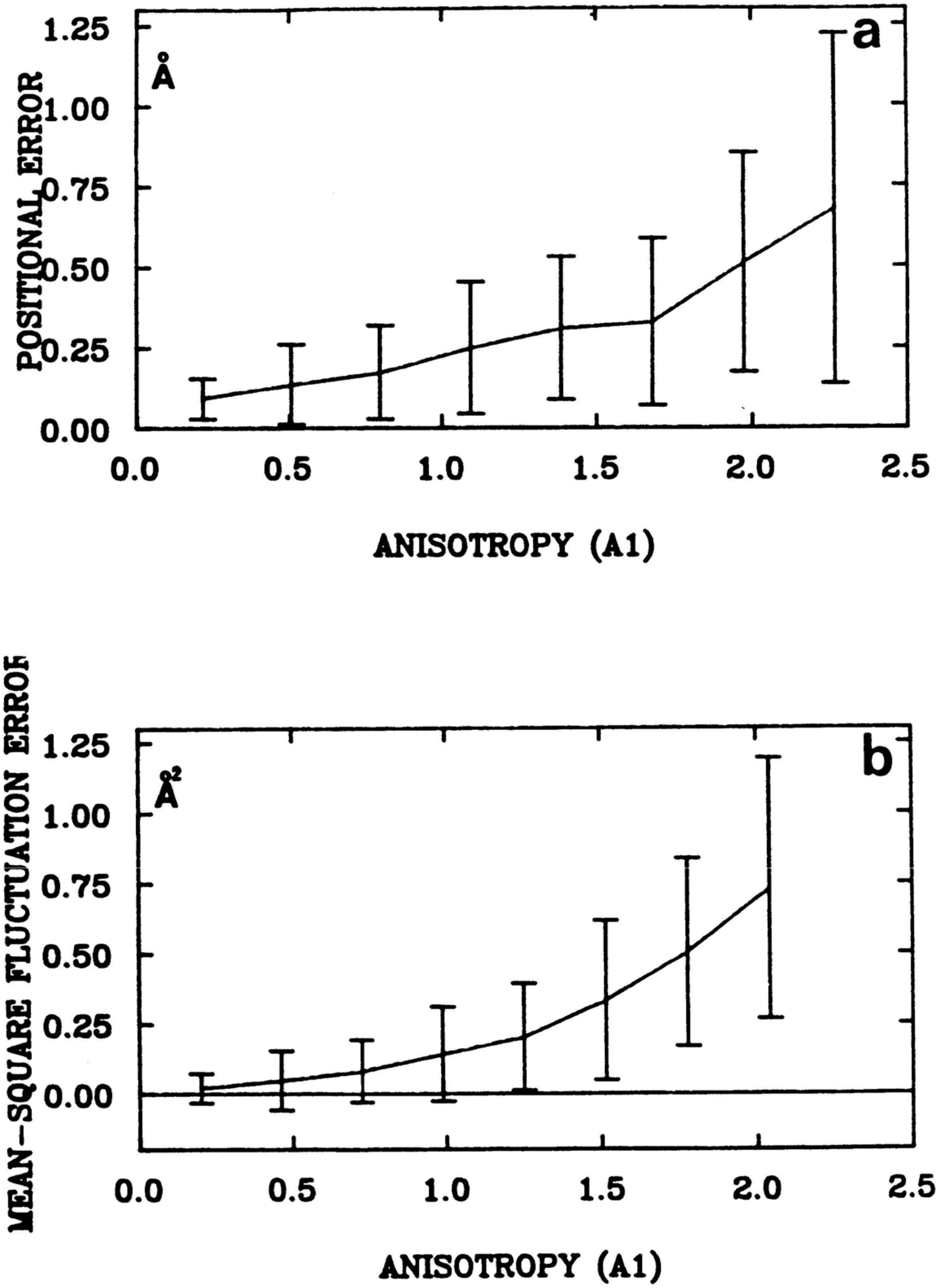
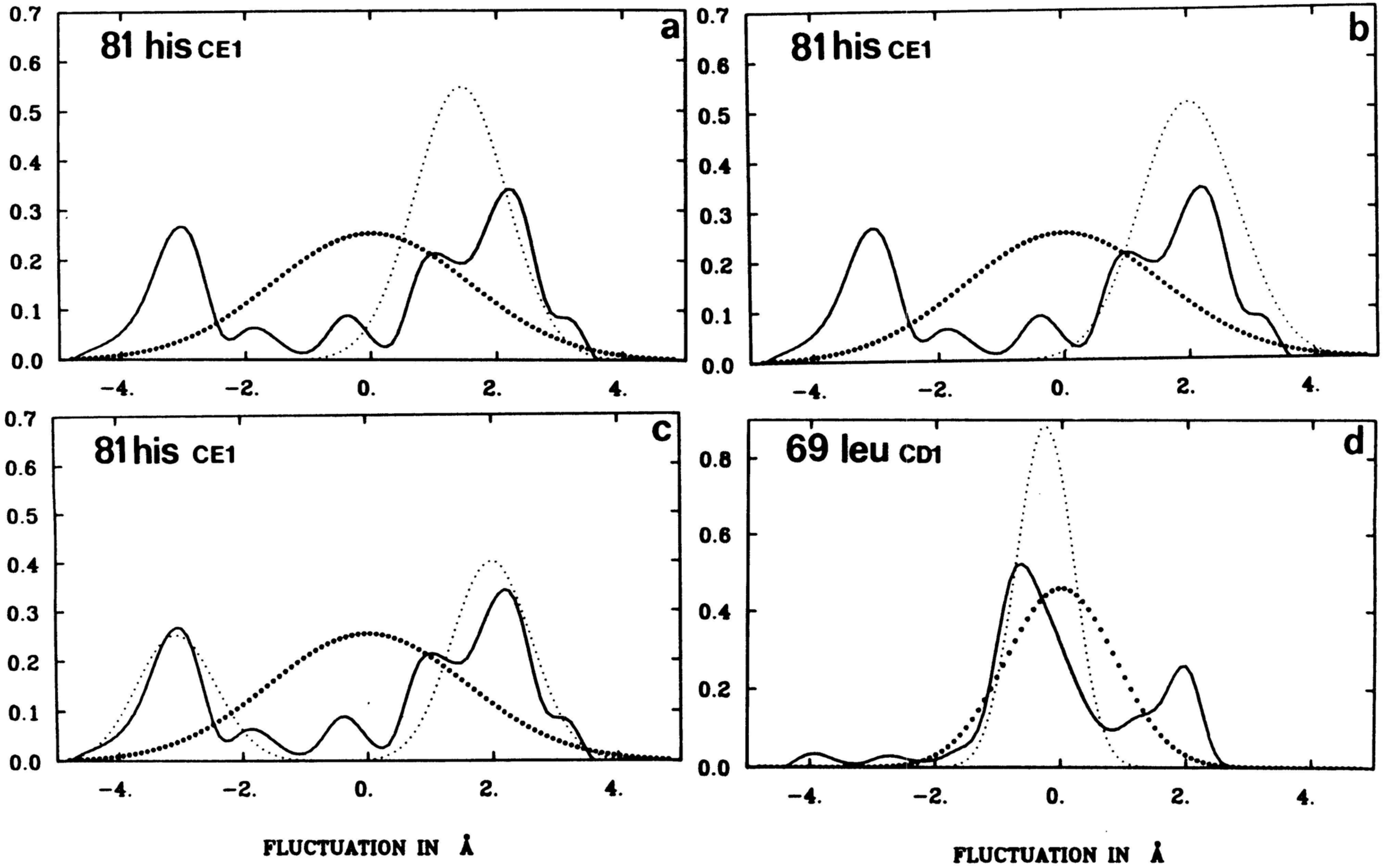


FIGURE 8


81 his CE1
a
81 his CE1
b
81 his CE1
c
69 leu CD1
d
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.8
-4.
-2.
0.
2.
4.
FLUCTUATION IN Å
FLUCTUATION IN Å

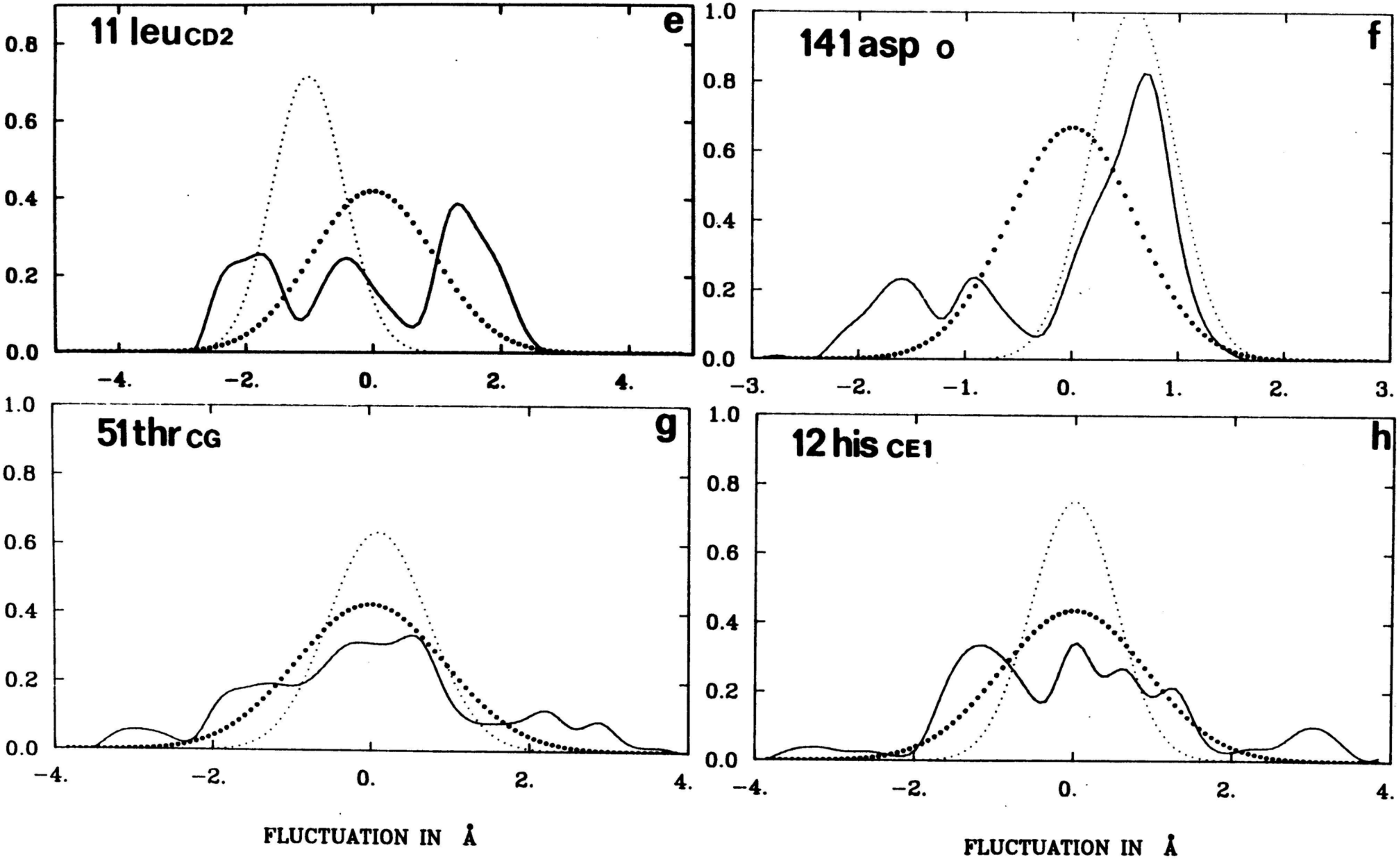


FIGURE 8, contd.

FIGURE 6

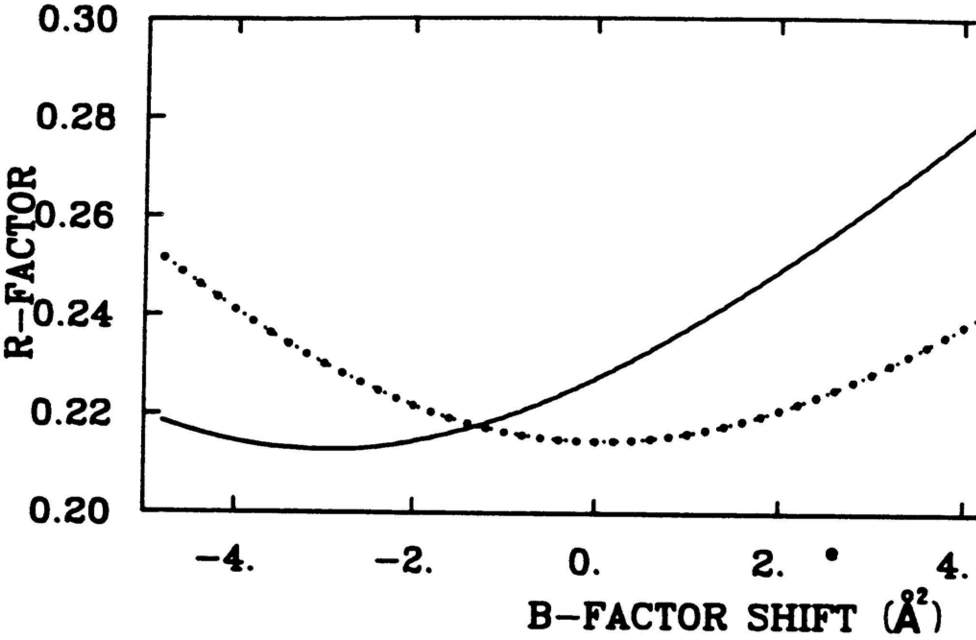
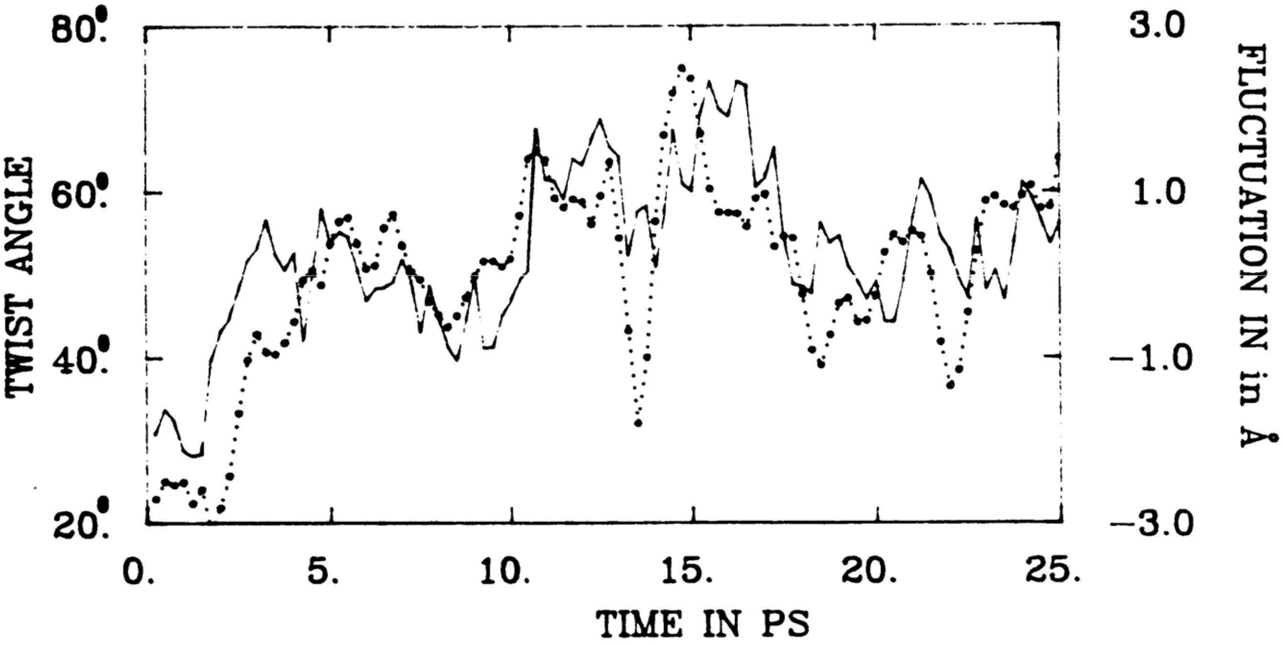


FIGURE 9



Chapter 3

The Thermal Expansion of a Protein: The Structure of Myoglobin at 80K and 255-300K

Abstract

The thermal expansion of a protein, met-Myoglobin, is studied by refinement of structural models against X-ray data at 80K and 255-300K. The results obtained are based on comparisons of two independently refined structures at 80K and four structures at 255-300K, thus reducing the uncertainty due to errors. The unit-cell expands by 5% in the temperature range studied and this expansion is correlated with changes in the structure of the protein; most noticeably, the atoms in the C and D helices move away from the rest of the protein on increasing the temperature. There is also an overall expansion of the molecule: the C_α atoms are, on average, further apart by 0.20Å at 255-300K than at 80K and the radius of gyration of the molecule increases by 0.20Å. The linear expansion coefficient between 80K and 255-300K is calculated to be $50X10^{-6}K^{-1}$ and is fairly constant over length scales from 5.0Å to 30.0Å. The expansion coefficient, however, does vary considerably over different regions of the molecule. This is presumably related to the degree of anharmonicity of the atomic potentials of mean-force, but the data are inadequate to draw definite conclusions about this. Finally, the radial distribution of atoms around the center of the molecule changes with temperature; the net change can be described as a motion of atoms away from the centroid when the temperature is increased.

Introduction

In 1979 Frauenfelder, Petsko and Tsernoglou published the first temperature dependent study of the atomic mobilities in a protein, met-Mb[1] (Frauenfelder et al., 1979, Frauenfelder and Petsko, 1980). This work covered the temperature range of 300K to 250K and in 1983 this was extended to 80K when Parak and Hartmann collected X-ray data on a flash-frozen crystal of met-myoglobin. A model was refined against these data, which extended to 2.0Å resolution, and a comparison of the resulting temperature factors with those at 300K showed that considerable atomic mobility apparently exists even at 80K (Hartmann et al.,1983).

Hartmann et al. (1983) focussed on a comparison of temperature factors at the two temperatures. However, during the course of the refinement, it was noticed that the distances between some atoms increased significantly on going from 80K to 300K. This raised the possibility of directly studying the thermal expansion of a protein by X-ray crystallography, and thereby perhaps learning something about the forces that hold the atoms in a protein together. Extensive comparisons of the 80K and 300K structures of met-Mb have been made by several workers (Frauenfelder et al., 1986). The analysis included estimation of the overall thermal expansion coefficient of the protein (the resulting value is between those for benzene and water) and an examination of the changes in internal cavities and packing defects.[2]

(1) Abbreviations used: met-Mb (sperm-whale Fe(III) myoglobin-OH2); CO-Mb (sperm whale Fe(II) myoglobin, carbon-monoxy); LT, low temperature (80K); RT, "room" temperature, 255K-300K; rms, root-mean-square;

2. It was shown that the expansion occurs by an increase in the volume of the small packing defects in the protein; the four or five large internal cavities do not change significantly (Frauenfelder et al., 1986).

In the original paper the conclusions about the expansion of met-Mb relied entirely on two structures, one at 80K and one at 300K. The magnitude of the changes observed between the two structures is small, and some doubts were raised about whether the observed differences are significantly above the noise level of the data. The reproducibility of the unit-cell parameters for myoglobin needs to be assessed. The observed change in unit cell constants between 80K and room temperature might be correlated with changes in the molecular structure and the significance of these changes needs to be investigated. An estimate of the error limits in the atomic coordinates must be made. Most of the analysis involves comparing distances between atoms at the two temperatures; the standard deviation of errors in the distances needs to be estimated. It has to be shown that the observed changes are significantly greater than the estimated errors in the distances.

The approach taken to estimate the reproducibility of the results is to obtain a new structure of met-Mb at 80K by an independent refinement of a new model against the X-ray data and also to collect three new sets of diffraction data in the temperature range of 255-300K. Having more than one structure at low and high temperature allows us to estimate the errors in atomic coordinates and also to average the results of structural comparisons.

This approach leads to a description of the effects of temperature with a knowledge of the inherant errors that permits its validity to be assessed. The unit-cell volume of met-Mb expands by 5% between 80K and 255-300K. This expansion of the unit-cell is correlated with changes in the structure of the protein that are quite localized, but nevertheless are the largest effects of the change in temperature. Apart from this, there is also a smaller overall expansion of the molecule; on average, the C_α atoms in the protein are further apart by about 0.20Å at 255-300K than at 80K. This effect is anisotropic and is only very roughly correlated with the changes in the temperature factors at the two temperatures. There does not seem to be any noticeable conformational change in the molecule; the expansion apparently occurs by small adjustments in the positions of all the atoms rather than by large, local, changes in the conformational torsion angles. In any case, large positional errors in the 80K structures preclude a description of the effect of temperature in terms of individual atoms at this stage. The anisotropic nature of the thermal expansion provides a picture of the deformable regions of the molecule; this complements that obtained by analysis of the X-ray temperature factors, and also provides an experimental framework within which to test the accuracy of empirical energy calculations on proteins.

In Section II the refinements of the various structures are briefly described and the positional errors in the coordinates are estimated. In Section III the significance and the structural consequences of the changes in the unit-cell on increasing the temperature are examined. In this section the structural changes are also described in terms of the helices and loops in the protein. In Section IV the local as well as the global effects of the thermal expansion are described in terms of the changes in the radius of gyration of the protein, the linear expansion coefficient and the radial distribution of atoms. The conclusions are summarized in Section V.

SECTION II: Refinement and Errors in the Structures

At temperatures above 250K it is straightforward to measure X-ray data on met-Mb to 2.0Å resolution and, in the course of various experiments, three independent data sets were collected, two at 290K and one at 255K. Structural models were independently refined against these data sets and, combined with the original 300K structure of Hartmann et al., (1983), this led to four independent structures of met-Mb in the temperature range of 255-300K. It was not feasible to collect X-ray data at 80K at M.I.T.; instead a new model was refined against the original 80K X-ray data of Hartmann et al. (1983). In the original refinement (Hartmann et al.,1983), the 300K met-Mb model of Frauenfelder et al. (1979) was used as the starting structure; in the new refinement we started with the recently refined model of CO-Mb (Chapter 4 of this thesis).

The four refined models of met-Mb in the 255-300K temperature range shall be collectively referred to as the met-Mb(RT) ("room" temperature) structures. There are now four RT structures and two LT (low temperature, 80K) structures, which makes eight structural comparisons possible instead of just the original one. All the structures used in the analyses were refined against X-ray data by the method of Konnert and Hendrickson (Konnert, 1976, Konnert and Hendrickson, 1980, Hendrickson and Konnert, 1980, Hendrickson, 1980). The refinement of the original 80K structure (met-Mb LT1) and the 300K structure (met-Mb RT1) have been described previously (Hartmann et al.,1983).

i) The new refinement against the 80K data (met-Mb LT2)

The X-ray data set used is identical to that used in the original refinement (Hartmann et al., 1983). It consists of the 4400 unique reflections considered to be statistically significant between the resolution limits of 7.0Å and 2.0Å. This includes only about half the total number of theoretically obtainable reflections and may be contrasted with the new RT data sets, all of which include about 9000 unique reflections between the same resolution limits.

The structure used as a starting model was the 255K refined structure of CO-Mb (Kuriyan et al.,1985). The carbon-monoxide was replaced by a water molecule and all the atoms were assigned an initial uniform temperature factor. The water molecules included in the CO-Mb structure were removed and replaced by those built during the original refinement. The strategy employed in the new refinement was to alternate a few cycles of least-squares structure factor refinement (Konnert and Hendrickson, 1980) with manual examination of difference electron density maps on a graphics system. Every residue in the protein was examined using difference maps with coefficients $(2F_o-F_c)$ (Blundell and Johnson, 1976). Regions where the $(2F_o-F_c)$ density was ambiguous were omitted from the model and difference maps with coefficients (F_o-F_c) were calculated and, if possible, the structure was rebuilt to fit the difference map. For the RT refinements, such a procedure finally results in the density being unambiguous and contiguous for virtually all the backbone atoms. This was unfortunately not true in the 80K refinement; there were several regions of the protein where the backbone density was very poor, even at an advanced stage of the refinement.

In all, three cycles of alternating least-squares refinement and manual rebuilding were done on the structure. The final R-factor is 20%

and the structure has good stereochemistry (rms deviations from ideality in all stereochemical parameters less than or equal to the target standard deviations). More than fifteen sidechains in the 80K structure were observed to be disordered, apparently existing in more than two conformations. No attempt was made to model them as such, because of the small number of X-ray reflections included in the refinement.

ii) The new RT structures (met-Mb RT 2-4)

The refinements of the three new RT structures (met-Mb RT 2-4) are described in Chapter 5 of this Thesis and will not be elaborated upon here.

iii) Errors in atomic positions and distance calculations

In characterizing the effects of temperature on the structure of the protein there would be little ambiguity if the deviations between the two LT structures were significantly less than the deviations between an RT structure and an LT structure. However, re-refinement of the 80K data has resulted in a structure which deviates almost as much from the old 80K (LT) structure as it does from any of the RT structures. One must, therefore, be cautious in assigning particular significance to any observed structural differences between the LT and the RT structures. In this section an attempt is made to arrive at some crude estimates of what would constitute a significant deviation between the two structures. The estimates will be based on the observed rms deviations between the two LT structures and also on the deviations between the four RT structures.

The rms deviations between the backbone and sidechain atoms of all the eight structures have been calculated and these are given in matrix form in Table 1, below. These deviations have been calculated by superimposing the backbone (N,C,C_α) atoms of residues 3 to 148 (the two N terminal and four C terminal residues are omitted from all comparisons). From Table 1 it can be seen that the backbone deviations between the LT structures (0.33Å) is more than twice as large as the average rms deviation of 0.14Å between the RT structures. The average rms deviation between an LT structure and an RT structure is only 0.35Å.

Table 1: Rms Deviations Between the Refined Structures

Upper entry: rms deviation of the backbone atoms, in Å. Only N,C, and C_α are included.

Lower entry: rms deviation, in Å, of all atoms not included in the above category (except solvent atoms).

	LT2 (80K)	RT1 (300K)	RT2 (255K)	RT3 (290K)	RT4 (290K)
LT1 (80K)	0.33 0.84	0.34 0.71	0.37 0.88	0.36 0.84	0.35 0.84
LT2 (80K)		0.37 0.91	0.35 0.88	0.35 0.86	0.35 0.85
RT1 (300K)			0.20 0.63	0.15 0.53	0.15 0.51
RT2 (255K)				0.12 0.36	0.11 0.37
RT3 (290K)					0.06 0.12

To examine this in more detail, in Fig. 1 the deviations between the backbone atoms of the two 80K structures are plotted as a function of residue number. The deviations are clearly not uniform throughout the structure and in certain regions the backbone atoms deviate by more than 1.0Å. These regions are seen to be those where the difference maps either show no density above noise-level or else are uninterpretable. These regions are not confined to the loop regions; for example the region around residue 25 in the B-helix has very poor density (see Fig. 1). The far better quality of the RT difference density maps probably reflects the superiority of the diffractometer data collected at the higher temperatures over the photographic data measured at 80K. Fig. 1 also shows the rms deviations between one of the 80K structures and an RT structure. The magnitude of the deviations are similar to that in the first plot, though the pattern of deviations is different.

If the simplifying assumptions that the errors in all the backbone atoms are independent and that these errors obey the same isotropic Gaussian distribution is made, then the rms deviations between the backbone atoms in Table 1 can be used to obtain an estimate of the standard deviation in the atomic coordinates. In the Appendix it is shown that the standard deviation, σ_x, in any one of the three coordinates of an atom is given by $\sigma_x^2 = \frac{1}{6}\langle\Delta^2\rangle$, where $\langle\Delta^2\rangle$ is the mean-square deviation averaged over all the atoms. In the Appendix it is also shown that the standard deviation, σ_r, in the distance between any two atoms is given by $\sigma_r^2 = \frac{1}{3}\langle\Delta^2\rangle$. When comparing distances in the 80K structure with those in the 300K structure, what is of interest is the standard deviation in the difference between two distances, $\sigma_{\Delta r}^2 = \sigma_{r1}^2 + \sigma_{r2}^2$. In Table 2 the estimated standard deviations in the coordinates, distances, and differences between distances for the LT and RT structures are presented.

Table 2: Standard Deviations inferred from rms deviations

	LT (80K)	RT (250-300K)
Coordinates:		
Å σ_x	0.13 0.34	0.05 (Backbone) 0.17 (Sidechains)
Distances:		
Å σ_r	0.18 0.48	0.07 (Backbone) 0.24 (Sidechains)
Difference in Distance:		
Å $\sigma_{\Delta r}$	0.19 (Backbone) 0.38 (sidechain)	

From this it can be seen that differences in LT and RT backbone distances which are less that about 0.20Å are not likely to be significant. The distances involving sidechains are much less accurate, especially for the surface sidechains, so just the backbone atoms will be used in most of the analysis which follows. Also, the error limits are further reduced in many cases by averaging over the eight possible pairs of LT and RT structures.

Thus, comparison of the two met-Mb(LT) structures shows that the errors in the 80K structure are two or three times larger than those in the met-Mb(RT) structures and that the rms deviation of atomic positions between the two LT structures is almost as large as the deviations observed between the LT structures and the RT structures. At first this seems to rule out the possibility of characterizing the effects of thermal expansion with any degree of certainty, but, as shall be shown below, it turns out that both 80K structures do exhibit certain marked differences from the RT structures. The positions of individual atoms cannot be described accurately enough to demonstrate the effect of thermal expansion on any given atom, but less local quantities such as the radius of gyration, the radial distribution functions and the relative packing of helices and loops show similar changes in all the eight comparisons which have been done.

SECTION III: Crystal Contacts and Secondary Structure Changes

(i) Crystal packing

In Table 3, below, the unit-cell parameters for the four met-Mb (RT) crystals, the met-Mb (LT) crystal and, for comparison, the CO-Mb crystal (Chapter 4 of this thesis) are presented.

Table 3: Unit-Cell Parameters

	a	b	c	β
	(Å)	(Å)	(Å)	
RT1(300K)	64.31	30.85	34.85	105.85°
RT2(290K)	64.53	30.96	34.85	105.79°
RT3(290K)	64.55	30.99	34.80	105.91°
RT4(255K)	64.46	31.02	34.82	105.92°
RT Average	64.46	30.95	34.83	105.87°
RT σ	0.09	0.06	0.02	0.27°
LT (80K)	63.44	30.45	34.05	105.60°
Δ(RT-LT)	1.02	0.50	0.78	0.27°
CO-Mb	64.18	30.84	34.69	105.84°
ΔCO-Mb - Met-Mb(RT)	0.28	0.11	0.14	0.03°

It has not been possible to re-measure X-ray data at 80K and therefore have no estimate of the errors in the low temperature unit-cell parameters. However, the increase in unit-cell lengths between 80K and 255-300K (1.0Å along a, 0.5Å along b, and 0.8Å along c) are at least an order of magnitude larger than the magnitude of the estimated standard deviation in the RT unit-cell lengths. Also, this expansion is about a factor of five larger than the increase in the unit-cell lengths observed on changing the ligation state of the molecule from an in-plane Fe(II)-(CO) heme to an out-of-plane Fe(III)-(OH$_2$) heme, which is known to involve significant changes in the molecular structure (Kuriyan et al., 1985). Thus, the change in unit-cell lengths between 80K and 300K, which corresponds to a 5% increase in cell volume, is likely to be correlated with changes in the molecular structure.

Myoglobin crystals at room-temperature are estimated to contain approximately 600 water molecules per asymmetric unit (Phillips, 1980). With good data and careful refinement, up to 250-300 water molecules (many with low apparent occupancy) can be located in the electron density map (Phillips, 1980); this is typical for crystals of proteins of this size (Blake et al. 1983, North and Smith, 1985). About half as many water molecules can be located at 300K as at 250K, presumably because of increased disorder in the solvent at the higher temperature (Kuriyan, unpublished). Unfortunately, electron density maps calculated using the present 80K data set (Hartmann et al., 1983) are too poor to locate more than about forty solvent molecules and so the effect of temperature on the solvent structure cannot be examined.

We examine whether the changes in the unit cell parameters are reflected in the protein structure by calculating the intermolecular crystal contacts between one protein molecule and all its neighbors in the monoclinic $P2_1$ lattice at low temperature and at high temperature. A crystal contact is defined as an intermolecular distance that is less than 4.0Å, and Fig. 2(a) shows the distribution of these distances for one of the RT structures in the appropriate RT unit cell. There are 96 contact pairs and most of the distances are greater than 3.5Å, the shorter ones being favourable ionic or hydrogen bonding interactions.

The effect of temperature can be seen by re-calculating the contacts for the same RT structure, using the 80K unit-cell instead of the RT one; i.e., the protein is kept rigid while the unit-cell is allowed to contract. The distribution of contact distances for this situation is shown in Fig. 2(b). The number of contacts has doubled and there is a pronounced peak in the distribution between 3.0Å and 3.5Å which was absent in the RT unit-cell contact distribution (Fig. 2(a)). Many of the contacts in this region are due to atoms in van der Waals collision and clearly, there must be changes in the low temperature structure which are correlated with changes in the unit cell. This is indeed the case, and Fig. 2(c) shows the distribution of crystal contact distances obtained from one of the 80K structures in the 80K unit-cell. The peak between 3.0Å and 3.5Å is considerably reduced, and the total number of contacts is only slightly more than that found for the RT structure in the RT unit-cell. There are, however, a somewhat larger number of short contact distances at 80K than at 300K, indicating that the intermolecular contacts themselves do increase between 80K and 300K. However, because of changes in the protein structure, this increase is much less than that calculated using a rigid protein.

The correlation between the changes in the unit-cell and the protein structure is now analysed. In Fig. 3(a) the number of contacts per residue for an RT structure in the appropriate unit-cell is plotted. All the helices and loops, except for the B helix, have at least a few residues that are involved in intermolecular packing interactions. The largest number of such interactions is in the C-D region and the N-terminal end of the E helix, which is interesting because this region is implicated in the formation of transient pathways for ligand entry into the heme pocket (Ringe et al., 1984, Chapter 4 of this thesis). In Fig. 3(b) the increase in the number of contacts per residue when the RT structure is packed into the 80K unit-cell is plotted. The largest increases are, again, in the C-D-E region and also in parts of the A helix.

In Fig. 3(b) the protein is kept rigid while the unit-cell lengths are decreased. In Fig. 3(c) the increase, per residue, in the number of contacts relative to the RT crystal when an 80K structure is packed into the 80K unit cell is plotted; in this case, the protein structure as well as the unit-cell changes relative to the RT crystal. Fig. 3(c) has a striking feature when compared to Fig. 3(b). The increase in contacts in the A helix and the C-D-E region in the latter is almost completely absent in the former; some residues in these regions actually have fewer contacts at 80K than at the higher temperature. We shall examine later whether this reduction in the number of contacts, over that predicted by the change in the unit-cell alone, is a result merely of sidechain rearrangements on the surface of the protein, or whether they involve more extensive backbone rearrangements as well.

In examining intermolecular contacts it is useful to know which parts of the structure are coupled through such contacts and how this coupling changes with temperature. In Fig. 4(a) a contact diagram for an RT met-Mb structure is presented. Each intermolecular contact is shown as a line connecting a residue on the left with a residue on the right. Residue pairs which have at least one contact less than 3.1Å are connected with a dashed line; all others are connected with solid lines. There are only six short contacts for the met-Mb RT structure in Fig. 4(a); five of them correspond to favourable ionic or H-bonding interactions (see below). In Fig. 4(b) we plot a similar contact diagram for an

80K structure in the 80K unit-cell. The over-all pattern is the same in both diagrams; however, there are more short contacts in the 80K structure. Notice that the C-D region is mainly coupled to itself through intermolecular contacts.

As mentioned above, the intermolecular contacts tend to be shorter at 80K than at 255-300K; this can be seen in Fig. 4, for example. To examine a few specific cases, in Table 4 the distances between all the atoms involved in ionic or H-bonding interactions at 255K and at 80K are given. The surprising thing is that there are very few such interactions; the molecular interfaces seem to be stabilized predominantly by solvent mediated interactions, rather than direct contact of protein sidechains.

Table 4: Ionic and H-bonding Intermolecular Contacts

Contact Pair	Interacting Atoms	Intermolecular Distance		Δ
		80K	300K	
	(Å)	(Å)	(Å)	
Ala 19(AB1) - Lys 63(E6)	O-NZ	2.32	2.71	0.39
Lys 62(E5) - Arg 118(G19)	NZ-O	2.40	3.40	1.0
Arg 139(H16) - Lys 147(H24)	NH2-O	2.23	2.53	0.30
Glu 136(H13) - Lys 147(H24)	OE2-NZ	*	2.33	-
Glu 109(G10) - Lys 147(H24)	OE2-NZ	3.17	*	-
Lys 96(FG1) - Glu 109(G10)	NZ-OE1	2.58	2.36	-0.22
Glu 18(A16) - Lys 50(CD8)	OE2-NZ	*	3.07	-
Glu 41(C6) - Lys 77(E20)	OE2-NZ	*	3.20	-
Glu 38(C3) - Lys 79(EF2)	OE1-NZ	1.90	3.0	1.1
Lys 140(H17) - Leu 149(H26)	N-O	3.30	*	-
His 48(CD6) - Glu 52(D2)	NE2-OE2	3.67	*	-

Only five of the eleven interactions are preserved between the LT and the RT structures. Of the five conserved interactions, one

* The distance between these two atoms is greater than 4.0Å.

apparently contracts between 80K and RT; the other four expand, two of them by more than 1.0Å. Many of the residues involved are disordered and alternate conformations have not been included in the analysis (for example, Lys 147(H24), which is observed to have two conformations at 80K, interacts with Glu 136(H13) in one conformation and Glu 109(G10) in another; only one of the interactions is listed in the Table above).

As discussed earlier, the increase in intermolecular distances on going from 80K to 255-300K tends to be smaller than that predicted by the changes in the unit-cell dimensions alone. Fig. 4(c) is a contact diagram which illustrates the adjustments made in the structure which compensate for the change in the unit-cell. Residue pairs where the shortest intermolecular contact at 80K is 0.25-0.75Å _greater_ than that predicted by the unit-cell contraction are connected by dashed lines; residue pairs where the change is greater than 0.75Å are connected by solid lines. The C-D region and the N-terminal end of the E helix show the largest effect: they have seven residue pairs for which the intermolecular contacts have adjusted by more than 0.75Å each.

To summarize, the unit-cell volume increases by 5% between 80K and 255-300K. The changes in intermolecular contacts are, however, smaller than that predicted by the unit-cell changes alone because the protein structure apparently adjusts to preserve the distribution of intermolecular contacts. The largest adjustments are seen to occur in the C-D region and the first few residues of the E helix.

(ii) Packing of secondary structural elements

The differences in the structure of met-Mb at 80K and at 255-300K are now examined. The rms deviation between the backbone atoms of the RT and LT structures does not prove to be a useful indication of the structural changes because the relatively large errors in the 80K structure obscure the analysis (see Section II). Likewise, the backbone torsional angles, which might be expected to provide an indication of the temperature induced conformational changes, are also difficult to analyse because of the large errors in the low temperature structure. In Fig. 5, for example, we plot the difference in the backbone ϕ and Φ angles between an RT structure and an LT structure. Though some of the changes are large, it has not been considered worthwhile to analyse this further.

We must, instead, rely on comparisons that are either averages over a large number of atoms, or which involve examining blocks of atoms together. The C_α atoms are generally the most accurately determined, and to study the temperature induced structural changes we have calculated the distances between all the C_α atoms in the protein at low and high temperature. The LT distances are subtracted from the RT distances and the *differences* are plotted in matrix form. These "$C_\alpha\Delta$-distance matrices" illustrate which regions of the protein move apart or come together on going from 80K to 255-300K.

Fig. 6 shows the $C_\alpha\Delta$-distance matrix for a comparison of the new 80K structure with an RT structure. The upper half of the matrix indicates C_α-C_α distances that expand between 80K and 300K and the lower half indicates those that contract. The matrix is shown in two parts. The first part has levels at 0.15Å and 0.25Å and the second part has levels at 0.30Å and 0.50Å. As discussed in the section on positional errors, we estimate that a change in distance less than 0.20Å is unlikely to be significant. The overall expansion of the molecule is apparent in Fig. 6. It is also clear that this expansion is not uniform. The most systematic feature is the motion of the C-D region and the N-terminal end of the E-helix away from the rest of the protein.

There are two ways in which we can demonstrate that the features of this $C_\alpha\Delta$-distance matrix are, by and large, determined by the X-ray data and not by biases introduced by the refinement or the initial structure. Firstly, the initial structure used in this refinement is the CO-Mb structure (Kuriyan et al.,1985). In Fig. 7 we show the $C_\alpha\Delta$-distance matrix between CO-Mb and the same RT structure that was used in the earlier matrix. The overall expansion in this matrix is very much less pronounced and the pattern of the expansion, which corresponds to the F-helix tracking the motion of the iron away from the heme plane in met-Mb (Kuriyan et al.,1985), is completely different from that seen in the met-Mb(RT-LT) $C_\alpha\Delta$-distance matrix (Fig. 6), where the motion of the F-helix is not a prominent feature. This comparison of the matrices before and after refinement of the structure against the 80K data shows that the features are very sensitive to the X-ray data used.

Secondly, the $C_\alpha\Delta$-distance matrices obtained by using the original 80K structure are very similar to those obtained using the new 80K structure. Fig. 8 shows the $C_\alpha\Delta$-distance matrix obtained by <u>averaging</u> over the eight matrices constructed from the two LT and four RT structures. The matrix clearly improves on averaging: the contraction half of the matrix becomes sparser and the expansion half becomes denser.

From the averaged matrix in Fig. 8 a few generalizations about the temperature induced structural changes can be made. The matrices indicate that, despite the large errors in the 80K data, the expansion of the protein is a systematic, reproducible effect. The most pronounced feature of the expansion is the movement of the C-D region and the N-terminal residues of the E helix as a unit, away from the A helix, the C-terminal end of the E helix, and the F-G-H region. Careful examination of the blocks along the diagonal of the matrices shows that apparently only the E helix expands internally (the N and C terminii move apart); if the other helices do expand, this is not discernible above the noise level of the data, contrary to what was suggested by Hartmann et al. (1983). The observed expansion of the protein occurs mainly by changes in the relative packing of the helices and loops.

In the previous section on crystal contacts it was pointed out that the C-D-E region adjusts the most to changes in the unit-cell. In this section we have shown that this region moves away as a block from the rest of the protein on going from 80K to 255-300K, suggesting that the changes in the backbone positions are correlated with the changes in the unit-cell. In the next section we shall examine less localized features of the expansion and also attempt to separate the effects of the flexible loops and the C-D region from the overall expansion of the rest of the protein.

SECTION IV: Overall measures of Expansion

(i) The radius of gyration

The radius of gyration and its three orthogonal components are measures of the size and shape of the molecule. We have calculated these for all the non-hydrogen atoms in the six met-Mb structures, assuming a unit mass for each atom. The results are given in Table 5 below:

Table 5: Radius of Gyration for all Atoms in the Protein

	R_g	R_x	R_y	R_z	Volume (Ellipsoid)
LT(1) 80K	21.15	10.87	11.79	13.79	
LT(2) 80K	21.11	10.88	11.73	13.77	
RT(1) 300K	21.35	10.98	11.89	13.92	
RT(2) 290K	21.35	10.97	11.90	13.92	
RT(3) 290K	21.34	10.95	11.91	13.91	
RT(4) 255K	21.31	10.93	11.89	13.89	
Average (LT)	21.13	10.87	11.76	13.78	7378.6
Average (RT)	21.34	10.96	11.90	13.91	7599.3
Δ	0.21	0.09	0.14	0.13	

The increase of 0.21Å in the overall radius of gyration corresponds to an overall linear expansion of approximately 1%. If the molecule is considered to be an ellipsoid, this corresponds to a volume increase of 3%. Table 5 above shows that the increase in the radius of gyration is systematic and reproducible and that the expansion is approximately the same in all directions. We now examine the contribution of the loops and the C-D region by excluding all the atoms in these regions from the radius of gyration calculation. The results are given in Table 6.

Table 6:

Radius of Gyration for all Atoms Except the Loops and the C-D Corner

	R_g	R_x	R_y	R_z	Volume (Ellipsoid)
LT(1) 80K	19.15	10.28	10.67	12.14	
LT(2) 80K	19.10	10.21	10.66	12.11	
RT(1) 300K	19.28	10.32	10.75	12.23	
RT(2) 290K	19.28	10.32	10.74	12.23	
RT(3) 290K	19.27	10.32	10.74	12.23	
RT(4) 255K	19.24	10.33	10.70	12.20	
Average (LT)	19.12	10.25	10.67	12.12	5552.4
Average (RT)	19.27	10.32	10.73	12.22	5668.1
Δ	0.15	0.07	0.06	0.10	

The overall expansion of the radius of gyration is now only 0.8% and the volume expansion is only about 2%. Nevertheless, the expansion is still quite marked. Thus, even though the motion of the C-D region away from the rest of the protein is the dominant feature of the thermal expansion, there is also a smaller overall expansion of the protein.

(ii) Thermal expansion coefficients from C_α-C_α distances

We have examined the magnitude of the thermal expansion over different length scales by using C_α-C_α distance scatter plots. In Fig. 9 we compare the C_α-C_α distances in an LT structure and an RT structure. The distances are grouped into three ranges: (a)5-10Å, (b)10-15Å and (c)20-25Å. All distances that differ by less than 0.25Å are excluded from the plot. In Fig. 9(a) (for distances between 5Å and 10Å) the expansion is seen as a small but perceptible increase in the number of points in the expansion side of the diagram. This feature is clearer in Fig. 9(b), for distances between 10Å and 15Å. Finally, in Fig. 9(c), for distances between 20Å and 25Å, the number of points on the expansion side is overwhelmingly larger.

The distributions of differences in C_α-C_α distances, Δ, for the three classes of distances are shown in Fig. 10. All three distributions are peaked at positive values of Δ, indicating the overall expansion of the molecule. The data in Figs. 9 and 10 are based on comparison of one LT and one RT structure. Similar results are obtained for all the other structures; the mean values of Δ and the standard deviations about the mean for the eight comparisons of LT and RT structures are given below in Table 7.

Table 7: Average Increase in C_α-C_α distances

	met-Mb(RT1) (300K)	met-Mb(RT2) (290K)	met-Mb(RT3) (290K)	met-Mb(RT4) (255K)
met-Mb(LT1)				
⟨Δ⟩	0.23	0.23	0.21	0.19
σ	0.29	0.28	0.28	0.29
met-Mb(LT2)				
⟨Δ⟩	0.20	0.20	0.19	0.16
σ	0.31	0.28	0.28	0.28

The fact that the average value of Δ scales with the distance indicates that the linear expansion coefficient might be constant over all length scales. This is approximately true, as we shall now show. We define a linear expansion coefficient, α_{ij}, for the C_α atoms, as follows:

$$\alpha_{ij}(T_2,T_1) = \frac{r_{ij}(T_2) - r_{ij}(T_1)}{\langle r_{ij}\rangle (T_2-T_1)}$$

where $r_{ij}(T)$ is the distance between the C_α atoms of residues i and j at temperature T, and $\langle r_{ij}\rangle$ is the average value of r_{ij} between the two temperatures. We have calculated the value of α_{ij} averaged over all the C_α-C_α distances in various ranges of distances. The results are given below, in Table 8.

Table 8: Linear Expansion Coefficient for C_{α} Distances

The values of α in this Table are given in units of $10^{6}K^{-1}$.

Range:	All	5-10Å	10-15Å	15-20Å	20-25Å	25-30Å

LT1(80K)
RT1(300)
	51	53	50	46	49	50

LT2(80K)
RT1(300K)
	47	42	50	41	41	43

LT1(80K)
RT2(290K)
	53	52	50	47	51	54

LT2(80K)
RT2(290K)
	48	40	51	42	43	45

LT1(80K)
RT3(290K)
	50	50	46	43	48	51

LT2(80K)
RT3(290K)
	45	37	47	39	39	42

LT1(80K)
RT4(255K)
	53	53	47	44	51	56

LT2(80K)
RT4(255K)
	47	39	48	39	41	46

Average	49.4	45.7	48.6	42.6	45.4	48.4
σ	2.9	6.4	1.7	2.8	4.6	4.4

The value of α is seen to be always lower when calculated using the new 80K structure (LT2), but is fairly constant over all the length scales used in the calculations. The average values of α are close to those obtained for the expansion of all the atoms with respect to the center of mass of the protein ($50 \times 10^{-6}K^{-1}$).

It would be interesting to relate α, as calculated here, to the linear expansion coefficients of substances such as water and benzene. For such a comparison, α has to be converted to the linear expansion coefficient at a particular temperature; in its present form it is the coefficient between two widely disparate temperatures. Such a conversion would require a knowledge of the temperature dependence of the expansion coefficient for myoglobin, which is not known at present. We shall not, in this work, estimate what the correction to α should be in order to relate it to the room-temperature thermal expansion coefficients of other substances.

It might not be unrealistic, given good data, to measure the temperature dependence of α from X-ray determinations of the structure of the protein at different temperatures. This is suggested by the fact that the structure of met-Mb at 255K seems to exhibit a very small but systematic shrinkage when compared to the structures of met-Mb at 290-300K (see Tables 5 and 6, the radii of gyration).

(*iii*) Variation of the Thermal Expansion Coefficient

The values of α in Table 9 are averaged over the whole structure. To look for local variation in the expansivity of the protein, we have calculated α in spheres of radius 5Å and 10Å, centered on each residue. The variation of α with residue number for both cases is shown in Fig. 11. For the smaller spheres, part of the large variation in α observed is due to the small number of distances in each sample. While the 10Å sphere calculation in Fig. 11(b) loses some of the local structure, it still shows a surprising amount of variation. One rough correlation that emerges is that α is lower in the middle of helices than at the ends, perhaps indicating more harmonic potentials in the centers. An interesting result is that the value of α is low in the C-D region in both Figs. 11(a) and 11(b); this is consistent with the C-D region moving as a relatively rigid body away from the rest of the protein.

Since the expansivity is related to the anharmonicity of the potential of mean-force for each atom, we have attempted to discover correlations between the variation in expansivity over the molecule with the

B-factors and, especially, the changes in B-factors between 80K and 255-3.00K. Such correlations turn out to be tenuous, at best, because of the poor quality of the data. We must once again emphasize that that we are limited to drawing only very general conclusions about the thermal expansion. The sensitivity of the B-factors to the errors in the data, among other things, makes it impossible at present to reliably correlate them with expansivity; such correlations would also require knowledge of the temperature dependance of the B-factors.

(iv) The radial distribution function

The distribution of differences in distances of all atoms from the center of mass of the protein at 255-300K and 80K are very similar to the distributions for C_α-C_α distances shown in Fig. 10 and, indeed, lead to the same estimate of the linear expansion coefficient. We have examined the nature of the expansion with respect to the center of mass by calculating the normalized radial distribution function around a central atom in all six structures. The distributions are calculated for the atom closest to the centroid of the protein, which is the CBB atom of the heme group. The radial distributions are averaged over the two 80K structures and also over the four RT structures. The resulting averaged distributions are shown in Fig. 12. They are quite structured, with six or seven peaks clearly seen.

What is interesting in Fig. 12 is that the second, third and fourth peaks are shifted outward by about 0.5Å in the RT distribution, corresponding to an expansion in the coordination shells on going from 80K to 300K. We can compare this with the observed increase in the radius of gyration by calculating the radius of gyration, R_g, for two radially symmetric spherical bodies with mass distributions given by the radial distributions in Fig. 12. On doing so we find that R_g is 21.25Å for the 80K distribution and 21.46Å for the RT distribution. This increase of 0.20Å in R_g is the same as that calculated using all the atoms in the structure.

SECTION V: CONCLUSIONS

Myoglobin exhibits a small but systematic thermal expansion in the temperature range of 80K to 300K. The challenge in the present work has been to characterize this expansion despite the relatively poor quality of the low temperature data. This has been done by re-refinement of a new model against the 80K X-ray data, using a different set of initial coordinates. The room temperature data have also been extended by collecting three new X-ray data sets in the temperature range of 255-290K. Certain overall features of the thermal expansion then emerge as systematic, reproducible, differences between the low temperature and room temperature structures.

The volume of the unit-cell increases by about 5% on going from 80K to 255-300K. However, expansion in the inter-molecular contacts is not as large as that predicted by the expansion in the unit-cell alone because the protein structure adjusts to the changes in the unit-cell. The largest adjustments occur in the C-D region and in the N-terminal residues of the E helix. On examining the temperature induced changes in the structure of the protein, the largest change observed involves the motion of the C-D-E region away from the rest of the protein. Thus one picture of the thermal expansion which emerges from this work is that of localized regions in the protein undergoing expansions that are correlated with changes in the packing of the molecule in the unit cell. The solvent probably plays an important role in this since the protein interfaces are mainly stabilized by protein-solvent interactions rather than protein-protein interactions; however, the data are not good enough to examine this phenomenon in any depth.

A complementary picture of the thermal expansion emerges on examining the radius of gyration of the molecules at the two temperatures: there is a small overall expansion of the molecule that is independent of the motion of the C-D region and the loop regions. On examining C_α-C_α distances it is seen that they move apart by about 0.20Å, on average, between 80K and 300K; the expansion in the distances increases with the distances, resulting in a linear expansion coefficient that is reasonably independant of length scale. The linear expansion coefficient, however, varies in different parts of the molecule. Careful temperature dependent studies are required to extract information about the atomic potentials of mean force from the variation in expansivity; the data used in the current work are by no means adequate.

On examining radial distribution functions around a central atom in the protein it is seen that there are several "coordination" shells around the atom. The structure of the coordination shells changes on increasing the temperature; some of the shells move outward by about 0.5Å. This motion outward is equivalent to an increase of 0.2Å in the radius of gyration and provides a picture of the thermal expansion as a global movement of atoms outwards from the center of the protein.

That such a clear picture of the shrinkage emerges from the present 80K data set bodes well for future temperature dependence studies of the structure of proteins. Another promising feature is that there seems to be a small, but noticeable, difference in the structure of met-Mb at 255K from that at 290-300K. This difference has largely been ignored in this work, but it indicates that X-ray diffraction studies at 2.0Å resolution might be sensitive to the changes in protein structure that occur due to a 40K change in temperature. These studies can readily be extended to at least 1.5Å resolution, especially with the use of area detectors, and so we look forward to some advances in this area.

Crystallographic studies in which thermodynamic parameters such as pressure or temperature are varied provide a means of identifying the deformable regions in proteins; in met-Mb, for example, the temperature dependence study described in this chapter has shown that the C-D-E region is very flexible and can move by about 0.5Å in response to changes in the temperature and/or crystal packing. Recent molecular dynamics simulations of CO-Mb (Kuriyan and Karplus, in progress) indicate that the fluctuations in this region are directly coupled to the ligand binding site and that the motion of the C-D region away from the protein might be a mechanism for the formation of transient channels for ligand entry into the binding pocket. In the crystal, large motions of this region are limited by the intermolecular contacts. The behaviour of the C-D region when the protein is in solution is of great interest.

Acknowledgements

I would like to thank Hans Frauenfelder for originally suggesting this work. This chapter is an outgrowtn of work that was originally done in collaboration with a number of workers (Frauenfelder et al., 1986).

APPENDIX

Positional Errors in Atomic Coordinates Inferred from Two Refinements of the Same Data.

Two independent refinements of the same diffraction data result in a mean-square deviation, $\langle\Delta^2\rangle$, averaged over all the atoms in the structure. In this appendix we relate this mean-square deviation to the standard deviation, σ_x, in the atomic coordinates and to the standard deviation, σ_r, in the distance between any two different atoms in the molecule. To do this we make the following assumptions:

i) The errors in each of the three coordinates of any atom are independent and the error in the position of any atom is uncorrelated with that of any other atom.

ii) The errors in the positions of all the atoms in the molecule obey the <u>same</u> isotropic normal distribution; alternatively, this could apply to certain classes of atoms, such as those in the backbone.

Results

a) The standard deviation, σ_x, in any one coordinate of an atom is related to the mean-square deviation, $\langle\Delta^2\rangle$, between the two refined structures by:

$$\sigma_x = \frac{1}{6^{1/2}} \langle\Delta^2\rangle^{1/2} \approx 0.41 \langle\Delta^2\rangle^{1/2}$$

b) The standard deviation, σ_r, in the distance between any two different atoms in the molecule is given by:

$$\sigma_r = \frac{1}{3^{1/2}} \langle\Delta^2\rangle^{1/2} \approx 0.58 \langle\Delta^2\rangle^{1/2}$$

Proof

a) The relation between σ_x and $\langle\Delta^2\rangle$

The two refinements result in two structures, 1 and 2. The square-deviation for any atom is defined as:

$$\Delta^2 = \sum_{\alpha=1}^{3} (X_{\alpha 1} - X_{\alpha 2})^2 \qquad (1a)$$

Denote the (unknown) true or mean value of the α^{th} coordinate by X_α. Then,

$$X_{\alpha 1} = X_\alpha + \Delta X_{\alpha 1} \quad \text{and} \qquad (1b)$$

$$X_{\alpha 2} = X_\alpha + \Delta X_{\alpha 2} \quad , \qquad (1c)$$

where $\Delta X_{\alpha i}$ is the error in the α^{th} coordinate of the atom in the first or second structure. Therefore:

$$\Delta^2 = \sum_{\alpha=1}^{3} (\Delta X_{\alpha 1} - \Delta X_{\alpha 2})^2 \qquad (2)$$

The mean-square deviation is obtained by averaging Δ^2 over the whole structure. Because of assumptions (a) and (b) given above, this is equivalent to averaging the errors over the single Gaussian distribution which describes the errors. Hence,

$$\langle\Delta^2\rangle = \sum_{\alpha=1}^{3} \left[\langle\Delta X^2_{\alpha 1}\rangle + \langle\Delta X^2_{\alpha 2}\rangle - 2\langle\Delta X_{\alpha 1}\Delta X_{\alpha 2}\rangle\right] \qquad (3)$$

Since the errors are assumed to be independent, the cross term in Eqn. 3 vanishes. Also, the error distribution is assumed to be isotropic, and so:

$$\sigma_x^2 = \langle \Delta X_{\alpha i}^2 \rangle = \langle \Delta X_{\beta j}^2 \rangle \quad \text{for } \alpha,\beta = 1,2,3 \text{ and for } i,j = 1,2 \qquad (4)$$

Hence,

$$\langle \Delta^2 \rangle = 6\sigma_x^2 \quad \text{and} \quad \sigma_x = \frac{1}{6^{1/2}} \langle \Delta^2 \rangle^{1/2} \qquad (5)$$

which proves the first proposition. The relation in Eqn. 5 was checked by a computer simulation in which two met-Mb structures were generated from one of the refined structures by introducing random, Gaussian, errors into the atomic coordinates. In this case the standard deviation, σ_x, in each atomic coordinate is known, and the rms deviation between the two "corrupted" structures is indeed that predicted by Eqn.5.

b) σ_r in terms of $\langle \Delta^2 \rangle$

Method 1 - Propagation of Errors:

Suppose there is a set of variables, $\{x_i\}$, for which the standard deviations, σ_i, for each x_i, are known. Then the standard deviation, σ_f, to first order, of any function f(x) is given by:

$$\sigma_f = \sum_i \left[\frac{df}{dx_i}\right]_T^2 \sigma_i^2 \qquad (6)$$

where the "T" indicates that the derivatives are evaluated at the "true" or mean values of the variables.

Let the distance between two atoms, 1 and 2, be denoted by r:

$$r = \left[\sum_{\alpha=1}^{3} (X_{\alpha 1}-X_{\alpha 2})^2\right]^{1/2}$$

Then,

$$\left[\frac{dr}{dX_{\alpha i}}\right]^2 = \frac{(X_{\alpha 1}-X_{\alpha 2})^2}{r^2}$$

Hence, using Eqn. 6:

$$\sigma_r^2 = \frac{1}{(r^2)}\sum_{\alpha=1}^{3}\left[(X_{\alpha 1}-X_{\alpha 2})^2 \sigma_{x_{\alpha 1}}^{\ 2} + (X_{\alpha 1}-X_{\alpha 2})^2 \sigma_{x_{\alpha 2}}^{\ 2}\right] \tag{7}$$

Using Eqn. 4, this simplifies to:

$$\sigma_r^2 = \frac{\sigma_X^2}{r^2}\sum_{\alpha=1}^{3} 2(X_{\alpha 1}-X_{\alpha 2})^2 \quad = 2\sigma_X^2 \tag{8a}$$

But, from Eqn. 5, $\sigma_X^2 = \frac{1}{6}\langle\Delta^2\rangle$, and so,

$$\sigma_r = \frac{1}{3^{1/2}}\langle\Delta^2\rangle^{1/2} \tag{8b}$$

which proves the second proposition.

Method 2: σ_r evaluated by Taylor expansion.

The error propagation formula, Eqn. 6, is derived using a Taylor expansion for f(x) in terms of the errors in x_i. We can also derive Eqn. 8 by directly applying the Taylor expansion method to the distances, which has the merit of providing a clearer geometrical picture. Let the true position of atom 1 be (X,Y,Z) and that of atom 2 be (0,0,0). Let the measured position of atom 1 be $(X+\Delta X_1, Y+\Delta Y_1, Z+\Delta Z_1)$ and that of atom 2 be $(\Delta X_2, \Delta Y_2, \Delta Z_2)$. Let the distance between the true positions of the atoms be r and that between the measured positions be r'. σ_r^2, the variance of the distance between atoms 1 and 2, is given by:

$$\sigma_r^2 = \langle (r-r')^2 \rangle \tag{9a}$$

Expanding the square, we get:

$$\sigma_r^2 = r^2 + \langle r'^2 \rangle - 2r\langle r' \rangle \tag{9b}$$

wnere r' is given by:

$$\begin{aligned} r'^2 = r^2 + &\sum_{\alpha=1}^{3}\left[(\Delta X_{\alpha 1})^2 + (\Delta X_{\alpha 2})^2\right] \\ &+ 2\sum_{\alpha=1}^{3} X_\alpha(\Delta X_{\alpha 1}-\Delta X_{\alpha 2}) \\ &- 2\sum_{\alpha=1}^{3} \Delta X_{\alpha 1}\Delta X_{\alpha 2} \end{aligned} \tag{10}$$

The last two terms in the above expression vanish on averaging, and so:

$$\langle r'^2 \rangle = r^2 + 6\sigma_X^2 \tag{11}$$

To evalaute the last term in Eqn. 9b, we need to expand r' in a Taylor series in $\{\Delta X_{\alpha i}\}$. The terms which are first order in $\Delta X_{\alpha i}$ will vanish on averaging. Likewise, among the second order terms, all the cross terms will also vanish on averaging. The Taylor expansion for r', keeping only the second order self terms, is:

$$r' = r + \frac{1}{2}\sum_{i=1}^{2}\sum_{\alpha=1}^{3}\left[\frac{d^2 r'}{d(\Delta X_{\alpha i})^2}\right]_{\Delta X_{\alpha i}=0}(\Delta X_{\alpha i})^2 + \dots \tag{12}$$

We can write, from Eqn. 10,

$$r' = (r^2 + \gamma)^{1/2} \tag{13}$$

where:

$$\begin{aligned} \gamma = &\sum_{\alpha=1}^{3}\left[(\Delta X_{\alpha 1})^2 + (\Delta X_{\alpha 2})^2\right] \\ &+ 2\sum_{\alpha=1}^{3} X_\alpha(\Delta X_{\alpha 1}-\Delta X_{\alpha 2}) \end{aligned}$$

$$- \ 2 \sum_{\alpha=1}^{3} \Delta X_{\alpha 1} \Delta X_{\alpha 2}$$

We can evaluate the derivatives as follows:

$$\frac{dr'}{d\Delta X_{\alpha i}} = \frac{1}{2}(r^2+\gamma)^{-\frac{1}{2}} \frac{d\gamma}{d\Delta X_{\alpha i}} \quad \text{and} \tag{14}$$

$$\frac{d^2r'}{d\Delta X^2_{\alpha i}} = \frac{1}{2}(r^2+\gamma)^{-\frac{1}{2}} \frac{d^2\gamma}{d\Delta X^2_{\alpha i}} - \frac{1}{4}\left[\frac{d\gamma}{d\Delta X_{\alpha i}}\right]^2 (r^2+\gamma)^{-3/2} \tag{15}$$

For i=1 (the first atom),

$$\frac{d\gamma}{d\Delta X_{\alpha 1}} = 2\Delta X_{\alpha 1} + 2X_{\alpha} - 2X_{\Delta}X_{\alpha 2} \tag{16a}$$

For i=2 (the second atom),

$$\frac{d\gamma}{d\Delta X_{\alpha 2}} = 2\Delta X_{\alpha 2} - 2X_{\alpha} - 2X_{\Delta}X_{\alpha 1} \tag{16b}$$

and, for i = 1 and 2,

$$\frac{d^2\gamma}{d\Delta X^2_{\alpha i}} = 2 \tag{17}$$

Using Eqns. 16 and 17 in Eqn. 15 and evaluating the expressions at $\Delta X_{\alpha i} = 0$,

$$\left[\frac{d^2r'}{d(\Delta X_{\alpha i})^2}\right]_{(\Delta X_{\alpha i}=0)} = \frac{1}{r} - \frac{X^2_{\alpha}}{r^3} \tag{18}$$

Hence, using Eqn. 18 in Eqn. 12,

$$r' = r + \frac{1}{2}\sum_{i=1}^{2} \sum_{\alpha=1}^{3} \left[\frac{1}{r} - \frac{X^2_{\alpha}}{r^3}\right](\Delta X_{\alpha i})^2 \tag{19}$$

Averaging, and using the assumption of a single isotropic error distribution for all the atoms, we get:

$$\langle r' \rangle = r + \frac{2\sigma^2_X}{r} \tag{20}$$

Using Eqns. 11 and 20 in Eqn. 9b we obtain:

$$\sigma_r^2 = 2r^2 + 6\sigma_x^2 - 2r\left[r + \frac{2\sigma_x^2}{r}\right]$$

$$=\rangle \quad \sigma_r^2 = 2\sigma_x^2 \qquad (21)$$

which is the same expression obtained using error propagation in Eqn. 8a.

REFERENCES

Blundell, T. and Johnson, L.N. (1976) "Protein Crystallography", Academic Press, New York.

Blake, C.C.F, Pulford, W.C.A. and Artymiuk, P.J. (1983) J. Mol. Biol. 167, 693-723.

Case, D.A. and Karplus, M. (1978) J. Mol. Biol. 123, 697-701.

Frauenfelder, H., Petsko, G.A., and Tsernoglou, D. (1979) Nature, 280, 558-563.

Frauenfelder, H. and Petsko, G.A. (1980) Biophys. J. 32, 465-483.

Frauenfelder, H., Petsko, G.A., Ringe, D., Kuriyan, J., Tilton, R.F., Parak, F., Hartmann, H., Kuntz, I.D., Max, N., Connolly, M.L., and Karplus, M. (1986) to be published.

Hartmann, H., Parak, F., Steigemann, W., Petsko, G.A., Ponzi, D.R., and Frauenfelder, H. (1983) Proc. Natl. Acad. Sci. (USA) 79, 4967-4971.

Hendrickson, W.A. (1980) in "Refinement of Protein Structures; Proceedings of the Daresbury Study Weekend" (ed. Machin, P.A. and Elder, M.) Science and Engineering Research Council, Daresbury Laboratory, U.K., pages 1-8.

Hendrickson, W.A. and Konnert, J. (1980) in "Computing in Crystallography", (ed. Diamond, R., Ramasheshan, S. and Venkatesan, K.) Indian Institute of Science, Bangalore, pages 13.01 - 13.23.

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

Konnert, J.H. and Hendrickson, W.A., (1980) Acta Cryst. A36, 344-349.

North, A.C.T. and Smith, J.C. (1985) Int. J. Biol. Macromol., 17, 223-225.

Phillips, S.E.V. (1980) J. Mol. Biol. 142, 531-554.

Ringe, D., Petsko, G.A., Kerr, and D. Ortiz de Montellano, P.R. (1984) Biochemistry 23, 2-4.

FIGURE LEGENDS

Figure 1

RMS deviations between backbone atoms of the two 80K structures (solid line) and an 80K structure and an RT structure (dotted line). The RMS deviations are averaged over residue. The structures being compared are superimposed on the backbone (N,C,C_α), atoms of residues 3 to 148. The circles mark the residue numbers.

Figure 2 a,b,c

Distribution of crystal contact distances. a) RT structure in RT unit cell. b) RT structure in 80K unit cell c) 80K structure in 80K unit cell.

Figure 3

TOP: Number of contacts vs. residue number for RT structure in RT unit cell.

MIDDLE: Increase in the number of contacts when the same RT structure is placed in the 80K unit cell.

BOTTOM: Increase in the number of contacts (with reference to the figure at TOP) when the 80K structure is placed in the 80K unit cell.

Figure 4 a,b,c Contact Diagrams

4a. Intermolecular contacts made between residues in the met-Mb 255K structure. The secondary structure elements are shown in the two vertical bars. Intermolecular contacts are represented as lines connecting residues in one bar to residues in the other. Solid lines represent contacts between 3.1Å and 4.0Å. Dashed lines represent residue pairs where at least one contact is less than 3.1Å.

4b. Intermolecular contacts between residues in an 80K structure, as in 4a above.

4c. Intermolecular distances which adjust to the change in unit-cell. The shortest intermolecular distance for every residue pair is considered for this Figure. Contacts in the 80K structure which are between 0.25Å to 0.75Å greater than that predicted by the change in the unit cell parameters are shown by a dashed line. Contacts which are more than 0.75Å greater are shown by a solid line.

Fig. 5 Φ and ϕ deviations

The deviations in the backbone torsion angles Φ and ϕ between an RT structure and an LT structure are given as a function of residue number. Solid line: deviation in Φ, Dashed line: deviation in ϕ.

Figure 6 a,b

$C_\alpha\Delta$ distance matrix for the new 80K structure vs. an RT structure. The entries below the diagonal represent contraction on going from 80K to 300K; the entries above the diagonal represent expansion on going from 80K to 300K. a) 0.15 and 0.25 Å levels. b) 0.30 and 0.50 Å levels.

Figure 7

$C_\alpha \Delta$ distance matrix, as above, for CO-Mb vs. the same RT structure used in Fig. 4. Levels are at 0.15 and 0.25 Å.

Figure 8 a,b

$C_\alpha \Delta$ distance matrix, averaged over the eight comparisons of 80K structures with RT structures. a) 0.15 and 0.25 Å levels. b) 0.30 and 0.50 Å levels.

Figure 9 C_α-C_α distance scatter plots

The three figures compare C_α-C_α distances in an LT structure with the corresponding distances in an RT structure. All distances which deviate by less than 0.25Å between RT and LT are excluded in the plots. The comparison is done in three ranges of distances: (a) 5-10Å (b) 10-15 Å and (c) 20-25Å.

Figure 10 a,b,c Distributions of differences in C_α distances

Histograms of the differences between C_α-C_α distances at 80K and 300K in the three ranges of distances used in Fig. 9.

Figure 11 a,b coefficients of expansion

The coefficient of expansion, α, as defined in the text is calculated and averaged over all the C_α atoms in spheres of radius 5.0Å (Fig. 11a) and 10.0Å (Fig. 11b), centered on each residue. The comparison is done for one RT structure and one LT structure.

Figure 12 averaged radial distribution functions

The spherically averaged radial distribution of atoms around the CBB atom of the heme is calculated and averaged over the two LT structures (solid line) and the four RT structure (dotted line).

FIGURE 1

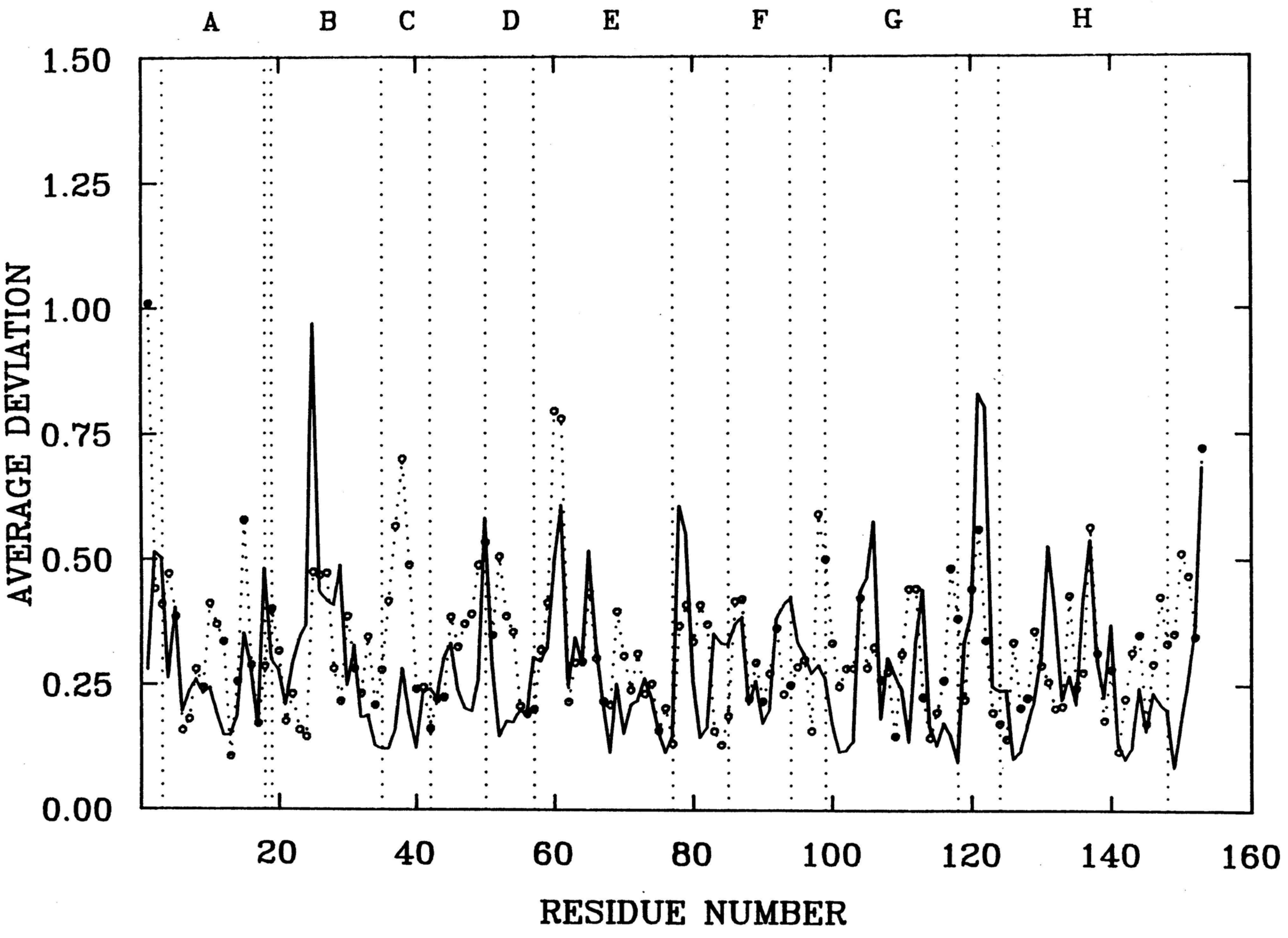

A
B
C
D
E
F
G
H
1.50
1.25
1.00
0.75
0.50
0.25
0.00
AVERAGE DEVIATION
20
40
60
80
100
120
140
160
RESIDUE NUMBER

FIGURE 2(a)

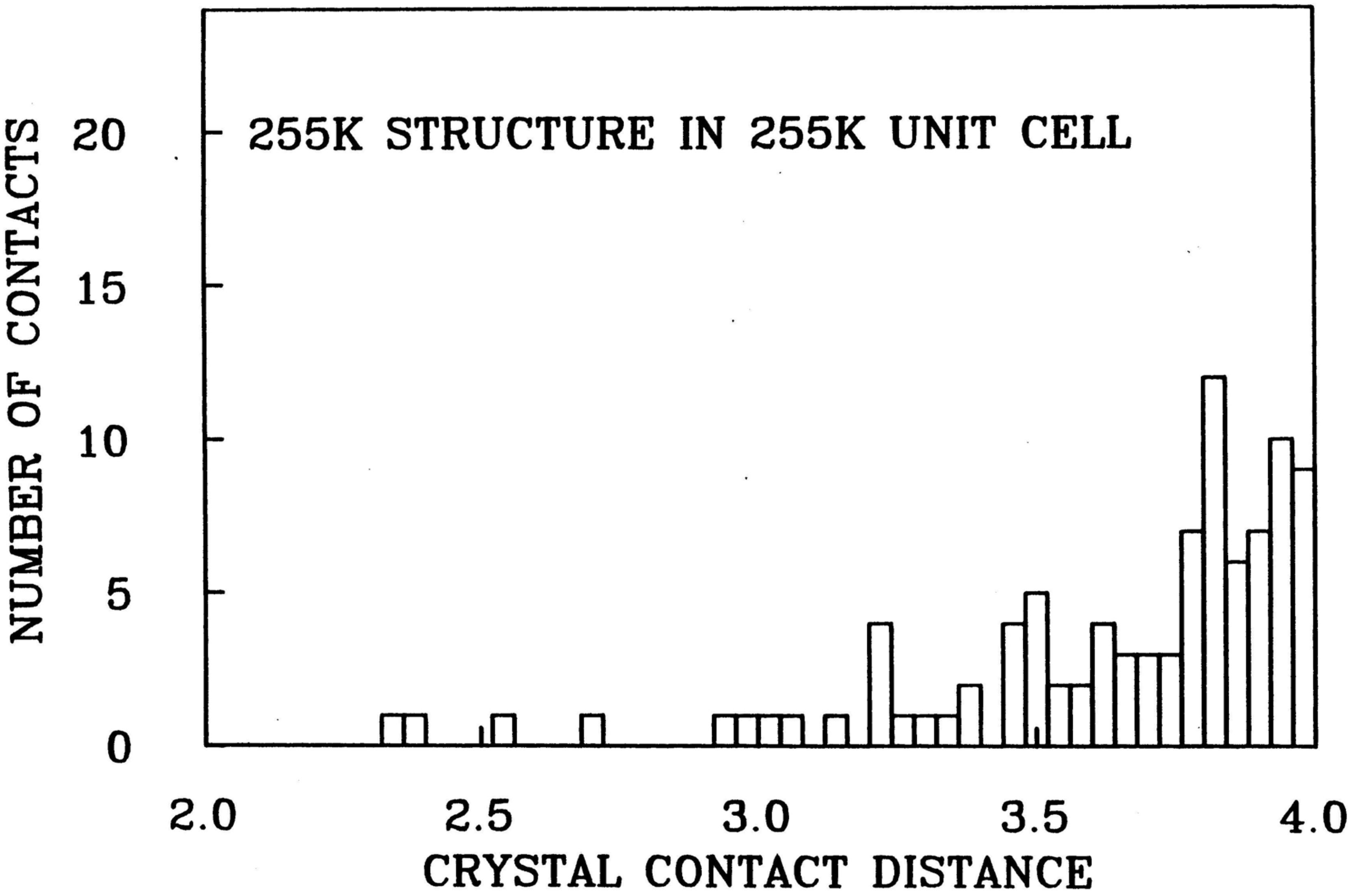

255K STRUCTURE IN 255K UNIT CELL
NUMBER OF CONTACTS
0
5
10
15
20
2.0
2.5
3.0
3.5
4.0
CRYSTAL CONTACT DISTANCE

FIGURE 2(b)

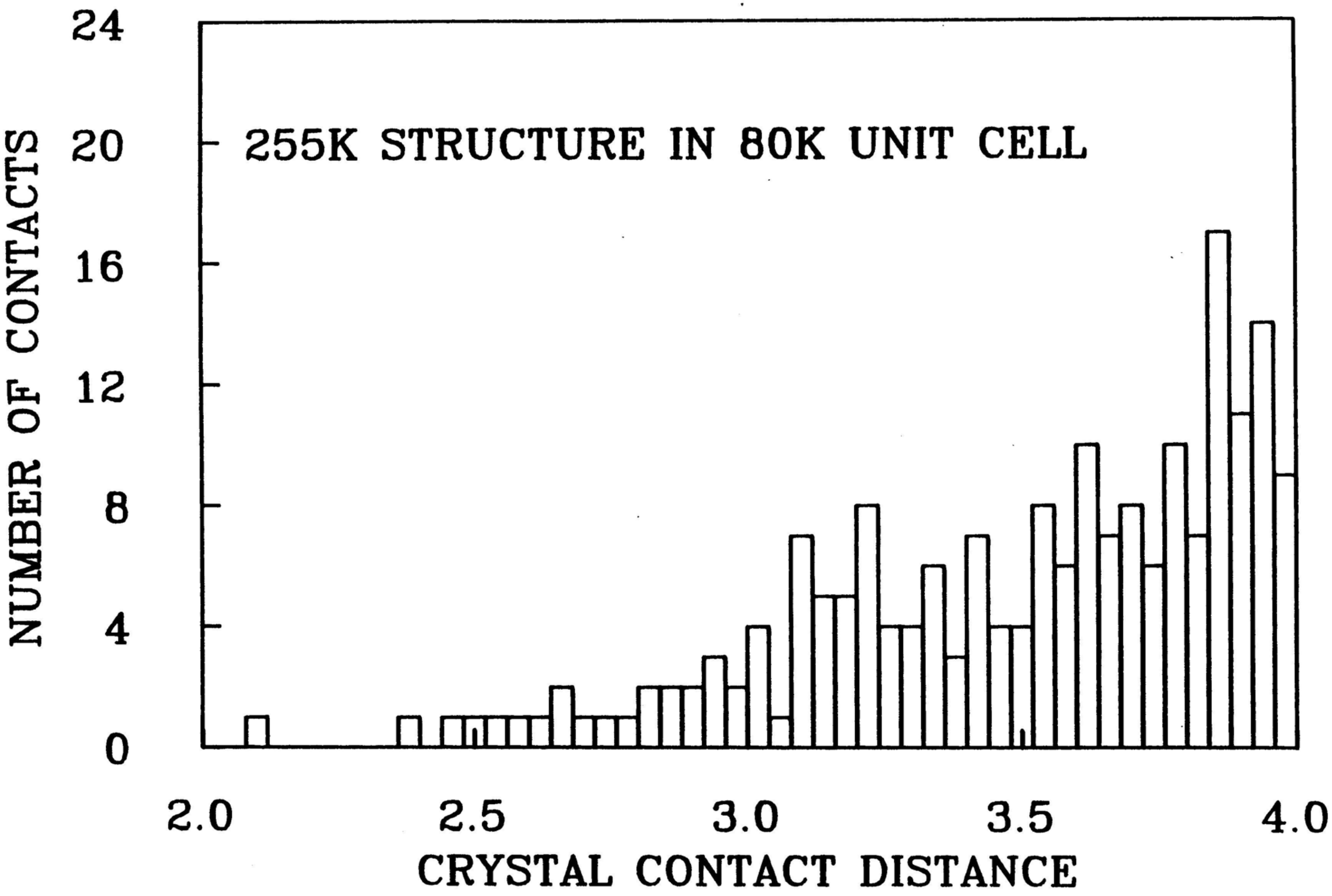

255K STRUCTURE IN 80K UNIT CELL
NUMBER OF CONTACTS
0
4
8
12
16
20
24
2.0
2.5
3.0
3.5
4.0
CRYSTAL CONTACT DISTANCE

FIGURE 2(c)

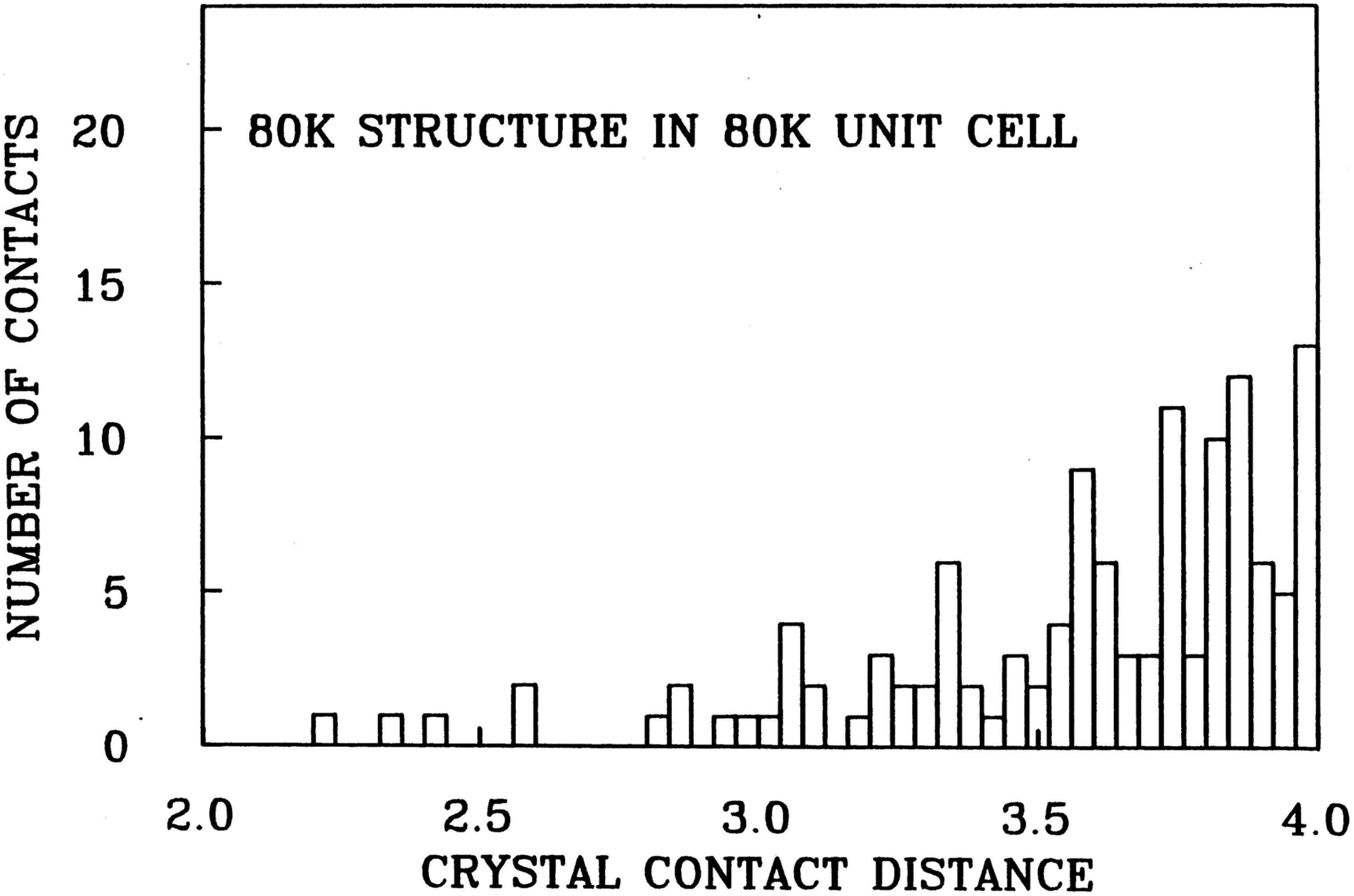

80K STRUCTURE IN 80K UNIT CELL
NUMBER OF CONTACTS
0
5
10
15
20
2.0
2.5
3.0
3.5
4.0
CRYSTAL CONTACT DISTANCE

FIGURE 3

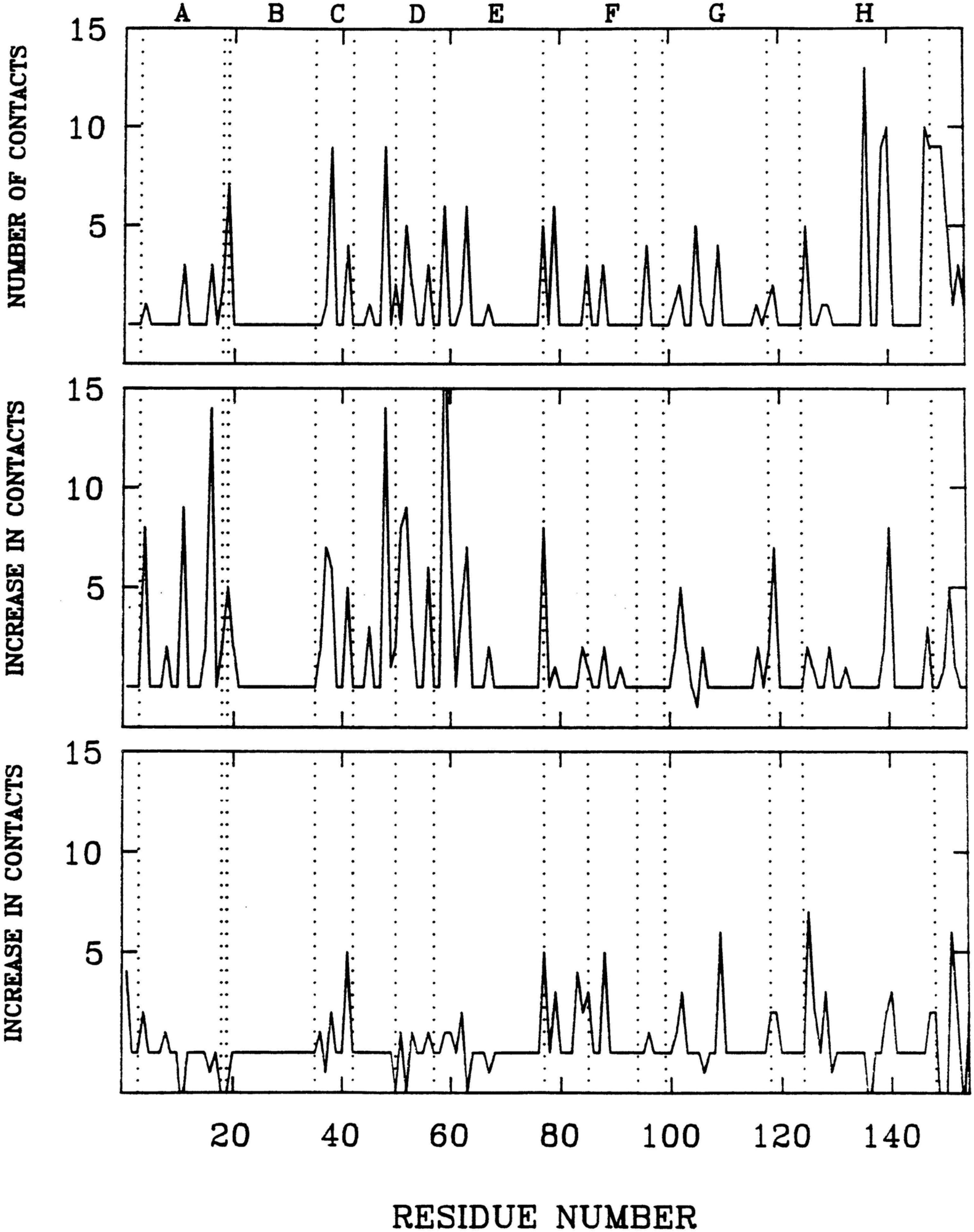

A
B
C
D
E
F
G
H
15
10
5
NUMBER OF CONTACTS
15
10
5
INCREASE IN CONTACTS
15
10
5
INCREASE IN CONTACTS
20
40
60
80
100
120
140
RESIDUE NUMBER

FIGURE 4(a)

INTER–MOLECULAR CONTACTS AT 255K

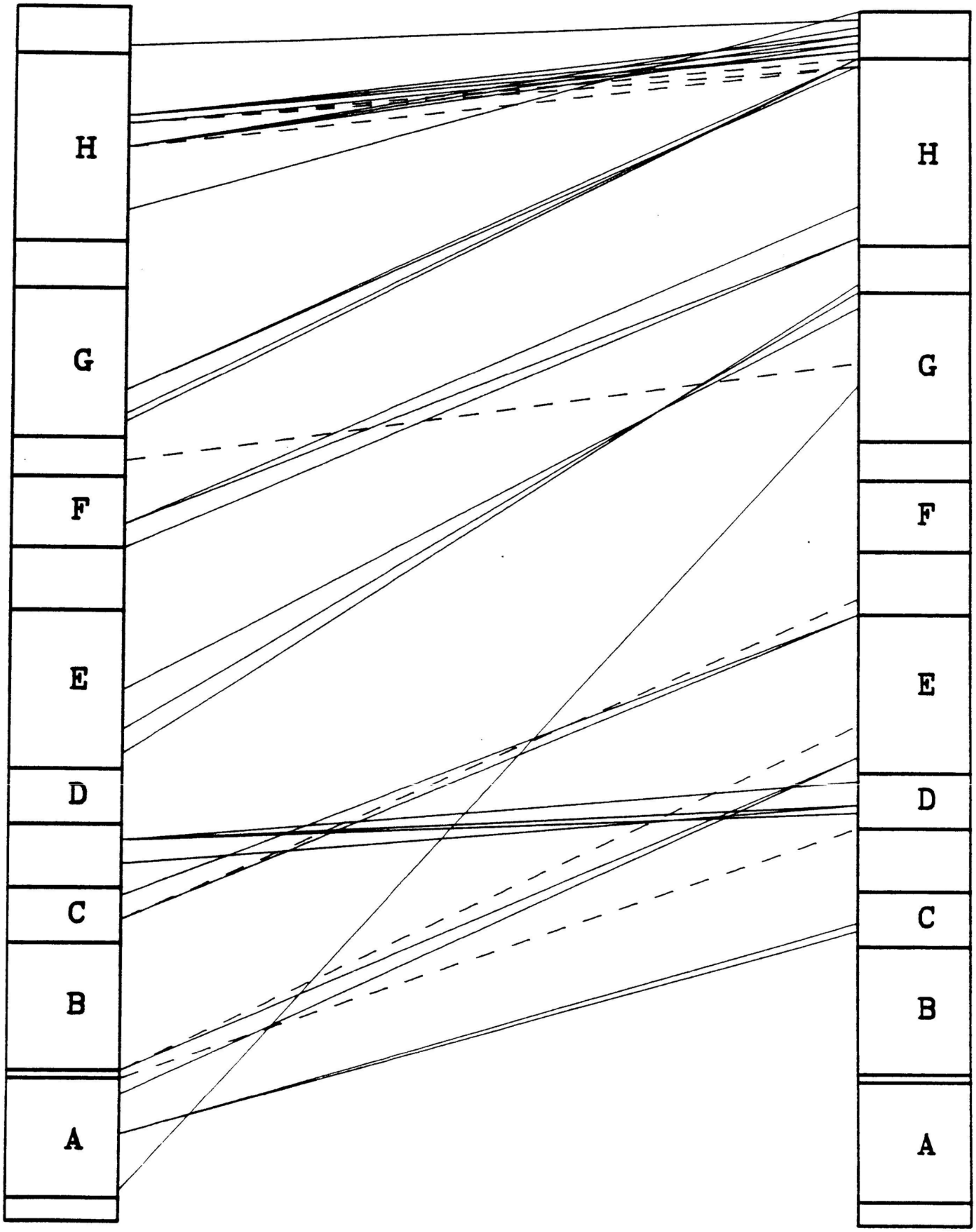


FIGURE 4(b)

INTER-MOLECULAR CONTACTS AT 80K

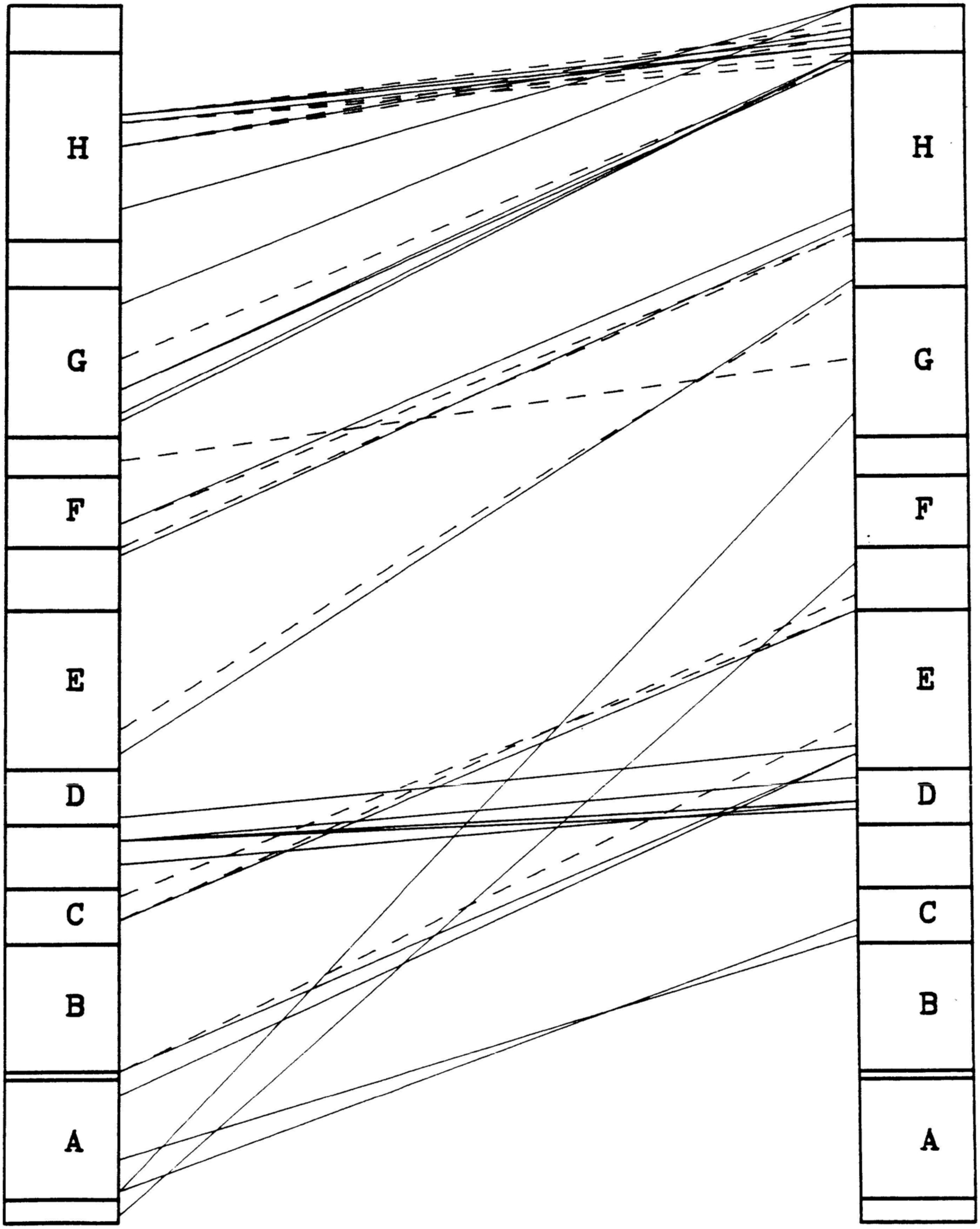


FIGURE 4(c)

INTER-MOLECULAR CONTACTS WHICH ADJUST TO CHANGE IN UNIT CELL

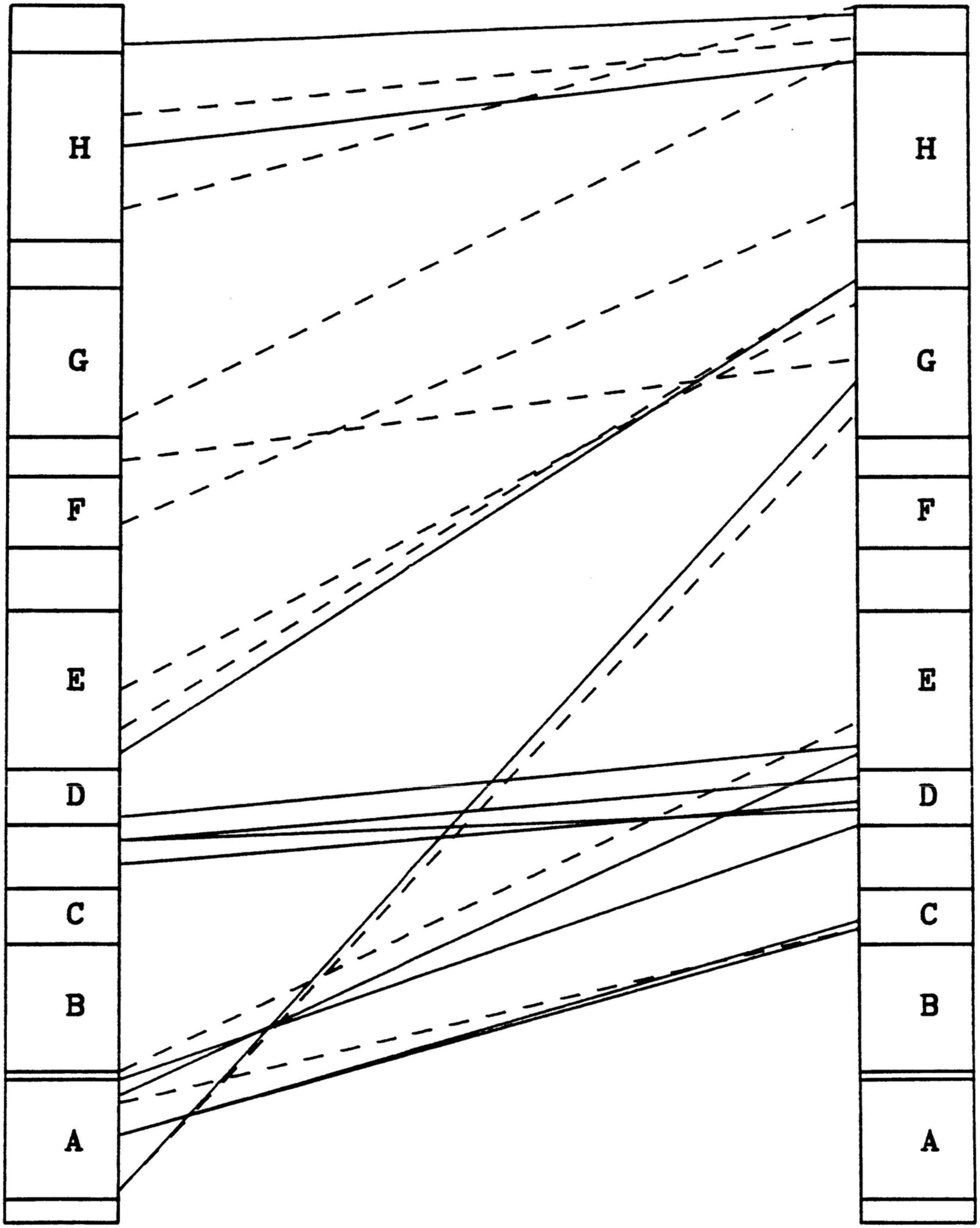


FIGURE 5

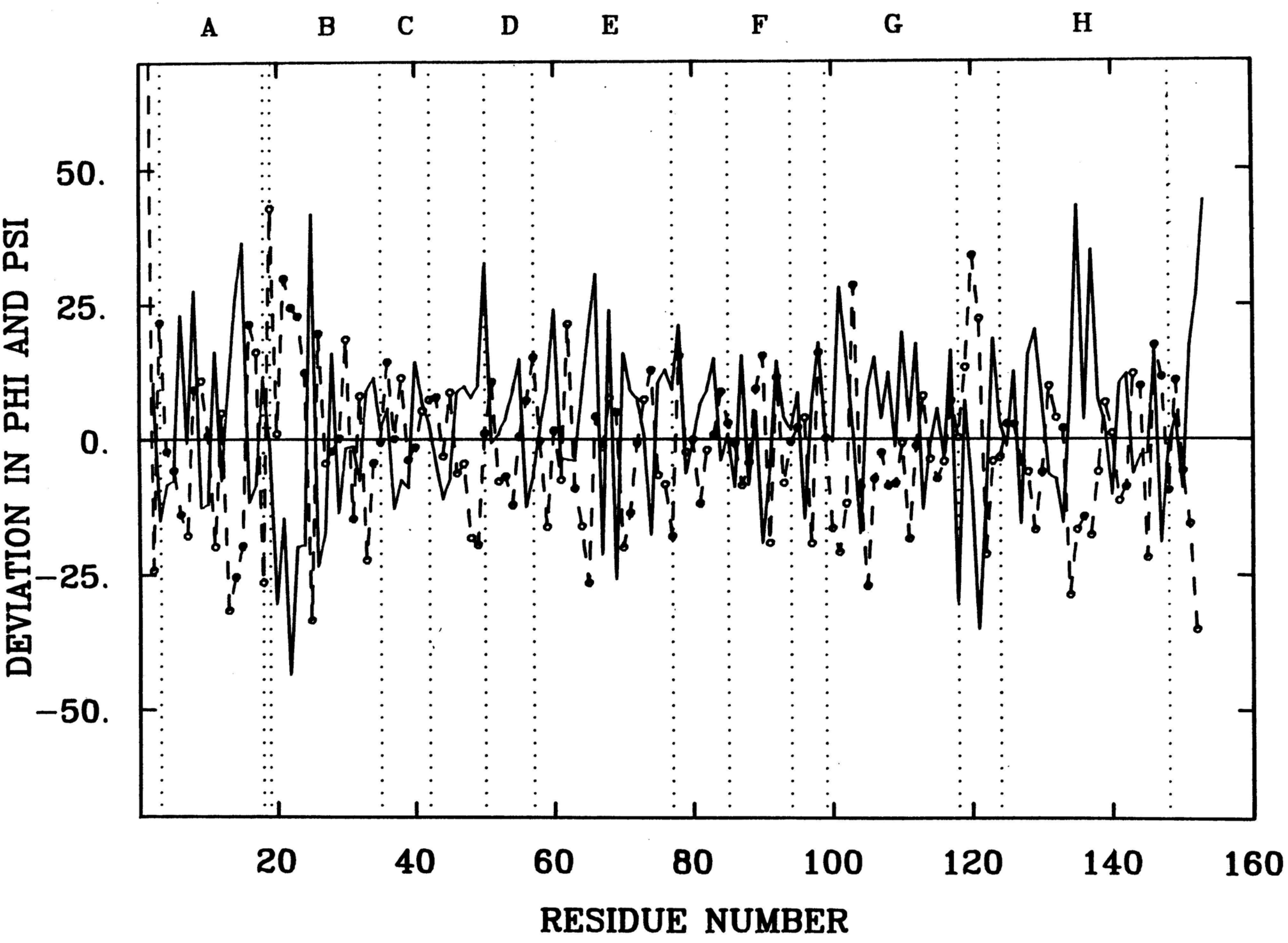
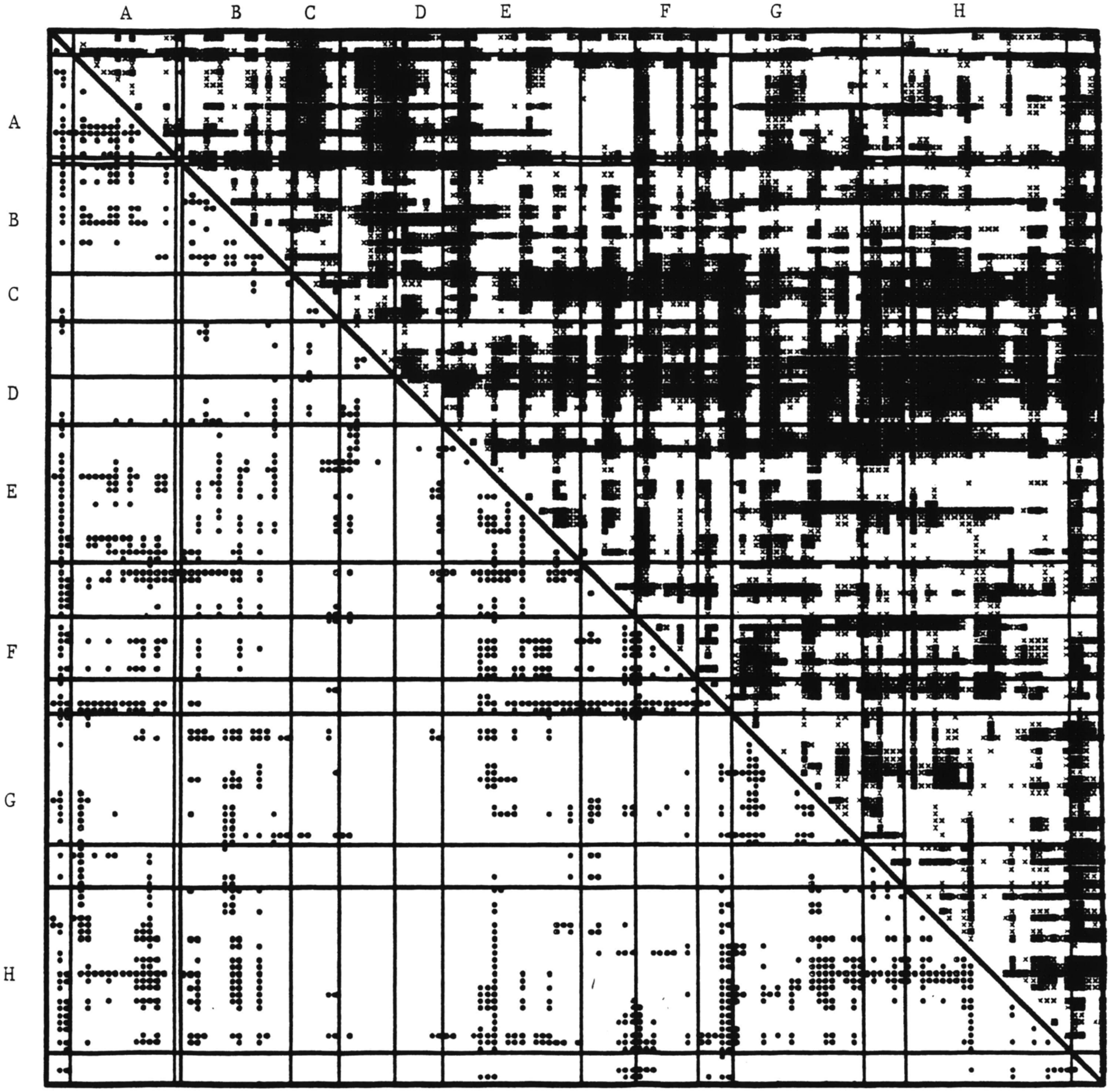
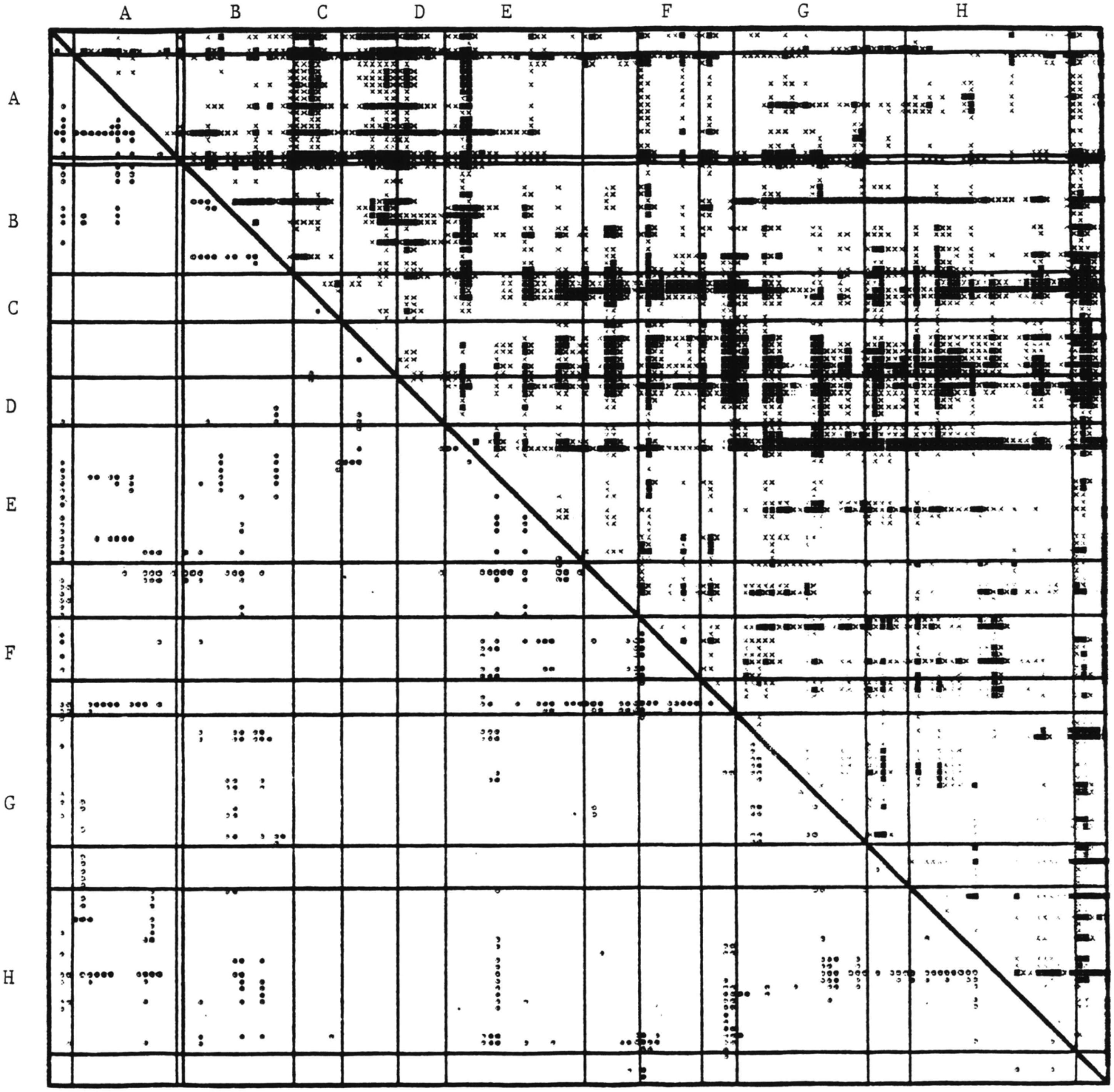

A
B
C
D
E
F
G
H
DEVIATION IN PHI AND PSI
50.
25.
0.
−25.
−50.
20
40
60
80
100
120
140
160
RESIDUE NUMBER

FIGURE 6(a)



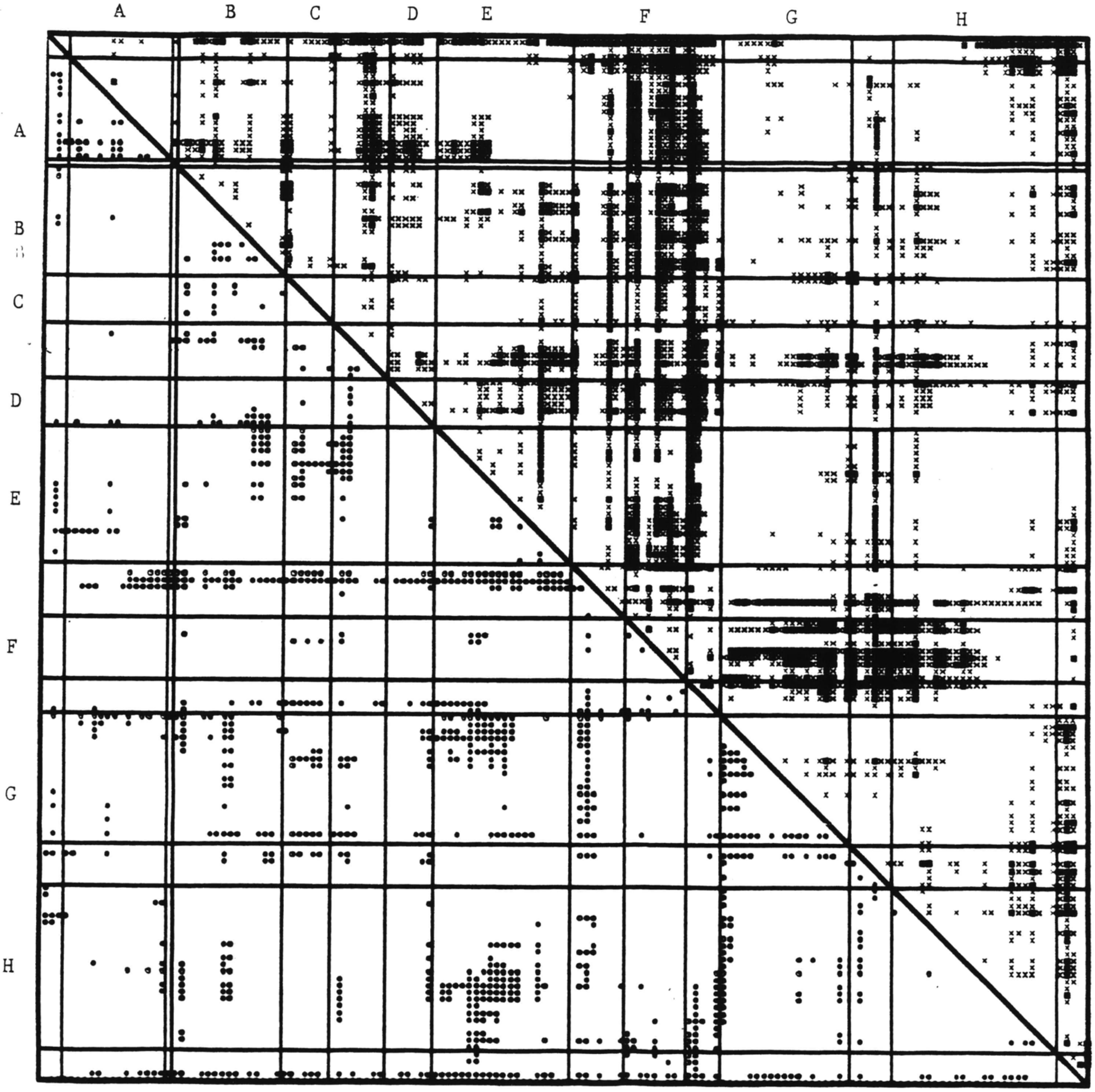
Expansion/Contraction matrix (see Text) between the new 80K structure and the 255K met-Mb structure. The upper half of the matrix corresponds to expansion and the lower half to contraction on going from 80K to 255K. In the lower half, all points represent distances that have decreased by more than 0.15 Å. In the upper half, the lighter points represent expansion by 0.15 Å and the darker points expansion by 0.25 Å. The horizontal and vertical lines demarcate the helix and loop regions.

FIGURE 6(b)



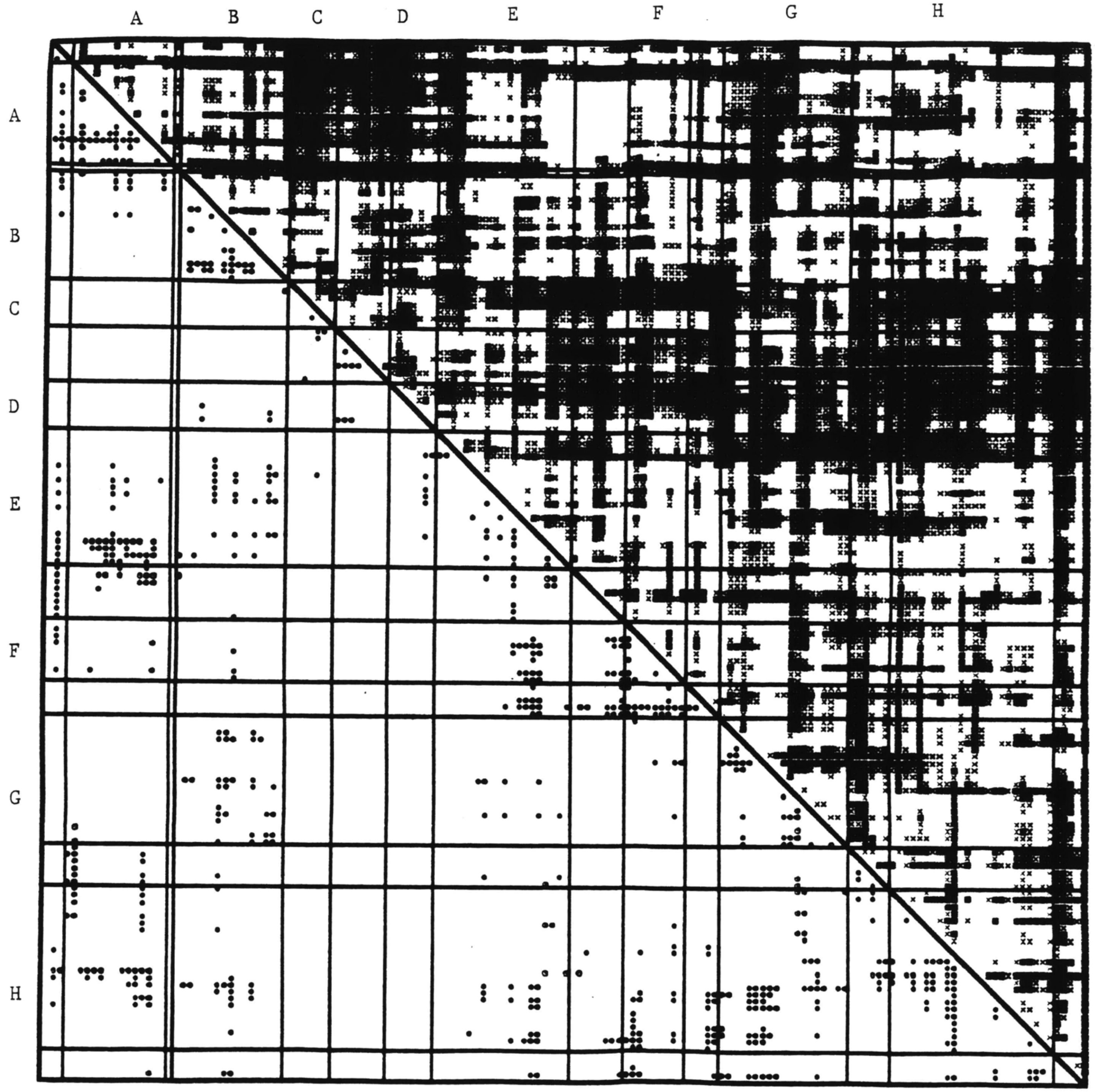
Contraction/Expansion matrix, as in Fig. 6 (a), between the new 80K structure and the 255K met-Mb structure. Levels are at 0.30 and 0.50 Å.

FIGURE 7



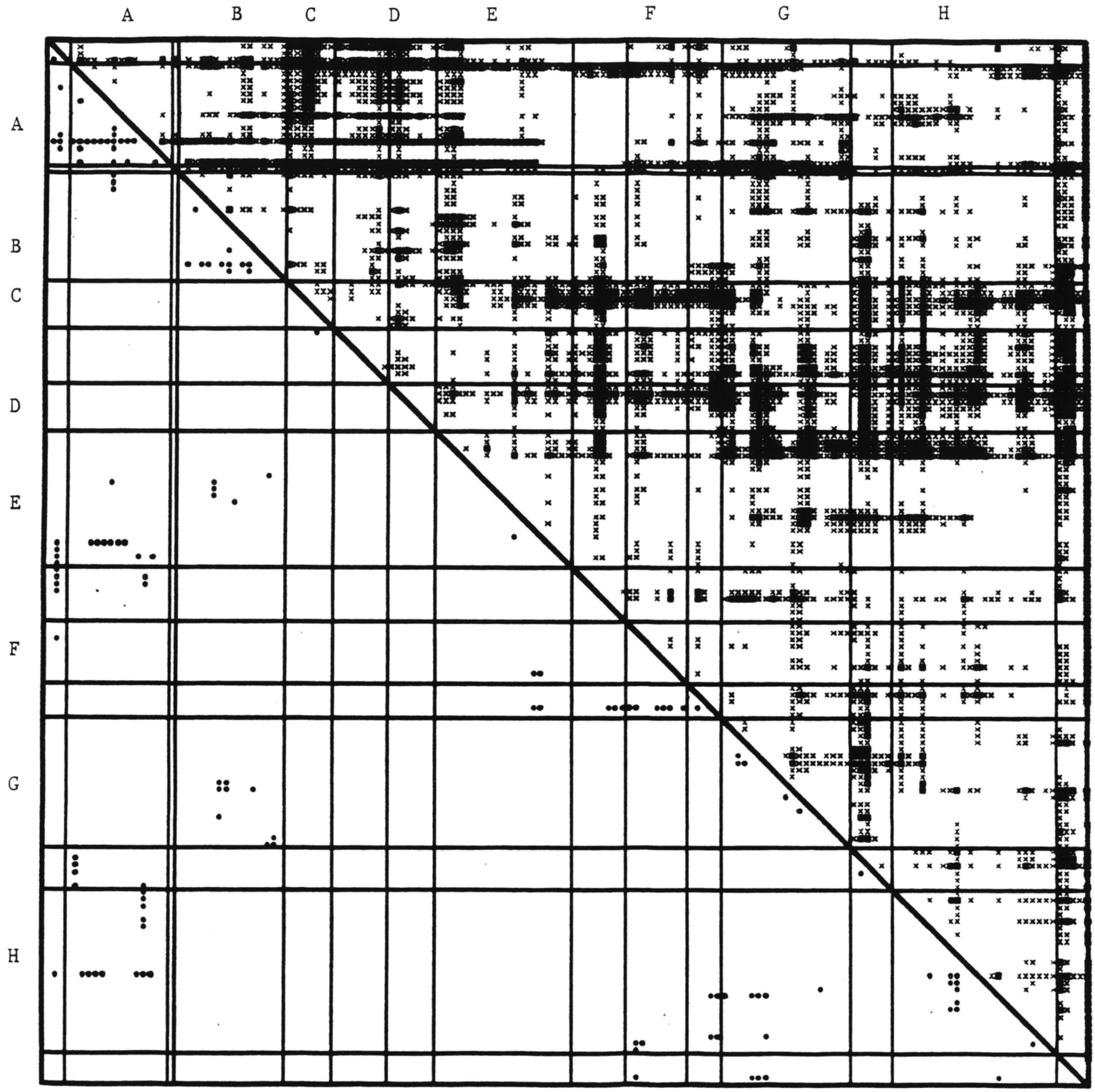
Contraction/Expansion matrix, as in Fig. 6 (a), between CO-Mb and the 255K structure of met-Mb. Levels are at 0.15 and 0.25 Å.

FIGURE 8 (a)



Contraction/Expansion matrix as in Fig. 6(a).
This matrix is the average of 8 matrices comparing two 80K structures and four RT structures. Levels are at 0.15 Å and 0.25 Å.

FIGURE 8 (b)



Contraction/Expansion matrix, as in Fig. 6 (a). This matrix is the average of 8 matrices comparing the two 80K structures and four RT structures. Levels are at 0.3 Å and 0.5 Å.

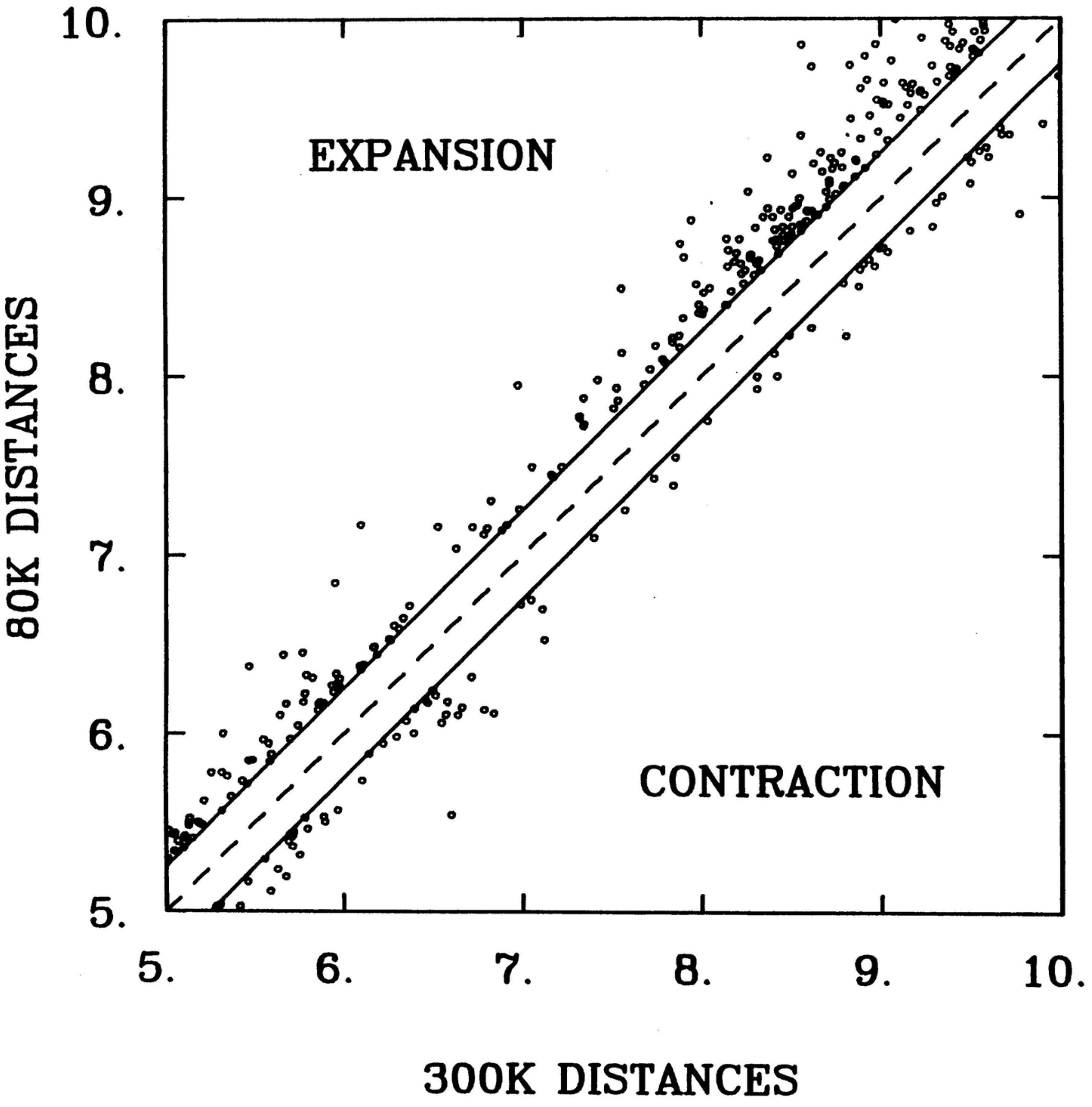
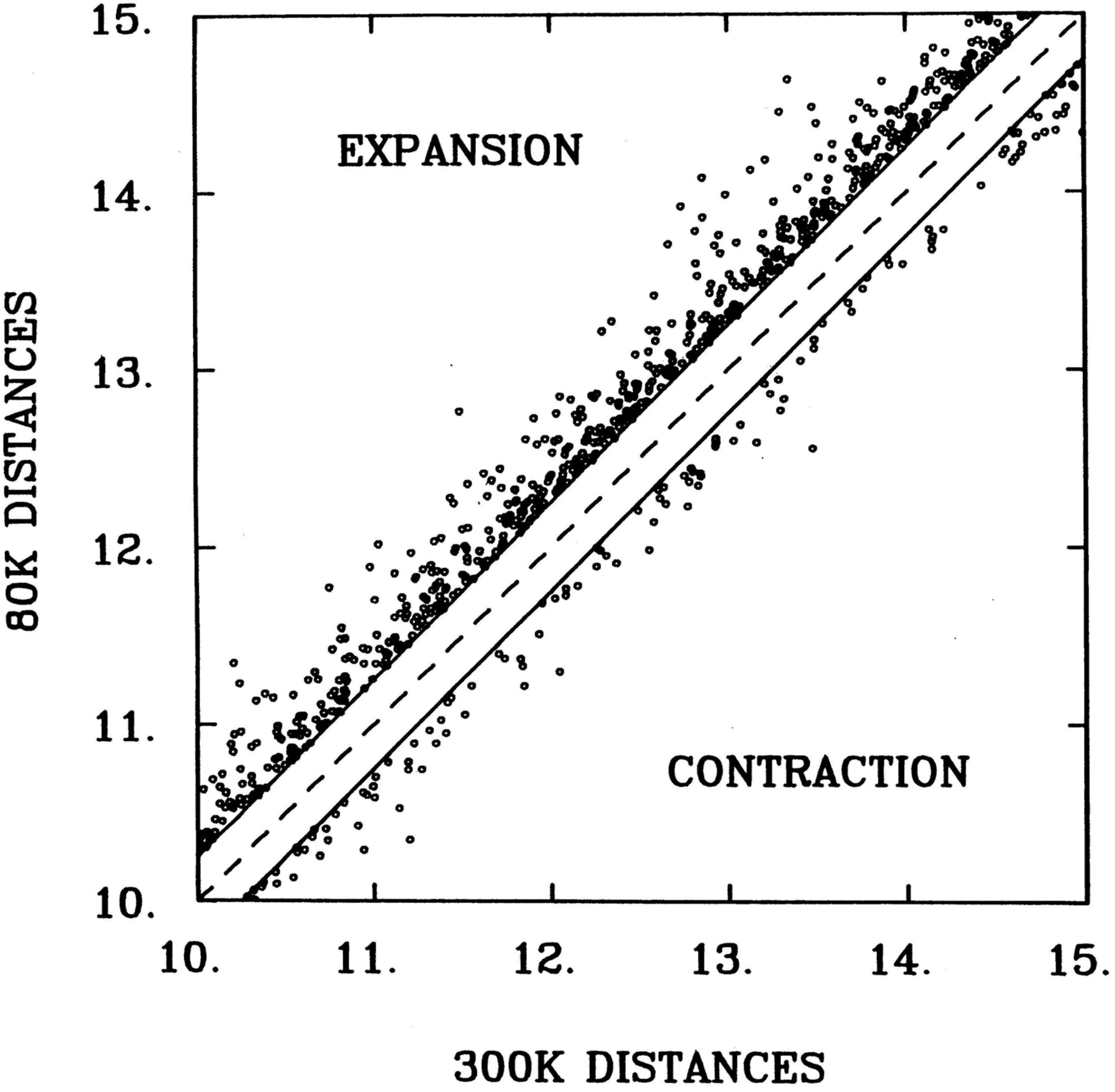

EXPANSION
CONTRACTION
80K DISTANCES
300K DISTANCES
5.
6.
7.
8.
9.
10.
5.
6.
7.
8.
9.
10.

FIGURE 9 (a)

FIGURE 9 (b)


EXPANSION
CONTRACTION
80K DISTANCES
300K DISTANCES
10.
11.
12.
13.
14.
15.
10.
11.
12.
13.
14.
15.

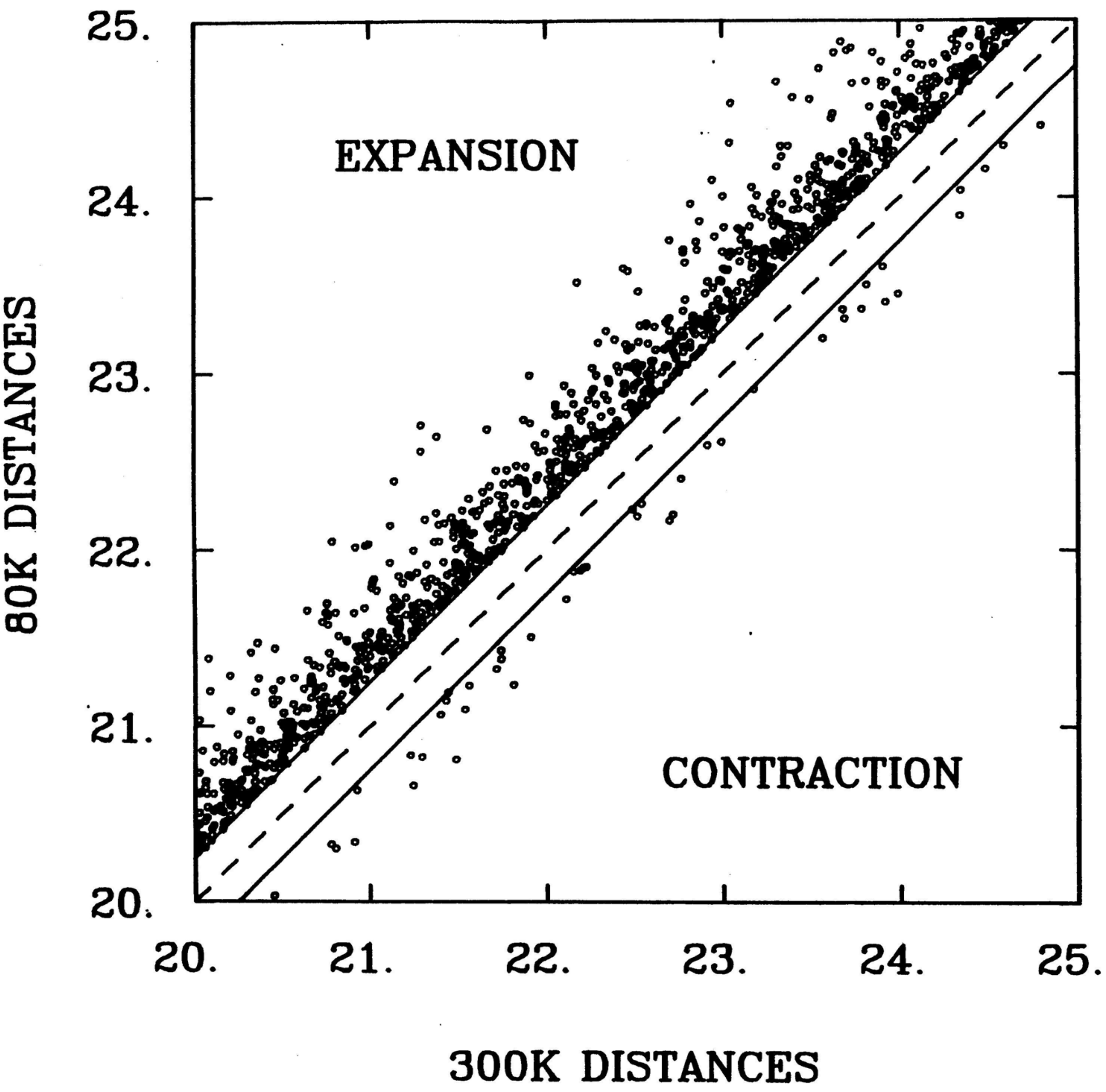

25.
24.
23.
22.
21.
20.
80K DISTANCES
EXPANSION
CONTRACTION
20.
21.
22.
23.
24.
25.
300K DISTANCES

FIGURE 9 (c)

FIGURE 10

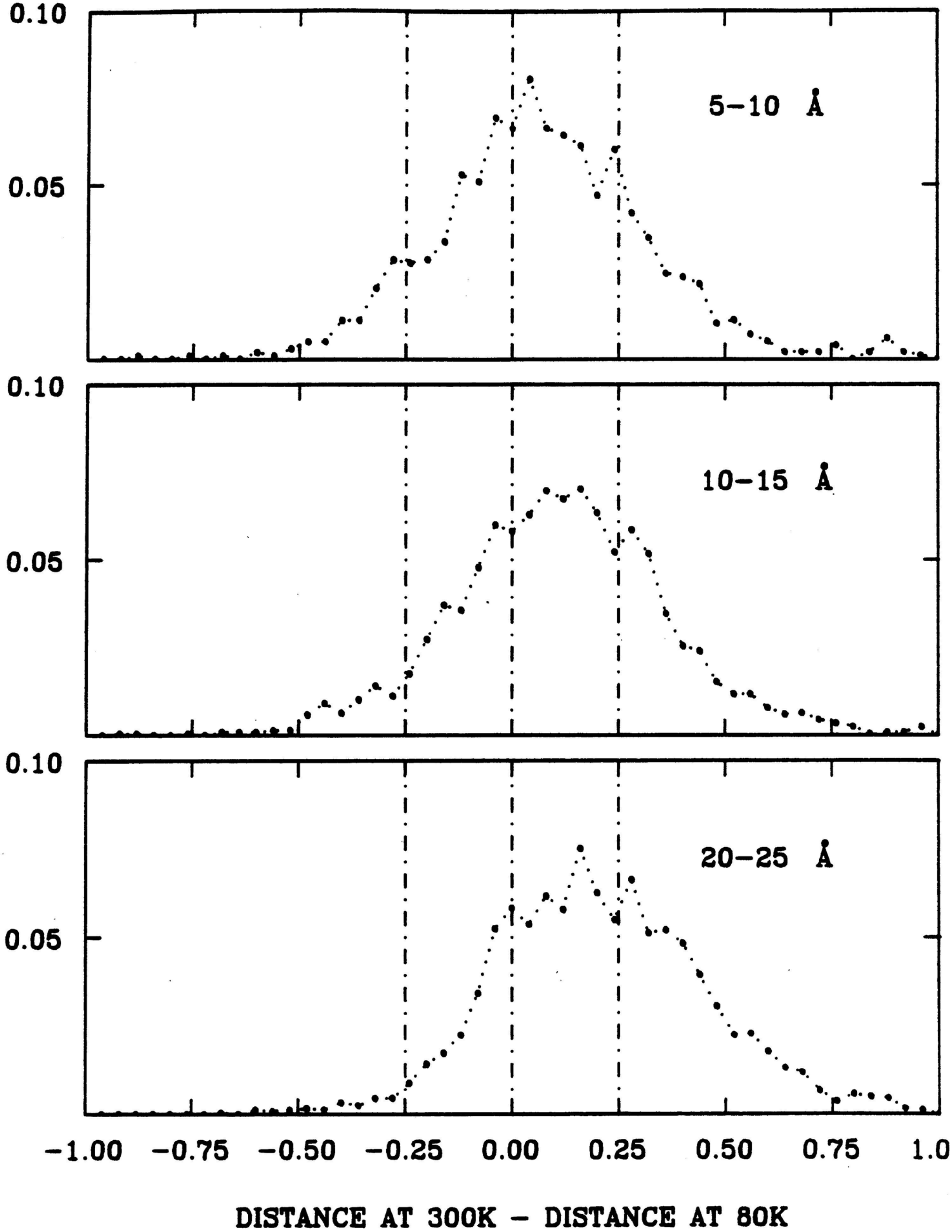

0.10
0.05
5–10 Å
0.10
0.05
10–15 Å
0.10
0.05
20–25 Å
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.0
DISTANCE AT 300K – DISTANCE AT 80K

FIGURE 11 (a)

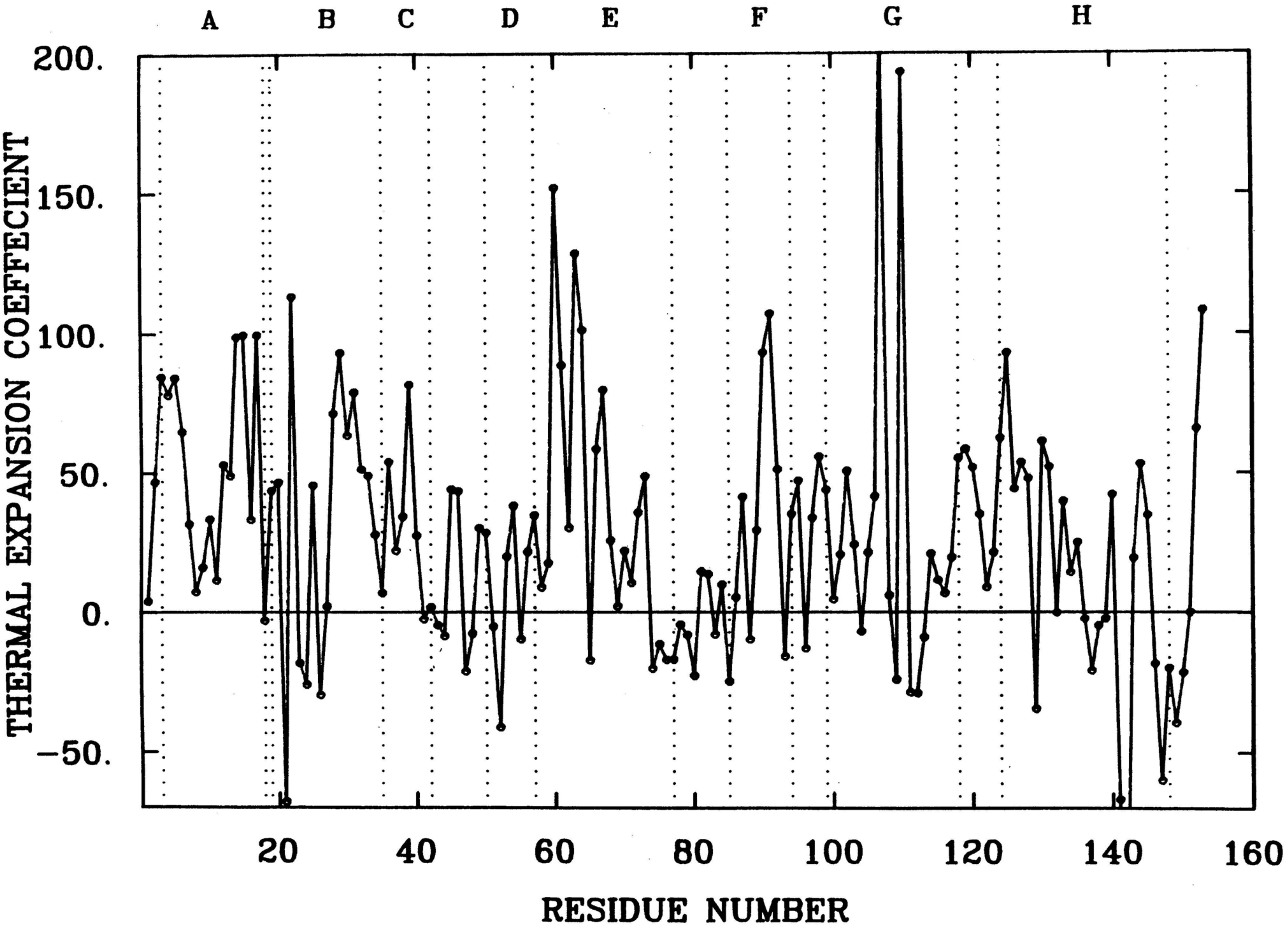

A
B
C
D
E
F
G
H
THERMAL EXPANSION COEFFECIENT
200.
150.
100.
50.
0.
−50.
20
40
60
80
100
120
140
160
RESIDUE NUMBER

FIGURE 11 (b)

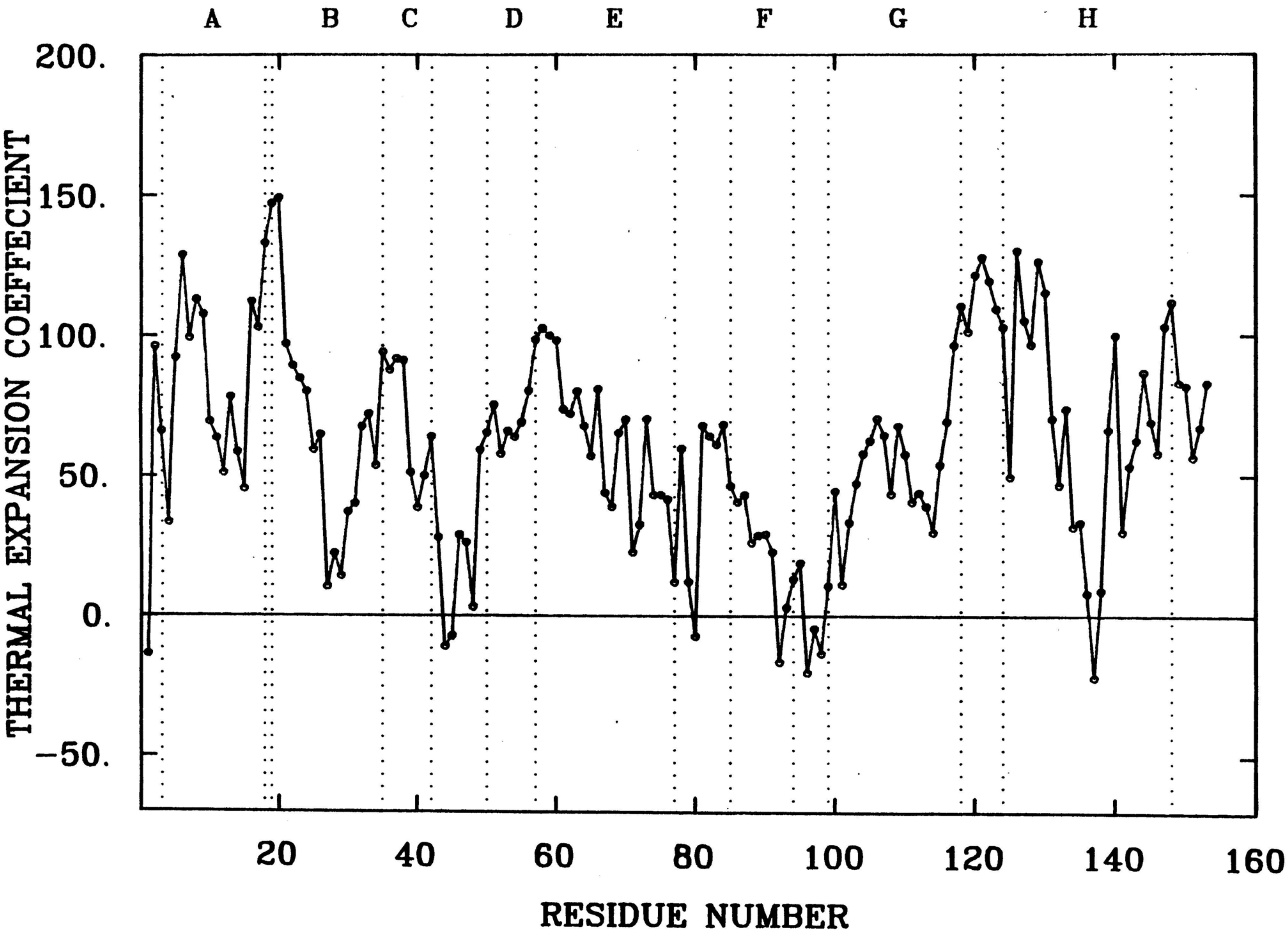
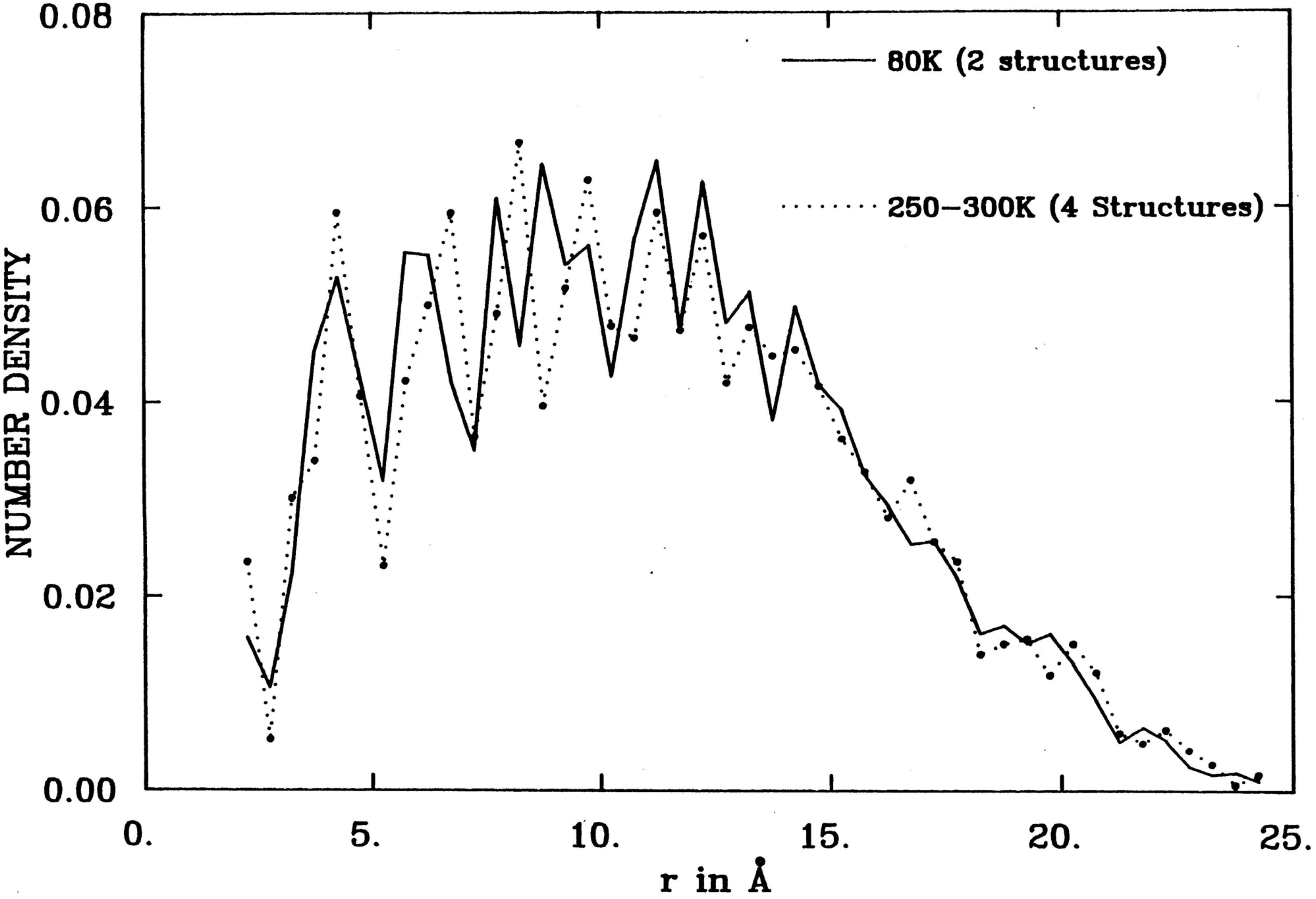

A
B
C
D
E
F
G
H
THERMAL EXPANSION COEFFECIENT
200.
150.
100.
50.
0.
−50.
20
40
60
80
100
120
140
160
RESIDUE NUMBER

FIGURE 12


0.08
0.06
0.04
0.02
0.00
NUMBER DENSITY
0.
5.
10.
15.
20.
25.
r in Å
80K (2 structures)
250–300K (4 Structures)

Chapter 4

The X-ray Structure and Refinement of CO-Myoglobin at 1.5 Å Resolution

Abstract

The structure of CO-myoglobin at 260K has been solved at a resolution of 1.5Å by X-Ray diffraction and a model refined against the X-ray data by restrained least-squares. The CO ligand is disordered and distorted from the linear conformation seen in model compounds. At least two conformations, with Fe-C-O angles of 140° and 120°, are required to model the system. The heme pocket is significantly larger than in deoxy-myoglobin because the distal residues have relaxed around the ligand; the largest displacement occurs for the distal histidine sidechain, which moves more than 1.4Å on ligand binding. The sidechain of Arg 45 (CD3) is disordered and apparently exists in two equally populated conformations. One of these does not block the motion of the distal histidine out of the binding pocket, suggesting a mechanism for ligand entry. The heme group is planar (root-mean-square deviation from planarity is 0.08Å) with no doming of the pyrrole groups. The Fe-$N_{\varepsilon 2}$ (His 93) bond length is 2.2Å and the Fe-C bond length in the CO complex

is 1.9Å. The iron is in the least-squares plane of the heme, and this leads to the proximal histidine moving by 0.4Å relative to its position in deoxy-myoglobin (Takano, T., J. Mol. Biol., 110, 569-84). This shift correlates with a global structural change, with the proximal part of the molecule translated towards the heme plane.

Introduction

Small, well characterized, molecules such as CO_2 and H_2 are model systems for the development of new theoretical and experimental techniques in physical chemistry. Small, well characterized, proteins such as bovine pancreatic trypsin inhibitor and sperm whale myoglobin (Mb)[1] are playing a corresponding role in biophysical chemistry. Myoglobin, in particular, is of great interest because of its known functional importance in the storage of oxygen and the ease of studying the dissociation reaction by photolysis techniques. More than twenty years after the three dimensional structure of the met form of sperm whale myoglobin was first described (Kendrew et al., 1960), it continues to be the subject of crystallographic investigation. New crystal structures of this form (with an S=5/2 ferric iron), as well as of the physiologically important high spin deoxy (S=2, FeII) and low spin oxy (S=0, FeII) forms have been obtained and refined (Takano, 1977a,b, Frauenfelder et al., 1979, Phillips, 1980, 1981). It has become clear that a knowledge of the atomic coordinates at the highest possible accuracy is essential for a functional description of the molecule. Moreover, it is only from care-

[1] Abbreviations used: Mb, myoglobin; CO-Mb, carbon-monoxy (Fe II) myoglobin; Oxy-Mb, O_2 (Fe II) myoglobin; met-Mb, OH_2 (Fe III) myoglobin; Hb, Hemoglobin.

fully refined high resolution structures that reliable information concerning the atomic fluctuations can be derived.

Because of the instability of the oxy form in aqueous solution at normal pH, a number of spectroscopic and kinetic studies have relied on the more stable carbon-monoxy (CO) form of the molecule as a model for the liganded, low spin, Fe II state of the protein (see,for example, Austin et al, 1975, Alben, 1978, Henry et al., 1983, Powers, et al., 1984). Such studies have raised questions about the geometry of the CO ligand when bound to the iron in the protein. There is evidence that the CO ligand binds to the protein in more than one conformation, (McCoy and Caughey, 1971, Churg et al., 1978, Makinen, et al., 1979, Brown et al., 1983) and one goal of a crystallographic structure determination is to delineate the possible stable structures of the ligand. It is also important to examine how the structure of CO-Mb differs from that of model compounds and the other liganded forms. Such information can provide a structural basis for the spectroscopic differences observed between the different liganded form. Moreover, it is important for understanding how ligand binding results in tertiary structural changes. Such tertiary structural changes are thought to be involved in the cooperativity of ligand binding in hemoglobin (Gelin and Karplus, 1977,Baldwin and Chothia, 1979, Gelin et al., 1983). Myoglobin serves as a simpler, non-cooperative, system for study and for comparison with the structural results for hemoglobin.

In the development of theoretical approaches to proteins, the availability of good structural data is important for testing simulation methods such as energy minimization and molecular dynamics. Accurate structures can also aid in the parametrization of the force fields used in such simulations. The magnitude of atomic fluctuations in the protein can be deduced from the crystallographic temperature factors (Willis and Pryor, 1975, Petsko and Ringe, 1984) and can be correlated with the fluctuations calculated from molecular dynamics (Karplus, 1981, Karplus and McCammon, 1983, Levy et al., 1985) and normal mode simulations of the protein (Brooks and Karplus, 1983, Go et al., 1983, Levitt et al., 1985). There is a fundamental interest in the structure and dynamics of myoglobin because migration of the ligand from the solvent to the buried binding site in the heme pocket, and subsequent binding to the iron, is the best studied example of a reaction involving potential barriers in the protein matrix (Austin et al., 1975, Case and Karplus, 1978,1979, Henry et al., 1983, Ansari et al., 1985).

A need clearly exists for a refined structure of CO-Mb that is comparable in accuracy to that of the available met , deoxy and oxy forms. The structure of CO-Mb has been solved previously by neutron diffraction, at a resolution of 1.8Å (Norvell et al., 1975, Hanson and Schoenborn, 1981), and refined by a real space constrained method (Diamond, 1971). The ability to locate hydrogen and deuterium atoms from the neutron data has provided a great deal of information about the protonation states of the titratable groups, the location of slowly exchanging protons and details about the solvent and protein hydrogen bonding. Unfortunately, apparent disorder at the ligand binding site prevented the unequivocal determination of the ligand geometry (Hanson and Schoenborn, 1981). Also, the use of real-space refinement makes it difficult to compare the results with those for met- and oxy-Mb which were refined in reciprocal-space.

We report here the determination of the structure of CO-Mb at a resolution of 1.5Å, by X-ray diffraction. The structure determination was done at low temperature (260K) to aid in characterizing the possible disordered ligand conformations. Refinement against the X-ray data was done by a restrained least-squares method (Konnert and Hendrickson, 1980) in which all internal degrees of freedom of the molecule are allowed to vary, subject to stereochemical and other restraints. Although a completely unrestrained structure determination is not possible at this resolution, the Konnert-Hendrickson method is expected to introduce less bias than constrained refinement, where the number of internal degrees of freedom are greatly reduced. Also, this method allows the assignment of individual isotropic atomic temperature factors (B-factors) which are more reliable than those obtained by real-space refinement.

This paper focuses on three aspects of the X-ray structure of CO-Mb. The conversion of met-Mb crystals to CO-Mb, the data collection and the refinement are described in Section I. The structure of the heme-CO complex and the local effects of CO binding are discussed in Section II. The more global changes in the tertiary structure observed on CO binding are examined in Section III. In the last two sections much of the analysis is based on comparisons of the X-ray structure of CO-Mb with that of deoxy-Mb (Takano, 1977b, Phillips, 1981), the coordinates of which were obtained from the Brookhaven Protein Data Bank (Bernstein et al., 1977). The conclusions are summarized in Section IV.

I Methods

i) Crystal Preparation

Sperm whale met-myoglobin (Sigma Chemical Company) was crystallized in the normal monoclinic form by the usual methods (Kendrew and Parrish, 1956). The crystals were stored in 75% saturated ammonium sulphate, pH 6.0 with 0.1 M phosphate buffer (hereafter referred to as the mother liquor). To convert these crystals to those of the carbon-monoxy derivative of the protein, they were suspended in 9 ml of the mother liquor in a round bottomed flask equipped with a side-arm and stopcock. The top of the flask was sealed with a rubber gasket. The atmosphere inside the flask was deoxygenated by flushing with nitrogen for one hour at room temperature. The nitrogen was then replaced by carbon-monoxide and the crystals were equilibrated overnight at room temperature. After twelve hours, the flask was brought to 4°C and a syringe was used to inject 1 ml of a 20 mg/ml deoxygenated, CO-equilibrated, solution of sodium dithionite through the rubber gasket into the mother liquor. Dithionite was introduced slowly, over a two hour period, to a final concentration of 2 mg/ml. During this time the colour of the smaller, thin crystals was observed to change from reddish brown to bright red. Following a second overnight incubation with CO at room temperature, a crystal of size comparable to that used for data collection was removed from the flask under a nitrogen atmosphere in a glove bag and dissolved in 0.1 M phosphate buffer, pH 6.0. Ultraviolet and visible absorption spectra were run immediately. The characteristic bands of CO-Mb (Hanania et al., 1966) were present and there was little indication of contamination from the met form (less than 2%).

A large flawless crystal was then removed from the flask under nitrogen in a glove bag that had been sealed around a stereo microscope. The crystal was mounted in a 1mm quartz capillary tube and a column of machine oil was inserted at either end of the tube to prevent contamination of the nitrogen atmosphere by room air if the seals leaked. Sealing of the tube was done with five minute epoxy; the assembly was left in the glove bag under nitrogen for three hours to insure proper hardening of the resin. The capillary tube was then mounted on a conventional goniometer head. On completion of the data collection the crystal was dissolved in pH 6.0 phosphate buffer; the UV and visible spectra (Hanania et al.,1966) revealed that no significant conversion to the met form had occurred during data collection.

ii) Data Collection

Data were measured on a Nicolet P3 diffractometer equipped with a modified LT-1 low temperature device. All measurements were made at 260K. Low temperature was used to reduce radiation damage (Petsko, 1975) and to minimize X-ray induced conversion of CO-Mb to the met form. It was decided not to employ a cryoprotective mother liquor, which would have permitted a lower temperature to be reached, since the use of one would have complicated comparisons of the structure with that of met-, oxy- and deoxy-Mb, all of which were solved in their normal aqueous mother liquor at temperatures between 255K and 300K. Since reflection profiles indicated that peak shape was uniform throughout reciprocal space, Wyckoff scans were used for reflection measurements (Wyckoff et al., 1967). A moving omega-step-scan of 11 steps of 0.03° each was taken across the top of the reflection profile; the highest 7 consecutive steps were summed as the peak intensity. The background correction was estimated from a curve of background vs. ϕ and 2θ, measured after the completion of the entire data set. The 2θ dependance was typical (Wyckoff, et al., 1967). The ϕ dependance matched that of the empirical absorption curve (North et al., 1968), so the 2θ curve was measured at the ϕ angle of maximum transmission, and the absorption correction was applied to the data prior to the background correction. Scan rates were selected so that the data acquisition rates averaged 2200 reflections per day.

Radiation damage was monitored from repeated measurement (every 300 reflections) of the intensities of five strong reflections chosen to be well distributed over reciprocal space. Decay of these reflections showed the radiation damage to be linear in time and isotropic. The total damage was observed to be 16% after 120 hours of data collection. We chose to collect the entire 1.5Å data set on one crystal rather than merge data sets collected on different crystals, thus avoiding errors caused by variation in the unit cell parameters and static disorder from crystal to crystal (Kuriyan, unpublished) at the expense of increased radiation damage. This is in contrast to other workers who have minimized the effects of radiation damage, usually by rejecting crystals with more than 10% radiation damage; for example, nine crystals were used to obtain the 1.6Å data set for oxy-Mb (Phillips, 1980), seventeen for the 2.0Å data set for the cyano derivative of a monomeric lamprey Hb (Honzatko et al.,1985) and five for the 1.4Å data set for CO derivative of the momomeric insect hemoglogin, erythrocruorin (Steigemann and Weber, 1979).

iii) Data Reduction

A total of 18520 measurements were made from a single crystal of CO-myoglobin; the data set covered a range of 2θ from 2° to 60°. These data were corrected for radiation damage (a linear decay model was used), absorption (North et al., 1968), background, and Lorentz polarization effects. Measurements where the reflection intensity, prior to any correction, was less than twice its estimated standard deviation were rejected as unreliable. The final reduced data set consisted of 10448 unique reflections from 10Å to 1.5Å resolution, corresponding to ??% of the total.

iv) Refinement

The refinement method used is a restrained-parameter least-squares method implemented in the program PROLSQ (Konnert, 1976, Konnert and Hendrickson, 1980). Three positional coordinates and one isotropic temperature factor (Willis and Pryor, 1975) were refined for every non-hydrogen atom in the protein. The refinement improves the agreement between experimentally observed structure factors, $F_o(h,k,l)$, and those calculated from the model, $F_c(h,k,l)$, subject to stereochemical, thermal factor and shift restraints (Konnert and Hendrickson, 1980). The various restraints and their associated weights are given in Table 1. The agreement between the model and the data is monitored by the crystallographic R-factor, R:

$$R = \frac{\sum_{h,k,l} \left| |F_o| - |F_c| \right|}{\sum_{h,k,l} |F_o|}$$

The initial model used was the met-Mb structure at 300K (Frauenfelder et al.,1979). A difference Fourier map using coefficients $(F_o-F_c)e^{2\pi i\alpha_c}$ (Blundell and Johnson, 1976) was calculated, where the structure factors, F_c, and phases, α_c, were obtained from the the met-myoglobin structure with the water bound to the iron deleted. The difference map clearly showed the CO ligand in the distal pocket. The density for the oxygen of the CO could not be fit by assuming a unique conformation for the ligand, and so it was modelled by two alternate positions for the oxygen atom. The Fe-C and C-O distances were restrained in the refinement to 1.85Å and 1.2Å[1], respectively. No other stereochemical restraints were placed on the ligand. The iron was not restrained to be in any plane, though the iron - pyrrole nitrogen distances were restrained to 2.01Å.[2] The four pyrrole groups were each separately restrained to be planar. No restraints were placed on the heme proximal-histidine linkage, nor on the overall planarity of the heme group.

Refinement against structure factor data between 10.0Å and 1.5Å was started by including only protein and heme atoms in the model. The refinement was started with loose stereochemical restraints and several cycles of refinement dropped the R-factor to 20.8%, from the initial value of 30.5%. Difference Fourier maps with coefficients (F_o-F_c) and $(2F_o-F_c)$ and phases from the refined model were calculated and examined

[1] These values, used in the restraint dictionary for refinement, are slightly in error. It was intended that the restraints on the Fe-C and C-O bonds be set to the values of 1.78Å and 1.12Å found in a model compound (Peng and Ibers, 1976).
The Fe - N (pyrrole) distance in a model compound (Peng and Ibers, 1976) is 2.02±0.03Å.

on a PS300 graphics system using the FRODO software (Jones, 1982). The strategy employed in examining difference Fourier maps was to first examine a $2F_o-F_c$ map on a residue by residue basis. Residues that seemed to be in doubt in this map were removed from the model and an F_o-F_c map calculated from the partial model, thus removing any bias in the phases for that region. Segments as large as ten residues could be removed from the model and useful information obtained from the difference Fourier map. A few sections of the protein needed rebuilding and many solvent molecules were placed on the basis of the difference maps. Several residues were seen to exist in more than one conformation and in seven of them the density was clear enough to warrant building in two conformations for the sidechain. The modelling of the disordered residues is discussed in more detail below.

Refinement was continued by alternating several cycles of least-squares refinement with examination of difference Fourier maps and manual rebuilding of sidechains or solvent molecules on the graphics system. After three such iterations the refinement was stopped at a stage where the difference maps within the protein boundary had no peaks above $0.2e/\mathring{A}^3$ and least-squares refinement resulted in no further drop in the R-factor, which had converged to a final value of 17.1%, with good stereochemistry. The final model includes 136 solvent molecules (presumably waters) and one sulphate ion, which is bound at the end of the E helix as in other myoglobin structures (Phillips, 1980). No sulphate ion is seen near the distal histidine, unlike met- or deoxymyoglobin. The model for the solvent could be improved since no variable occupancies were refined at any stage for the solvent and no solvent correction (Phillips, 1980) was added to the calculated structure factors. No hydrogen atom positions were refined, even though this has been shown to improve the accuracy of the internal coordinates (Phillips, 1980).

One problem with restrained refinements such as this is that it is difficult to decide the relative weights to be assigned to the restraints and the structure factor data. The usual goal is to reduce deviations of the model from ideal stereochemistry as much as possible. However, atomic motion causes the time averaged structure to exhibit deviant stereochemistry (Karplus, 1981, Kuriyan et al., 1985). If the restraints on stereochemistry are very loose, the refinement tends to move well determined atoms that are near regions of high mobility away from their true positions. The use of tight stereochemical restraints prevents this and has been shown to yield better structural results than loose restraints (Kuriyan et al.,1985).

The statistics of the restrained parameters in the final model are shown in Table 1. The rms deviation from ideality for angles, planar 1-4 distances, chirality and sidechain temperature-factors are slightly higher than the target values, while those for bonds, planar groups and mainchain temperature-factors are less than or equal to the target values. Continuing the refinement for a few more cycles with higher weights on the restraints results in a structure in which the rms deviations from ideality for all classes of parameters are less than the target values, but the R-factor increases to 18%. The overall shift in atomic position is small (less than 0.05Å rms), and the significance, if any, in the difference between the two structures is not clear. We base all subsequent analysis on the structure with lower R-factor (17.1 %).

v) Modelling sidechain disorder

Seven sidechains and the CO ligand were modelled with two distinct conformations. The refinements were done using a modified version of PROLSQ, which allows the occupancies of the disordered sidechains to vary, but refines only one variable occupancy, Q, per residue (Kuriyan, unpublished). The sidechains were modelled as being disordered only beyond the C_β atom, and all the atoms in a particular conformation had the same variable occupancy, Q, and the occupancies of both conformations were constrained to sum to unity; i.e. one conformation had the occupancy Q and the other (1-Q).

Because of the coupling between occupancy and temperature factors, they are difficult to refine simultaneously (Watenpaugh et al., 1980). In an attempt to minimize this problem, shifts for the occupancy and the temperature factors were applied in alternate cycles, as suggested by Hendrickson (1980). Nevertheless, the refined occupancies and temperature factors for disordered groups are approximate. Test refinements of computer generated "perfect" data has shown that the B-factors and the occupancies compensate for errors in each other. The calculations indicate that the occupancies and B-factors for disordered groups are correct, at best, to within 15% and 10%, respectively (Kuriyan, unpublished).

The bonds, angles and planar groups in the two conformations for the disordered residues were restrained in the usual way, and no non-bonded contact restraints were allowed between the atoms in the same sidechain. Torsional restraints were applied to only one of the conformations. This was necessitated by the fact that the program allowed disorder only from the C_γ atom outwards. A few of the alternate conformations required small changes in the backbone positions as well and the lack of flexibility in the backbone resulted in some sidechain dihedrals taking on unrealistic values (see Table 2 below).

The seven residues that were modelled with two alternate conformations for the sidechains are Arg 45 (CD3)[1], Leu 61 (E4), Ile 75 (E18), Gln 91 (F6), Lys 96 (FG2), Met 131 (H7) and Lys 147 (H23). The average B-factors, refined occupancies, and torsional angles for the disordered parts of the sidechains are given in Table 2, below. One of the conformations for Met 131 (H8) has an occupancy of 14%, which is close to estimated noise level of the density. The significance of this conformation is therefore questionable, but it has been retained in the model because the difference electron density indicates that the residue is disordered to some extent.

(vi) Refinement of Arg 45 (CD3)

Since the possible consequences of the disorder of Arg 45 (CD3) form the subject of discussion below, we describe in detail the modelling of this residue. The two refined conformations of this residue, as well as the electron density in a difference map (calculated by omitting the residue from the model) are shown in Fig. 1. [In this, and all the maps that follow, the electron density is given on an absolute scale of

[1] The alpha-numeric codes refer to the position of the residue within the helices and loops (Dickerson and Geis, 1983); see Table 7.

electrons/Å^3 (e/Å^3) by scaling F_o so that $\langle F_o \rangle = \langle F_c \rangle$. The F_c's are calculated on an absolute scale.]

At the contour level of 4σ (0.4e/Å^3), shown in Fig. 1(a), there is no density for the atoms beyond the C_δ atom, an indication that the residue is disordered. At a contour level of 0.3e/Å^3 there are several peaks near the C_δ atom. These can be fit by modelling the residue with two conformations. On dropping the contour level to 2σ(0.2e/Å^3), both conformation are seen to be in observable density (Fig. 1b).

The refined occupancies and average temperature factors for the disordered part of the sidechain are 55% and 11.3Å^2, respectively, for one conformation (conformation A, which is similar to the conformations seen in met-, deoxy- and oxy-Mb) and 45% and 12.7Å^2 for the other (conformation B, which has not been observed before). The sidechain B-factors for both conformations are close to the average sidechain B-factors of 12.6Å^2 for the whole protein. These refined occupancies and temperature factors indicate that the new conformation is appreciably populated. An alternative interpretation is that the extra density is due to ordered solvent and not to an alternate conformation for the sidechain. We have tentatively ruled out this possibility because the density at 2σ is contiguous with the sidechain density and is not close to the polar atoms of the sidechain.

The electron density in a difference Fourier map for a region not included in the phase calculation is reduced to approximately half its value (Blundell and Johnson, 1976). This, coupled with the fact that the true electron density for the weaker alternate conformation is already reduced by 50% or more, can make it difficult to interpret the electron density for disordered regions of the protein. To test the reliability of the refined occupancy, the following control refinements were done.

The sidechain of another arginine, Arg 139 (H15), was modelled as having two conformations. The difference Fourier maps did not show any unaccounted density around this residue. The test conformation was built into a crevice on the surface of the protein in such a way that the atoms in the sidechain did not have any bad contacts. Refinement was started in the usual way, with both conformations assigned initial occupancies of 50%. In just two cycles the occupancy of the test conformation dropped to 13%. Further refinement did not reduce the occupancy, suggesting that an occupancy of 13% is indistinguisable from the noise level of the map and that the new conformation for Arg 45 CD3, at an occupancy of 45%, represents some real feature of the electron density.

In another test, Arg 45 (CD3) and Arg 139 (H15) were both modelled by single conformations. For Arg 45 (CD3) this was the conformation at 45% occupancy (conformation B) and for Arg 139 (H15) this was the (non-existent) test conformation. Refinement against the X-ray data was started with B-factors of 10.0Å^2 assigned to the atoms in both sidechains. In three cycles the sidechain B-factors for Arg 139 (H15) increased to more that 35Å^2 for the atoms beyond C_γ. However, after eleven cycles of refinement, the average B-factor, beyond C_γ, for Arg 45 (CD3) in conformation B is 20Å^2. This is another indication that this conformation is significantly populated.

The existence of two or more significantly populated conformations for a protein sidechain is not uncommon (Steigemann and Weber, 1979, Palmer et al.,1984, Honzatko et al., 1985). In fact, as the resolution of protein structures improves, increasing numbers of alternative conformations are observed. In oxy-Mb, Phillips (1980) has reported alternate conformations for four sidechains (Val 13 (A11), Leu 86 (F1), Leu 89 (F4) and Gln 128 (H5)). Based on the neutron data, Hanson and Schoenborn (1981) have also reported alternate conformations in CO-Mb, but for four different sidechains (Lys 16 (A14), Lys 47 (CD5), Lys 78 (EF1) and Asp 122 (GH4)).

There are no residues in common between the eight reported earlier and the seven reported as being disordered in this work. Difference electron density maps, with coeffecients $(F_o-F_c)e^{2\pi i\alpha_c}$ were calculated from the X-ray CO-Mb data and the refined structure with these residues omitted. Lys 16 (A14) and Asp 122 (GH4) have no density for the terminal atoms even at low contour levels thus making it impossible to model these residues adequately. Val 13 (A11), Lys 47 (CD5) and Leu 86 (F1) show evidence for disorder but in each case the alternate conformations are only weakly populated. Lys 78 (EF1), Leu 89 (F4) and Gln 128 (H5) are in good density and there is no evidence for any disorder. Of the alternate conformations in the X-ray CO-Mb model (this work), that for Met 131 (H8) is of doubtful significance since the refined occupancy of 14% is barely above the noise level. The electron density, however, cannot be satisfactorily fit by a single conformation.

II Local Structural Effects of Ligand Binding.

There are two major questions regarding the ligand binding site in CO-Mb. The first concerns the stereochemistry of the CO ligand in Mb and its effect on the packing of distal atoms around the binding site. The second concerns the structure of the heme group and the heme proximal-histidine linkage.

In crystal structures of carboxy-porphyrins, the Fe-C-O group is linear and perpendicular to the heme plane (Peng and Ibers, 1976); this is presumed to be the minimum energy conformation of the CO. In contrast to this, the CO is bent away from the heme normal in various globins such as the annelid bloodworm CO-Hb (Padlan and Love, 1974), CO-erythrocruorin (Steigemann and Weber, 1979), human CO-Hb (Baldwin, 1980) and CO-Mb (Hanson and Schoenborn, 1981). EXAFS measurements on CO-Mb have determined that the FE-C-O angle is 127±4°, (Powers et al., 1984), similar to that in a "pocket" porphyrin, FePocPiv(1-MeIm)(CO) (Collman al., 1983).

The details of the CO deformation are of interest because it represents a balance between the strain in the protein and in the ligand (Case and Karplus, 1978) and also because it is believed to be responsible for the lower affinity of Mb and Hb for CO relative to that of isolated porphyrins. Oxygen, on the other hand, binds to both heme proteins and isolated porphyrins in a bent conformation (Phillips, 1980); Mb and Hb therefore appear to discriminate against CO in favour of O_2. The structure of the CO ligand in the X-ray structure of CO-Mb is discussed in Section II (i) below. The extent to which the protein limits the conformations of the ligand is examined using empirical energy functions with the program CHARMM (Brooks et al., 1983) in Section II (ii) below. In Sect II (iii) the disorder of Arg 45 (CD3) is related to a possible pathway for ligand entry into the heme pocket.

In Hb the allosteric transition between the T (deoxy) and R (liganded) quaternary structures is triggered by the motion of the iron towards the heme plane on ligand binding (Perutz, 1978). In oxy-Mb the iron is 0.18Å from the mean heme plane (Phillips et al., 1980). Baldwin and Chothia (1979) have speculated that the lower oxygen affinity of Mb with respect to R-state Hb is due to the fact that the structure of Mb prevents the iron from being in the heme plane. However, it is shown below that in X-ray structure of CO-Mb, as in the neutron structure (Hanson and Schoenborn, 1981), the iron is in the least squares plane of the heme.

In this connection it is important to know the extent to which the motion of the iron on ligand binding is tracked by the proximal histidine and how the ligand binding changes the heme proximal-histidine linkage. In Hb this linkage is implicated both in the transmission of the effects of the ligand binding to the subunit interface and in the reduced oxygen affinity of the T-state of the molecule with respect to that of the R-state (Gelin and Karplus, 1977, Baldwin and Chothia, 1979, Gelin et al., 1983, Friedman 1985).The structure of the heme group and the heme proximal-histidine linkage in the X-ray structure of CO-Mb are described in Section II (iv) below.

The changes in Mb on CO binding are analysed with reference to the structure of deoxy-Mb. Most of the analysis is based on the 2.0Å structure of deoxy-Mb at pH 6.0, which has been refined by Takano (1977b) to an R-factor of 23.3% using the real-space refinement procedure of Diamond (1971). Phillips (1981) has refined a model for deoxy-Mb at pH 8.5 by the reciprocal-space method of Jack and Levitt (1978) against X-ray data to 1.4Å resolution. This structure is expected to be more accurate, but the details regarding the refinement and the final R-factor have not been published yet. The structure has been obtained from the Brookhaven Protein Data Bank and is used to verify the conclusions drawn from the Takano (1977b) structure. The r.m.s. deviation in backbone (N, C, C_α) atom positions between the two structures is 0.25Å, excluding the N- and C-terminal regions.

II (i) The stereochemistry of the CO ligand.

Fig. 2 shows the heme group and the nomenclature used to identify the atoms. We characterize the possible CO conformations by the dihedral angle ϕ between the O-C-Fe and C-Fe-NC planes (see Fig. 2) and the angle θ(Fe-C-O); $\phi = 0^\circ$ corresponds to the CO eclipsing the Fe-NC bond and $\theta = 180^\circ$ corresponds to a linear Fe-C-O angle. Initially two conformations were built with $\phi = 10^\circ$ and $\phi = 90^\circ$. These refined to structures with $\phi = 20^\circ$ and $\theta = 154^\circ$ (conformation A) and $\phi = 93^\circ$ and $\theta = 120^\circ$ (conformation B). At a later stage of the refinement it became clear that a conformation at $\phi = -60^\circ$ was also required to fit the density (see Figs. 3 and 4(a)). Rather than refine more than two conformations for the CO, conformation B was removed from the model and a new conformation at $\phi = -60^\circ$ was added. Refinement was continued and the final values of ϕ and θ are $\phi = 60^\circ$ and $\theta = 141^\circ$ (conformation C, with a refined occupancy of 78%) and $\phi = -62^{\circ}$ and $\theta = 120^{\circ}$ (conformation D, with a refined occupancy of 22%). The carbon atom was never modelled with more than one position. The value of ϕ for the oxygen ligand in oxy-Mb is 20°, close to conformation A of the CO ligand. Phillips (1980) has noted that the terminal oxygen atom might have more than one position in oxy-Mb.

Fig. 3 shows the electron density at the ligand binding site, calculated using a difference Fourier synthesis with coefficients $(F_o - F_c)e^{2\pi i \alpha_c}$. The phases, α_c, and the structure factors, F_c, are calculated from the final refined structure without including the atoms of the CO. The density in Fig. 3 is at 3 and 5.5 σ (0.3 and 0.55e/Å^3) above the average density of the map; at these levels other regions of the difference map within the protein boundary are completely featureless. We have modelled the oxygen of the CO with two positions; the two final conformations of the CO group are shown in Fig. 3.

There is no density for one of the positions of the oxygen atom at 5.5 σ. This conformation has refined to an occupancy of 22% which is close to the noise level of the map (13%, see above). However, the B-factor of the oxygen atom is low (6.3Å^2). Electron density maps calculated using an independantly collected 2.0Å CO-Mb data set (Kuriyan, unpublished) with the CO atoms omitted from the phase calculation, show very similar density around the the ligand. The refined occupancies for the two conformations in the independantly refined 2.0Å structure are 65% and 35%, verifying that the population of the second conformation is significant. There is no evidence for a water molecule in the heme cavity in either data set, in contrast to oxy- and deoxy-Mb (Takano 1977b, Phillips, 1980).

The only stereochemical restraints on the CO ligand involved the Fe-C and the C-O bond lengths, which were restrained to values of 1.85Å and 1.20Å (see note in Section I). The carbon atom has refined to a position directly above the iron, on the heme normal, even though no such restraint was placed on it. The refined Fe-C distance is 1.92Å. EXAFS measurements on CO-Mb at 4K indicate that the Fe-C bond length is 1.93±0.02Å (Powers et al., 1983). The agreement between the value inferred from EXAFS and the X-ray diffraction value is probably coincidental because the latter is sensitive to the restraints used. On continuing the refinement for a few cycles with no restraints on the Fe-C bond, the bond length increased to 2.27Å.

Steigemann and Weber (1979) have remarked on the anomalously large Fe-C bond length (2.4Å) in CO-erythrocruorin and suggest that this is due to repulsion between the C atom (which is 0.3Å away from the heme normal in their structure) and the N atom of pyrrole ring 2 (see Fig. 2 for heme nomenclature). However, Yu et al. (1984) claim that the Fe-C bond length reported by Steigemann and Weber (1979) is in error. Resonance Raman measurements indicate that an upper limit on the Fe-C bond for CO-erythrocruorin in solution is 1.8Å (Yu et al., 1984). Inadequate modelling of the ligand disorder could cause the refinement to move the carbon away from its true position; such behaviour for atoms near disordered regions of the protein has been noted earlier (Kuriyan et al., 1985) and might be responsible for the apparent lengthing of the bond in both CO-Mb and CO-erythrocruorin.

The Fe-C-O bond angles were not restrained. The final refined values are 141° for conformation C and 120° for conformation D. The oxygen atoms in these conformations are displaced by 0.8Å and 1.0Å, respectively, from the heme normal. The CO bond length is 1.17Å in conformation C and 1.20Å in conformation D. EXAFS measurements (Powers et al., 1984) indicate a value of 127±4° for the Fe-C-O angle, which is between that found in the two refined conformations. Recent experiments using X-ray near-edge spectroscopy (XANES) indicate that the Fe-C-O angle is 150° in MbCO crystals and in solution (Bianconi et al., 1985).

The situation with modelling the CO can be summarized by plotting the overlap of calculated and observed density as a function of θ and ϕ. This is done in Fig. 4(a) for a difference electron density map calculated using coefficients F_o-F_c, ommitting the CO atoms from the calculation of F_c. The conformation at $\varphi = 60^\circ$ and $\theta = 140^\circ$ is seen to be dominant one.

An alternative model for the CO could have the Fe-C-O angle linear, but tilted with respect to the heme normal. If the oxygen positions are kept the same as in conformations C and D above, the carbon atom would have to be displaced by 0.5Å and 0.75Å, respectively, in order to form linear Fe-C-O angles. Such displacements lead to shorter Fe-C bond lengths (1.7Å and 1.4Å), but would place the carbon atom outside the electron density. Bent Fe-C-O structures are more consistent with the electron density, but at this resolution X-ray diffraction alone cannot conclusively establish which model is better.

II (ii) Empirical Energy Calculations of Ligand-Protein Interactions.

Case and Karplus (1978) have examined the stereochemistry of carbon-monoxide binding to myoglobin by using empirical energy functions to calculate the potential energy as a function of ligand position in a plane within the heme cavity. Their calculations showed that a linear perpendicular conformation would be unfavourable because of van der Waals repulsion between the oxygen of the CO ligand and the $N_{\varepsilon 2}$ atom of His 64 (E7), the distal histidine. Phillips (1980) has calculated the protein - O_2 interaction energy in oxy-Mb as a function of rotation around the Fe-O bond (i.e., as a function of ϕ) and has showed that the ligand non-bonded energy is low in the region $-60^\circ < \phi < +60^\circ$, which is consistent with the extent of disorder observed in CO-Mb.

In order to compare the results of Case and Karplus (1978) and Phillips (1980) with our results, we have done two types of calculations. In one we calculate the van der Waals interaction energy between the ligand and the protein as a function of both ϕ (NC (heme) - Fe - C - O) and θ(Fe-C-O). The protein is kept rigid while ϕ and θ are varied through $0-360^\circ$ and $90-180^\circ$, respectively. In this calculation we ignore electrostatic effects (because of uncertainty regarding the partial charges on the CO), as well as the internal energy of the ligand; In CO-Mb the distal histidine does not form a strong H-bond with the ligand, unlike in met- or oxy-Mb (Norvell et al., 1975, Phillips and Schoenborn, 1981). The consequences of this are discussed under Ligand Binding Pathways, Section II (iii).

The other calculation is similar to that of Case and Karplus

(1978), where the potential energy of the oxygen atom of the CO is calculated in a rigid protein at various positions in a plane 2.5Å above, and parallel to, the heme plane. This is the elevation of the oxygen atom in conformation D and is somewhat lower than that used by Case and Karplus since it corresponds to a bent ligand. While Case and Karplus (1978) used the structure of met-Mb (Takano,1977a) to calculate their energy maps, we have used the final refined coordinates of CO-Mb obtained in this work as well as the coordinates for deoxy-Mb (Takano, 1977b). This allows us to study the effect of the ligand on the protein, since, as pointed out by Case and Karplus (1978), the strain on ligand binding that leads to distortion of the CO is likely also to result in structural changes in the protein. However, they did not examine the extent to which the protein would relax.

II (ii) a. Protein-Ligand Energy for CO rotation and bending.

Fig. 4(b) shows the two dimensional van der Waals energy surface for the bending and rotation of the CO ligand in CO-Mb, with the protein kept rigid. The non-bonded energy is low for values of ϕ between 60° and -60° i.e., the ligand is restricted to point towards pyrrole ring C of the heme. This is similar to the result obtained for oxy-Mb by Phillips (1980). The minimum energy conformation has the oxygen directly over the Fe-NC bond ($\phi = 0^\circ$) and bent ($\theta = 130\text{-}140^\circ$). The potential energy minimum is very shallow along both θ and ϕ; the two refined conformations, though separated by 120° in ϕ are within 3 and 6 Kcal/mole of the minimum energy conformation (Fig. 4b). The broad, shallow minimum in the ϕ-θ energy map is consistent with the observed disorder in the CO ligand (Fig. 4a). The difference in van der Waals energy between the minimum energy bent conformation and the linear conformation is about 12 kcal/mole. Though higher than for the bent conformations, this is much lower than the value of 90 kcal/mole obtained by Case and Karplus (1978) for a linear CO in met-Mb. The energy of CO rotation has also been calculated using the deoxy-Mb structure where both conformations are in regions of very high energy (Fig. 4c), indicating that the protein atoms have relaxed around the CO in CO-Mb.

II (ii) b. Energy Surfaces for the ligand in a plane above the heme.

Fig. 5(a) shows a 10Å by 10Å section of the energy map calculated for an oxygen atom inside the deoxy-Mb structure (the map was calculated without including the carbon atom of the ligand). The ligand was moved through the region of the map in steps of 0.25Å along X and Y, and the van der Waals energy of interaction between the protein and the oxygen atom was evaluated at every step by the program CHARMM using standard energy parameters (Brooks et al.,1983). The projections of the distal histidine (64 E7) and Val 68 (E11) in the plane of the map, as well as that of the two initial and two final conformations of CO, are also shown in the Figure. The heme pocket is quite restrictive in deoxy myoglobin, and the range of conformations apparently accessible to the CO ligand would be of high energy if the protein matrix did not relax.

Quite different results are obtained when the rigid protein energy map is calculated using the coordinates of CO-Mb instead of the deoxy structure (Fig. 5(b)). The region of low potential energy in the map is significantly larger, indicating that the protein has relaxed around the binding site.

It is of interest to compare the interaction energies of the four conformations for the CO (A,B,C and D, see above) in deoxy-Mb and in CO-Mb. In the deoxy-Mb structure, B and C ($\phi = 93^{\circ}$ and 60°) have very high interaction energies (2400 Kcal/mole and 270 Kcal/mole respectively). The other two conformations, A and D, ($\phi = 20^{\circ}$ and -62°,respectively) have smaller, but still repulsive, energies, mainly due to nonbonded contacts between the oxygen atom of the ligand and the sidechain atoms of residues His 64 (E7) and Val 68 (E11). In the deoxy-Mb structure, the most stable conformations are A and D (Fig. 5a). With the refined CO-Mb coordinates conformations A and C have net attractive van der Waal's energies of -3.0 Kcal/mole and -2.8 Kcal/mole respectively. The energy of conformation D is slightly repulsive, due to a close contact with Val 68 (E11), but at 1.0 Kcal/mole it is somewhat more stable than in the deoxy structure. Conformation B is relatively unstable, with an energy of 6.0 Kcal/mole, but its energy is more than 2000 Kcal/mole lower than in the deoxy structure.

Thus, in the CO-Mb structure, the two final refined conformations (C and D) are quite stable with respect to interactions with the protein and their relative energies (-2.9 Kcal/mole and 1.0 Kcal/mole) are consistent with their refined occupancies of 78% and 22%. The observed disorder of the CO is also consistent with the larger binding site in CO-myoglobin. Relaxation of the protein around the binding site has resulted in all four conformations examined being relatively stable (their energies lying within approximately 10 Kcal/mole of each other). These energies are only approximate, however, because no minimization of the protein-ligand system was done to relax strain due to errors in the atomic coordinates (Gelin et al., 1983).

II (ii) c. Changes in protein structure around the ligand.

The heme groups in deoxy- and CO-Mb were superimposed by least-squares in order to calculate the energy maps in Fig. 5a and 5b, i.e., the maps in Figs. 5a and 5b are in the same orientation with respect to the heme. On binding CO, the center of the heme group moves by 0.24Å into the heme pocket and the entire heme group rotates by a small amount (2°) about an axis close to the NB-ND vector. The effect of this small motion is to move the ligand away from the distal histidine and Val 68 (E11). This change is similar in magnitude to that seen in the monomeric globin CO-erythrocruorin (Steigemann and Weber, 1979). It is, however, smaller than that observed in the R-T transition in hemoglobin where the hemes move into the heme pocket by 0.5Å in the α subunit and 1.5Å in the β subunit (Baldwin and Chothia, 1979).

Superimposed on the movement of the heme into the heme pocket is the effect of changes in the sidechain packing around the heme group. The largest of these is the motion of the distal histidine away from the ligand. On superimposing the main chain atoms of CO- and deoxy-Mb (Takano, 1977b), the $N_{\varepsilon 2}$ atom of the distal histidine moves by 1.7Å, most of this motion being away from the ligand (see Fig. 6).

The results discussed so far have been based on comparisons between the X-ray CO-Mb structure and the deoxy-Mb structure of Takano (1977b). Similar results are obtained when using the high pH deoxy-Mb structure of Phillips (1981). In the Table below the positional changes that occur between deoxy-Mb and CO-Mb in residues within 8Å of the CO carbon in

CO-Mb are given. The heme groups of the two deoxy-Mb structures were individually superimposed on the CO-Mb heme group; the deviations between the CO- and deoxy-Mb structures given in the Table therefore have contributions due to the shift of the heme group into the heme pocket on ligand binding, as well as due to changes in the relative positions of the atoms. The shifts have not been decomposed into components away from the ligand since the ligand disorder makes such a decomposition difficult.

The magnitudes of the deviation observed for each residue are very similar for the two deoxy-Mb structures. Most of the residues surrounding the heme pocket have deviations that are much larger than the 0.24Å shift of the heme group into the heme pocket, i.e., the contribution due to sidechain rearrangement is significant. The largest changes are seen in Phe 46 (CD4), Arg 45 (CD3), His 64 (E7) and Thr 67 (E10).

II (iii) Pathways for Ligand Entry Into the Heme Pocket.

It is well known that in the X-ray structures of myoglobin and hemoglobin there are no paths for a ligand to enter the binding pocket from the solvent (Case and Karplus, 1978). In Fig. 5 (a) and 5 (b), for example, passage from the binding cavity to the protein exterior in the upper left hand corner of the Figure is blocked by His 64 (E7) in both deoxy- and CO-myoglobin. Case and Karplus (1978), on the basis of energetic calculations on the flexibility of His 64 (E7) and Val 68 (E11), suggested that a channel could be opened for ligand entry between the distal histidine and Val 68 (E11) by rotations of 15° and 100° in χ_1 for these two residues. On binding phenylhydrazine, a different channel opens up in the protein to accomodate the bulky ligand (Ringe et al., 1984). The distal histidine swings out of the distal cavity, opening a channel between itself, Phe 46 (CD4) and pyrrole rings 1 and 4 of the heme group. Normally, this outward motion of the histidine is prevented by steric interaction with the sidechain of Arg 45 (CD3). In the binding of phenylhydrazine, the histidine is displaced by interaction with the ligand, which in turn "pushes" the arginine outwards into the solvent (Ringe et al., 1984).

Figure 7a is an energy map of the χ_1-χ_2 surface for His 64 (E7) in CO-Mb. The energy of the sidechain as a function of χ_1 and χ_2 is calculated with the rest of the protein kept rigid. This map is similar to that obtained by Case and Karplus (1978) using the structure of met-Mb (Takano, 1977a). The value of χ_1 is restricted to lie between approximately 180° and 250° because of collisions with the ligand (for χ_1 less than 180°) and Arg 45 (CD3) (for χ_1 greater than 250°).

As discussed in Section I, the sidechain of Arg 45 (CD3) is disordered in CO-myoglobin (see Fig. 1).The positions of the terminal atoms of the sidechain are separated by more than 4.0Å in the two conformations because of large changes in the dihedral angles (χ_1 through χ_5 change by 18°, 57°, 161°, 157°, and 10°, respectively). The new conformation seen in CO-Mb is quite different from that seen in the phenylhydrazine-Mb structure, in which the arginine sidechain moves away from the distal histidine into the solvent. In CO-Mb the conformaion B of the sidechain packs into a crevice on the surface of the protein, close to the distal histidine, and is much more buried than the original (i.e., that reported in all myoglobins) conformation A.

Both conformations have potential H-bonding interactions with surrounding atoms in the protein and the solvent; these are listed in Table 4. Some of the contacts are very short, which could be a consequence of the disorder. Apart from these interactions, the NH1 atom in conformation B interacts closely with the face of the aromatic ring of Phe 46 (CD4) (Fig. 8). It has recently been shown, using ab-initio quantum-mechanical calculations, that charged groups such as the ammonium ion interact favourably with aromatic molecules like benzene with stabilization energies comparable to those of conventional H-bonds (Deakyne and Meot-Ner, 1985). The interaction between the positively charged guanidinium group, with a 0.25 electron charge on the nitrogen atom, and the aromatic ring of the phenyl group is therefore likely to stabilize the B conformation of Arg 45 (CD3). It is interesting to note that the distance of the terminal nitrogen (NH1) of Arg 45 (CD3) (conformation B) from the center of the aromatic ring of Phe 46 (CD4) is 3.0Å, which is within the range of 2.91-3.10 Å reported by Deakyne and Meot-ner (1985) as being the optimal distance of approach by the ammonium nitrogen along the normal to the ring. Since conformations A and B for the arginine are on the surface of the protein, calculation of their relative stabilities would require treatment of the solvent contribution to the free energy. Such a calculation is planned.

Arg 45 (CD3) is a residue in the flexible CD corner of myoglobin. Some of the residues in the loop regions (BC, CD, EF, FG and GH) undergo large shifts in position in CO-Mb relative to that in deoxy-Mb. On superimposing the two structures by their backbone atoms, the average backbone displacement for Arg 45 (CD3) is 0.5Å; larger displacements are seen only in a few other loop residues. In conformation A, which is the closer of the two to the conformation in deoxy-Mb, the sidechain atoms are displaced, on average, by 0.9Å relative to their positions in deoxy-Mb. For conformation B, the average sidechain shift (from deoxy) is 2.3Å, with the terminal nitrogen atoms displaced by 5.0Å and 3.5Å. The two conformations, some of the neighboring residues, and the heme group are shown in Fig. 8.

A consequence of this alternative conformation for Arg 45 (CD3) is that the motion of the distal histidine away from the heme pocket is no longer blocked when the arginine is in conformation B. Fig. 7b shows the χ_1,χ_2 energy map for the distal histidine, this time with Arg 45 (CD3) in conformation B. In this map the value of $\chi 1$ for the distal histidine is no longer restricted to be less than 250° and, in fact, a low energy region of the map has opened up around $\chi_1 = 280°$ and $\chi_2 = 275°$. This conformation for the histidine is close to that seen in the phenylhydrazine-Mb structure (Ringe et al., 1983). A map of the ligand energy in a plane above the heme with the histidine in this conformation shows that a channel opens up between the surface of the protein and the binding site (Fig. 5(c)).

While difference electron density maps for CO-Mb show no evidence for a significantly populated alternative conformation for the distal histidine, comparison with maps for met-Mb indicate that the histidine is relatively more mobile in CO-Mb than in met-Mb. Difference electron density maps for met-Mb (Kuriyan, unpublished) also indicate that Arg 45 (CD3) is <u>not</u> disordered in the met form of the protein. A sulphate ion which is bound to the distal histidine and Arg 45 (CD3) in met-Mb is absent in the neutron structure of CO-Mb (Schoenborn et al., 1981) as well as in the two X-ray structures of CO-Mb. There is no significant change in the Arg 45 (conformation A) - distal histidine interaction between met-Mb and CO-Mb in the region where the sulphate is bound, and so the disappearance of the sulphate ion is probably connected with the disorder of Arg 45 and the increased flexibility of His 64 (E7). Neutron diffraction studies have shown that the distal histidine is hydrogen bonded to the ligand in met- and oxy-Mb, but not in CO-Mb (Norvell, et al., 1975, Phillips and Schoenborn, 1981).

The increased mobility of the histidine that is implied by the lack of hydrogen bonding to the ligand, the disorder of Arg 45 (CD3), the poor electron density and the lack of a bound sulphate ion is confirmed by a comparison of the atomic temperature factors for His 64 (E7) in met-Mb (Kuriyan, unpublished) and the two CO-Mb structures (Table 5). The average backbone B-factors of the new CO-Mb structure are very much higher than those for the first two structures, presumably because of a larger static disorder contribution; this can, however, be treated as a constant offset in the B-factors (Kuriyan, unpublished). The difference in the sidechain mobility between met-Mb and CO-Mb is made clearer by subtracting the average backbone B-factor for the residue in each structure from the sidechain B-factors. The results are given in Table 5.

The trend is clear. In met-Mb the sidechain B-factors are essentially the same as for the backbone. In both CO-Mb structures, the sidechain B-factors are increased by about 5.5Å^2. This increase in B-factors is greater than the average backbone B-factors for the distal histidine in met-Mb.

The results presented in this section suggest that a possible reaction coordinate for ligand entry might involve a combination of Arg 45 (CD3) and His 64 (E7) sidechain torsions. It would be of interest to estimate the relative free energies of the different conformations for the two residues, as well as the barrier between the two conformations, taking into consideration the effects of solvent and crystal environment.

II (iv) The heme group.

When a least-squares plane of the 24 central porphyrin atoms is defined (the heme plane), the iron is within 0.03Å of the plane in CO-Mb, and 0.50Å from the plane in met-Mb (Takano, 1977b). The rms deviation from planarity for the 24 porphyrin atoms is 0.08Å in CO- and 0.12Å in deoxy-Mb The iron is exactly in the least-squares plane of the four porphyrin nitrogens in CO-Mb, and 0.45Å from this plane in deoxy-Mb. Fig. 2 shows all the atoms that deviate more than 0.1Å from the heme plane in CO-Mb, and compares these deviations with those seen in deoxy-Mb (Takano,1977b). The lack of any doming in CO-Mb results in very few atoms deviating significantly from the heme plane. The outer atoms of pyrrole 1 tip towards the proximal side and those of pyrrole 3 tip towards the distal side. These displacements are similar to those seen in deoxy-Mb (see Fig. 2), indicating that they arise from similar packing constraints in the heme pocket. The two vinyl groups are significantly non-planar, with the C_β atoms twisted from the heme plane by about 1Å. This non-planarity of the vinyl groups is due to van der Waals repulsion between their terminal atoms and the adjacent methyl substituents in the planar conformation.

Various parameters for the heme group and the proximal histidine are compared for two deoxy-Mb structures, (Takano, 1977b, Phillips, 1981) CO-Mb (this work), oxy-Mb (Phillips, 1980) and met-Mb (Kuriyan, 1985) in Table 6. The asymmetry of the heme-proximal histidine linkage in hemoglobin has been implicated as an important element in the initiation of the tertiary and quaternary structural changes on ligand binding (Gelin et al., Freidman, 1985). In the human deoxy-hemoglobin A α chain, for example, the difference between the His 93 F8 $C_{\varepsilon 1}$ - NA (heme) and His 93 F8 $C_{\delta 2}$ - NC (heme) distances is 0.8Å, and this difference is considerably reduced in CO-hemoglobin due to tilting of the heme group as well as a motion of the F-helix across the face of the heme (Baldwin and Chothia, 1979, Gelin et al., 1983).

In Mb this asymmetry is small in the deoxy state (the two distances being 3.50Å and 3.44Å, respectively, in the Takano (1977b) structure and 3.20Å and 3.46Å in the Phillips (1981) structure; the significance of the difference between the two structures is not known since the refinement of the latter structure has not yet been published). In CO-Mb both distances decrease to 3.10Å and 3.07Å, respectively, showing that the proximal histidine follows the motion of the iron towards the heme plane; its orientation with respect to the heme normal is also somewhat more symmetric in CO-Mb than in deoxy-Mb.

The distance between the centroid of the imidazole ring of the proximal histidine and the heme plane decreases by 0.4Å in CO-Mb with respect to the Takano (1977b) structure of deoxy-Mb. If the backbone atoms (N,C_{α},C) of deoxy-Mb and CO-Mb are superimposed by least squares, the vector between the centroids of the heme groups is almost entirely in the plane of the two hemes (the out-of-plane component is less than 0.02Å while the in-plane component is 0.24Å). Thus the change in the proximal-histidine heme linkage on ligand binding is best described as a motion of the histidine toward the heme plane rather than the reverse. Since the proximal histidine is rather rigidly attached to the F-helix, both through the backbone and by a hydrogen bond between the N_δ atom and the carbonyl group of Leu 89 (F4) (Gelin et al., 1983), the motion of the proximal histidine is expected to be coupled to the F-helix. It is interesting to note that Leu 89 (F4) is one of five residues in the protein that are conserved between the globins of organisms as diverse as sharks and man (Dickerson and Geis, 1983); the carbonyl group involved in the H-bond is not, of course, unique to leucine and one might speculate that the packing of the residue against the heme group is somehow important for function.

III Tertiary Structural Changes

The binding of CO to Mb initiates two specific local changes: (i) a perturbation in the proximal side of the heme caused by the out-of-plane deoxy iron moving into the heme plane and (ii) one on the distal side caused by the atoms surrounding the distal pocket moving away from the ligand, the largest motion being that of the distal histidine. The motion of the iron into the heme plane is tracked by the proximal histidine; with respect to Takano's (1977b) deoxy-Mb structure the histidine moves by 0.4Å towards the heme. This motion is about twice as large as that seen in oxy-Mb or met-Mb, where the liganded iron is not in the heme plane (see Table 6). The consequences of these specific

The Libraries
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Institute Archives and Special Collections
Room 14N-118
(617) 253-5688

There is no text material missing here.
Pages have been incorrectly numbered.

Page 244

local changes on the tertiary structure of the protein are now examined.

The structures of deoxy- and CO-Mb were examined on a PS300 graphics system using the software HYDRA (Hubbard, 1985). This program allows one to interactively superimpose various parts of the molecules by least-squares, and simultaneously display the structures. The ability to flash between the two structures being compared greatly facilitates the analysis of the large-scale changes in the structure.

The motion of the distal histidine and other residues near the ligand, though large, is essentially local in nature. Except in the loop regions (see below) there are no significant systematic changes in the distal part of the molecule. However, on the proximal side, the motion of the iron and the proximal histidine does result in a large scale and widely distributed shift of atoms. In CO-Mb, relative to deoxy-Mb, an overall motion of most of the atoms in the proximal side towards the plane of the heme was perceptible on the graphics system. This effect was seen relative to both deoxy-Mb structures (Takano, 1977b, Phillips, 1981).

The N-terminal and C-terminal regions of the protein (residues 1-3 and 149-153) are excluded from all the calculations that follow because these are usually not well determined. Fig. 9 shows the average backbone (N, C, C_α) displacements, per residue, obtained by superimposing the backbone atoms of CO-Mb and deoxy-Mb (Takano, 1977b). The rms deviation of backbone atoms is 0.29Å. From Fig. 9 it is seen that the regions of large displacement are the CD loop, the F helix, the FG loop, the GH loop and the H helix. To determine whether these deviations, which range from 0.34Å to 1.0Å, are significant or merely represent errors in the positional coordinates, the secondary structural elements are superimposed separately. The results are shown in Table 7.

All the helices, when superimposed individually, show uniformly low rms deviations (average value 0.17Å) and the deviations of atoms in the F helix are not higher than in others. The loop regions also have low rms deviations, (the average value, excluding the single residue AB loop, is 0.24Å). This indicates that the larger deviations seen in Fig. 9 are due to changes in the relative packing of the helices and loops in deoxy- and CO-Mb. This is in accord with the results of Chothia and Lesk (1985) for different hemoglobins and that of Elber and Karplus (1986) for a molecular dynamics simulation of Mb.

The large displacement of the CD corner has already been remarked upon as having changed the position of Arg 45 (CD3) relative to the heme group. The GH loop, which also undergoes a large shift (Fig. 9), is well removed from the heme group. However, in the crystal packing of Mb (in the $P2_1$ space group), the GH loop of one molecule packs close to the CD loop of another molecule. In the crystal the loops do not make direct contact, but rather interact through several ordered water molecules. This suggests the possibility that the water transmits the effect of structural transitions in one molecule to the other. It would be interesting to investigate this by molecular dynamics simulations of the protein in the crystal lattice. However, it should be noted that the loop regions are the least well-determined parts of the molecule, with the highest temperature factors and concomitant positional errors.

When all the backbone atoms of the two molecules are used in the least-squares superpositioning (instead of just a few helices) the effect is to decrease the apparent shift of regions with large differences and increase the apparent shift in others, although the trends are unaltered (Baldwin and Chothia, 1979). Consequently, it is useful to superimpose the structures on the basis of only parts of the molecule. Table 7 lists the rms deviations of the secondary structural elements of deoxy- and CO-Mb based on least-squares superpositioning of the backbone atoms of the A,B,C and G helices, the B,C and G helices and the distal part of the molecule. In all of these, the deviations between the 500 atoms in the proximal part of the molecule involve a large component that is towards the heme plane in going from deoxy- to CO-Mb. The residues in the proximal part of the molecule form part of the A-helix (4 A2 -5 A3), the E helix (74 E17 -77 E20), the G helix (100 G1 -105 G6), the H helix (133 H9 to 148 H24) and the entire EF loop, the F helix and the FG loop.

To demonstrate the overall shift of this region towards the heme plane, the two structures are superimposed on the B,C and G helices (which are almost entirely in the distal part of the molecule and do not appear to undergo large changes in structure) and the component of the shift along the normal to the heme plane in CO-Mb is calculated. Fig. 10 shows histograms of this component for all the atoms on the proximal side of the heme, and also for just the backbone atoms. Both histograms are peaked at about 0.25Å towards the heme and more than 75% of the atoms have shifts that are in the direction of the heme plane in going from deoxy- to CO-Mb.

This large scale change in the structure can also be demonstrated by a method which is independent of the relative orientations of the two molecules. The distances between all the C_α atoms in the two structures are calculated and every residue pair for which the C_α-C_α distance changes by more than 0.25Å between deoxy-Mb and CO-Mb is identified in a C_α Δ-distance matrix (Fig. 11). On going from deoxy-Mb to CO-Mb extremely few C_α-C_α distances increase by more than 0.25Å (these are shown in the lower half of the matrix). However, a large number decrease by more than 0.25Å (the upper half of the matrix); the pattern of changes is similar for comparisons of CO-Mb with both deoxy-Mb structures (Takano, 1977b, Fig. 11a, Phillips, 1981, Fig. 11b) and corresponds to atoms in the E-F corner, the F-helix, the F-G, the G-H corner, parts of the H helix and the C terminal end of the molecule moving towards the rest of the protein.

On comparing the structures of oxy-Mb (Phillips, 1980) or met-Mb (Kuriyan, unpublished) to deoxy-Mb and CO-Mb, the former two structures are seen to be intermediate between CO-Mb and deoxy-Mb. This is consistent with the out-of-plane iron in both structures (Table 6). To demonstrate this the CO-Mb and met-Mb structures are compared in a C_α Δ-distance plot in Fig. 11c. The motion of the F-helix can be discerned, but the effects are smaller than in the comparison of deoxy-Mb and CO-Mb.

The changes in the tertiary structure on CO binding to Mb described here, though significant, are much smaller than the tertiary structural changes associated with the R-T quaternary transition in Hb, where the F-helix is translated across the face of the heme by about 1.0Å as well as being tilted with respect to it (Baldwin and Chothia, 1979). The changes described here might be comparable in magnitude to the differences between the ligated and unligated state of an Hb subunit in one quarternary structure. However, significant differences would be expected due to the fact that Mb is unconstrained while in a given quaternary structure the tertiary structure of a given Hb chain is expected to be constrained. There is considerable interest in understanding these changes as a means for interpreting the results of fast-time scale photo-dissociation experiments in Hb and Mb (Martin et al., 1983, Friedman, 1985).

Conclusions

This X-ray diffraction study of CO-myoglobin has shown that the ligand binds to the protein in more than one stable orientation. This makes it difficult to obtain reproducible structural parameters for the ligand since such estimates depend on how the ligand is modelled during refinement. Nevertheless, energy calculations as well as refinement of ligand conformations have shown that the range of conformations accessible to the ligand are limited to values of ϕ(NC-Fe-C-O) between, approximately, -60° and 60° and Θ(Fe-C-O) between 120° and 150°. Two conformations at $\phi = 60^\circ$ and -60°, respectively, are adequate to fit the electron density.

The distortion from the preferred linear perpendicular binding geometry is a consequence of steric interaction with the residues lining the distal pocket, especially His 64 (E7), Thr 67 (E10) and Val 68 (E11). The strain introduced on ligand binding results in changes in the protein as well as the bending of the ligand. The sidechain of the distal histidine shifts by more than 1.4Å relative to its position in deoxy-myoglobin (Takano, 1977b). This, coupled with other smaller changes in the region of the binding site, results in a larger distal cavity in CO-Mb than in deoxy-Mb. The disorder inferred from the electron density for the ligand is consistent with the larger binding cavity in CO-myoglobin; the range of conformations observed would not be possible if the tertiary structure remained that of deoxy-myoglobin.

There are large changes in the positions of atoms in the CD loop region, including Arg 45 (CD3), which is disordered. Two conformations for the sidechain of this residue have been identified in difference Fourier maps and refinement indicates that both are significantly populated. One conformation, which is similar to that seen in deoxy-myoglobin, blocks the movement of the distal histidine out of the distal pocket, whereas the other does not. The static crystal structure does not have any pathway for the ligand to enter the distal pocket; the disorder of Arg 45 CD3 could be connected with a mechanism for ligand entry by allowing the distal histidine to move out of the distal pocket. There is no evidence in the electron density for two discrete conformations for the distal histidine; however the sidechain is significantly more mobile in CO-Mb than in met-Mb. The hydrogen bond between the distal histidine and the ligand in met-Mb is absent in CO-Mb. Arg 45 (CD3) is not observed to be disordered in met-Mb.

The iron is in the least-squares plane of the heme, as observed in the neutron diffraction study of CO-myoglobin (Hanson and Schoenborn,1981). This shift of the iron atom 0.40Å relative to its out-of-plane position in deoxy-myoglobin (Takano, 1977b) is tracked by the proximal histidine and the F-helix. The result is that a large number of atoms in the proximal half of the molecule shift towards the heme plane, relative to their positions in deoxy-myoglobin. This overall change in structure is larger than that seen in oxy- or met-myoglobin, but is smaller than that observed in the T to R transition in hemoglobin (Baldwin and Chothia, 1979).

Acknowledgements

We are grateful to R.F. Tilton, Jr. for discussions and Rod Hubbard for help with the drawing program, HYDRA.

REFERENCES

Alben, J.O. (1978) in "The Porphyrins", Vol. II pages 323-345, Academic Press, New York.

Ansari, A., Berendzen, J., Bowne, S.F., Frauenfelder, H.F., Iben, I.E.T., Sauke, T.E., Shyamsunder, E. and Young, R.D. (1985) Proc. Natl. Acad. Sci. (U.S.A.) 82, 5000-5004.

Austin, R.H., Beeson, K.W., Eisenstein, L., Frauenfelder, H. and Gunsalus, I.C. (1975) Biochemistry 14, 5355

Baldwin, J.M. (1980) J. Mol. Biol. 136, 103-28.

Baldwin, J.M. and Chothia, C. (1979) J. Mol. Biol., 129,175-220.

Blundell,T. and Johnson, L.N. (1976) "Protein Crystallography", Academic Press, New York.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M. (1983) J. Comp. Chem., 4, 187-217.

Brooks, B.R. and Karplus, M. (1983) Proc. Natl. Acad. Sci. (U.S.A.) 80, 6571.

Brown, W.E., Sutcliffe, J.W. and Pulsinelli, D. (1983) Biochemistry, 22, 2914-23.

Case, D.A. and Karplus, M. (1978) J. Mol. Biol., 123, 697-701.

Case, D.A. and Karplus, M. (1979) J. Mol. Biol., 132,343-68.

Chothia, C. and Lesk, A.M. (1985) Trends in Biochem. Sci. ??,116-118.

Churg, A.K., Danziger, R.S. and Makinen, M.W. (1978), in "Biochemical and Clinical Aspects of Hemoglobin Abnormalities", pages 323-334, Academic Press, New York.

Collman, J.P., Brauman, J.I., Collins, T.J., Iverson, B.L., Lang, G., Pettman, R., Sessler, J.L. and Walters, M.A. (1983) J. Am. Chem. Soc. 105, 3038-52.

Dalvit, C. and Ho, C. (1985) Biochemistry, 24, 3398-407.

Diamond, R. (1971) Acta Cryst. A27 436-52.

Dickerson, R.E. and Geis, I. (1983) "Hemoglobin: Structure, Function, Evolution and Pathology", Benjamin/Cummings, Menlo Park, CA.

Deakyne, C.A. and Meot-Ner, M. (1985) J. Am. Chem. Soc. 107, 474-479.

Frauenfelder, H., Petsko, G.A., and Tsernoglou, D. (1979) Nature (London), 280, 558-63.

Friedman, J.M. (1985) Science (Washington, D.C.), 228, 1273-280

Gelin, B.R. and Karplus, M. (1977) Proc. Nat. Acad. Sci. (U.S.A.) 74, 801-5

Gelin, B.R., Lee, A. W.-M. and Karplus, M. (1983) J. Mol. Biol. 171, 489-559.

Go, N., Noguti, T. and Nishikawa (1983) Proc. Natl. Acad. Sci. (U.S.A.) 80, 3696-3700.

Hanania, G.I.H, Yeghiayan, A. and Cameron, B.F. (1966) Biochem. J. 98, 189-192.

Hanson, J.C. and Schoenborn, B.P. (1981) J. Mol. Biol.,153, 117-146.

Hendrickson, W.A. (1980) in "Refinement of Protein Structures; Proceedings of the Daresbury Study Weekend" (ed. Machin, P.A. and Elder, M.) Science and Engineering Research Council, Daresbury Laboratory, U.K., pages 1-8.

Henry, E.R., Sommer, J.R., Hofrichter, J. and Eaton, W.A. (1983) J. Mol. Biol. 166 443-451.

Honzatko, R.B., Hendrickson, W.A. and Love, W.A. (1985) J. Mol. Biol. 184, 147-64.

Hubbard, R.E. (1985) "Harvard-York Drawing Program: HYDRA", to be published.

Jack, A. and Levitt, M. (1978) Acta Cryst. A34, 931-35.

Jones, T.A. (1982) in "Computational Crystallography" (ed. Sayre, D.) Calerendon, Oxford, 3303-17.

Karplus, M. (1981) Ann. of the N.Y. Acad. Sci. 367 407-18.

Karplus, M. and McCammon, J.A., (1981), C.R.C. Critical Reviews in Biochemistry, 9, 293-349.

Karplus, M. and McCammon, J.A. (1983) Ann. Rev. Biochem. 53, 263-300.

Kendrew, J.C. and Parrish, R.G. (1956) Proc. Roy. Soc. ser. A, 238,305-324.

Kendrew, J.C., Dickerson, R.E., Strandberg, B.E., Hart, R.G., Davies, D.R., Phillips, D.C., and Shore, V.C. (1960) Nature (London), 185, 422-27.

Konnert, J.H. (1976) Acta Cryst. A32, 614-17.

Konnert, J.H., Hendrickson, W.A., (1980) Acta Cryst. A36, 344-49.

Kuriyan, J., Petsko, G.A., Levy, R.M. and Karplus, M. (1985) J. Mol. Biol., in press.

Levitt, M., Sander, C. and Stern, P.S. (1985) J. Mol. Biol. 181, 423-447.

Levy, R.M., Sheridan, R.P., Keepers, J., Dubey, G.S., Swaminathan, S. and Karplus, M. (1985) Biophys. J. (in press).

Martin, J. L., Migus, A., Poyart, C. Lecarpentier, Y. Aster, R., Antonetti, A. (1983) Proc. Natl. Acad. Sci. (U.S.A.) 80, 173-177.

McCoy, S. and Caughey, W.S. (1971) in "Probes of Structure and Function of Macromolecules and Membranes, Vol II. Probes of Enzymes and Hemoproteins", B. Chance, T. Yonetani and A.S. Mildvan, Editors, Academic Press, New York. page 289.

Makinen, M.W., Houtchens, R.A. and Cougher, W.S. (1979) Proc. Nat. Acad. Sci. (U.S.A.) 76, 6042-46.

North, A.C.T., Phillips, D.C. and Mathews, F.S. (1968) Acta Cryst. A24, 351-59.

Norvell, J.C., Nunes, A.C. and Schoenborn, B.P. (1975) Science (Washington, D.C.), 190, 568-570.

Padlan, E.A. and Love, W.E. (1974) J. Biol. Chem. 249, 4067-78.

Palmer, R.A., Moss, D.S., Haneef, I. and Borkakoti, N. (1984) Biochim. et Biophys. Acta 785 81-88.

Peng, S.M. and Ibers, J.I. (1976) Biochemistry, 12, 134-9.

Perutz, M.F. (1978) (December) Scientific American, 92-123.

Petsko, G.A. (1975) J. Mol. Biol. 96, 381-92.

Petsko, G.A. and Ringe, D. (1984) Ann. Rev. Biophys. Bioeng. 13, 331-71.

Phillips, S.E.V. (1980) J. Mol. Biol. 142 531-54.

Phillips, S.E.V. (1981) The X-ray structure of deoxy-Mb (pH 8.5) at 1.4Å resolution. Brookhaven Protein Data Bank.

Phillips, S.E.V. and Schoenborn, B.P. (1981) Nature (London), 292, 81-82.

Powers, L., Sessler, J.L., Woolery, G.L. and Chance, B. (1984) Biochemistry, 23, 5519-23.

Ringe, D., Petsko, G.A., Kerr, D., Ortiz de Montellano, P.R. (1984) Biochemistry, 23, 2-4.

Shulman, R.G., Wuthrich, K., Yamane, T., Patel, D.J. and Blumberg, W.E. (1970) J. Mol. Biol. 53 143.

Steigemann, W. and Weber, E. (1979) J. Mol. Biol. 127, 309-338.

Takano, T. (1977a) J. Mol. Biol. 110, 537-68

Takano, T. (1977b) J. Mol. Biol. 110, 569-84

Watenpaugh, K.D. Sieker,L.C., and Jensen L.H. (1980) J. Mol. Biol. 138, 615-33.

Willis, B.T.M. and Pryor, W. (1975) "Thermal Vibrations in Crystallography", Cambridge Univ. Press, London. 280 pages.

Wyckoff, H.W., Doscher, M., Tsernoglou, D., Inagami, T., Johnson, L.N., Hardman, K.D., Allewell, N.M., Kelly, D.M., and Richards, F.M. (1967) J. Mol. Biol. 27 563-78.

Yu, N.-T., Benko, B., Kerr, E.A. and Gersonde, K. (1984) Proc. Natl. Acad. Sci. (U.S.A.), 81, 5106-5110.

Table 1: Restraints in the refinement.

This Table gives the target standard deviations (σ) and the final standard deviations for the deviations from ideality in the restrained parameters. See Hendrickson (1980) for a discussion of the various terms. The values of the target standard deviations are those of Hendrickson (1980) with minor modifications.

	Target σ	Final σ	Unit
Bond distances	0.030	0.030	Å
Angle distances	0.040	0.047	Å
Planar distances	0.052	0.057	Å
Planar groups	0.025	0.016	Å
Chiral groups	0.150	0.170	Å^3
Torsion Angle Restraints:			
Planar			
(0,180)	5.0	6.2	degrees
Staggered			
(±-60,180)	15.0	20.0	
Orthonormal	15.0	34.0	
B-factor Restraints:			
Backbone bonds	1.0	0.9	Å^2
Backbone angles	1.0	1.1	Å^2
Sidechain bonds	1.0	1.1	Å^2
Sidechain angles	1.5	1.9	Å^2
Shift Restraints:			
Positional	0.3	-	Å

B-factors	3.0	-	Å^2
Occupancies	0.05	-	-

TABLE 2 Occupancies and average temperature factors of disordered sidechains

Only two conformations were mod~~e~~lled for each disordered residue. The occupancies (Q) were constrained to sum to unity every atom in the same conformation had equal occupancy, thus only one occupancy factor was refined per residue. The temperature factors are averages from the $ C sub gamma $ atom out. Notice the very low B-factors for certain residues. The overall temperature factor of the structure is lower than that for other myoglobin structures.

	Conformation 1		Confromation 2	
	Q	<B>	Q	<B>
Arg 45	0.55	11.3	0.45	12.7
Leu 61	0.51	0.9	0.49	0.9
Ile 75	0.29	3.5	0.71	3.6
Gln 91	0.58	10.5	0.42	10.4
Lys 96	0.41	10.5	0.59	12.4
Met 131	0.86	5.5	0.14	3.6
Lys 147	0.41	7.9	0.59	9.9

Table 3: Shifts in atomic positions around the ligand binding site

	Average Backbone Deviation Å		Average Sidechain Deviation Å		Maximum Deviation Å		Atom with Maximum Deviation
Residue	1	2	1	2	1	2	
Leu 29 (B11)	0.7	0.6	0.7	0.6	0.9	0.7	CD1
Leu 32 (B13)	0.5	0.6	0.5	0.5	0.6	0.7	N
Phe 33 (B14)	0.7	0.7	0.7	0.7	0.8	0.8	C
Phe 43 (CD1)	0.5	0.5	0.3	0.4	0.5	0.5	O
Arg 45 (CD3)	0.7	0.3	1.1	1.1	1.6	2.6	NH2
Phe 46 (CD4)	0.9	0.6	0.9	0.6	1.2	0.9	O
His 64 (E7)	0.7	0.7	1.4	0.9	2.0	1.2	NE2
Gly 65 (E8)	0.7	0.6	0.8	0.6	0.8	0.7	N
Thr 67 (E10)	0.4	0.5	0.8	0.8	1.4	0.9	CG2
Val 68 (E11)	0.5	0.4	0.5	0.4	0.5	0.5	CG1
Leu 69 (E12)	0.6	0.5	0.3	0.8	0.8	0.7	CD2
Ala 71 (E14)	0.5	0.5	0.5	0.6	0.5	0.5	CB
Ile 107 (G8)	0.4	0.3	0.5	0.4	0.9	0.3	CG2

Notes: The comparisons in the columns marked (1) are with respect to Takano's structure of deoxy-Mb (Takano ,1977b) and those in the columns marked (2) are with that of Phillips (1981).

Table 4: Potential H-bonding interactions for Arg 45 (CD3)

	distance in Å
Conformation A:	
45 NH2 - Heme O2D	2.9
45 NH2 - Water 259	3.2
45 NH1 - Water 259	2.3
45 NH1 - Asp 60 OD2	3.4
45 NH1 - Asp 60 OD1	3.2
Conformation B:	
45 NH1 - Water 260	3.4
45 NH2 - Water 260	2.6
45 NH2 - Heme O1D	2.2
45 NH2 - Heme O2D	3.3

Table 5(a): Temperature factors for the Distal Histidine in met-Mb and CO-Mb

The B-factors being compared are from the 1.5Å structure of CO-Mb (at 260K, this work), the 2.0Å structure of CO-Mb (255K, Kuriyan, 1985) and a new 2.0Å structure of met-Mb at 255K (Kuriyan, 1985). For the last two structures, the met-Mb data were collected first and then the crystal was converted to CO-Mb.

	met-MB (255K, 2.0Å)	CO-Mb (260K, 1.5Å)	CO-Mb (255K, 2.0Å)
	B(Å^{*2})	B(Å^{*2})	B(Å^{*2})
ATOM			
N	5.36	5.88	13.88
C_{α}	5.23	7.30	14.12
C	5.15	6.76	13.42
O	5.02	6.44	13.64
C_{β}	5.97	9.35	16.14
C_{γ}	6.04	11.02	18.40
$N_{\delta 1}$	5.12	12.11	19.66
$C_{\delta 2}$	5.55	12.02	18.94
$C_{\varepsilon 1}$	5.71	12.30	19.63
$N_{\varepsilon 2}$	5.81	12.54	19.41

Table 5(b): Sidechain B-factors with Backbone B-factors subtracted.

ATOM	met-MB (255K, 2.0Å) B(Å^2)	CO-Mb (260K, 1.5Å) B(Å^2)	CO-Mb (255K, 2.0Å) B(Å^2)
Average Backbone	5.20	6.56	13.77
Backbone Subtracted:			
C_β	0.77	2.75	2.37
C_γ	0.84	4.42	4.63
$N_{\delta 1}$	-0.08	5.52	5.89
$C_{\delta 2}$	0.35	5.42	5.17
$C_{\varepsilon 1}$	0.51	5.71	5.86
$N_{\varepsilon 2}$	0.61	5.94	5.64

Table 6: Heme geometry in various myoglobins

	deoxy	deoxy	CO	oxy	met
	Mb	Mb	Mb	Mb	Mb
	(1)	(2)	(3)	(4)	(5)
Distances: (Å)					
Fe -NA	2.07	2.00	1.99	2.00	1.99
Fe -NB	2.06	1.97	1.90	1.90	1.99
Fe -NC	2.06	2.03	1.99	1.91	2.01
Fe -ND	2.06	2.09	2.00	2.00	2.01
Fe- 93NE2	2.12	2.10	2.19	2.06	2.01
93CE1 - NA	3.50	3.20	3.10	3.32	3.36
93CD2 - NC	3.44	3.46	3.07	3.20	3.32
rms deviation of Plane 1	0.12	0.07	0.08	0.11	0.09
Fe - Plane 1	0.50	0.35	0.03	0.19	0.21
Fe - Plane 2	0.42	0.30	0.00	0.18	0.18
93 Imidazole ring centroid - heme normal	0.45	0.28	0.22	0.15	0.35
93 Imidazole ring centroid - Plane1	3.72	3.61	3.35	3.43	3.60
Angles (degrees)					
93CE1 - 93NE2 -Fe	128	130	130	130	137
93CD2 - 93NE2 -Fe	122	120	119	121	113
93NE2 - Fe - NA	102	100	87	103	88
93NE2 - Fe - NB	92	96	91	91	96
93NE2 - Fe - NC	102	100	93	94	101
93NE2 - Fe - ND	111	100	89	94	95
Dihedral (degrees)					
93CD2-93NE2-Fe-NC	24	2	-4	2	2

Notes: deoxy-Mb (1) is the structure of Takano (1977b) and deoxy-Mb (2) is the structure of Phillips (1981). Plane 1 refers to the mean plane of the 24 porphyrin atoms and Plane 2 refers to that of the four pyrrole nitrogens alone.

Table 7. Backbone positional shifts by secondary structure

The final refined structure of CO-Mb is compared with the structure of deoxy-Mb (Takano, 1977b). Only the backbone atoms (N,C and C_{α}) are included in comparison. The helices and loops are defined by the following residue ranges: A helix (3-18), AB loop (19), B helix (20-35), C helix (36-42), D helix (51-57), E helix (58-77), EF loop (78-85), F helix (86-94), FG loop (95-99), G helix (100-118), GH loop (119-124) and the H helix (125-148). Each column in the Table lists the deviations for the loops, helices, the distal part of the molecule, the proximal part of the molecule and the whole molecule (excluding the terminal regions) for superimposing backbone atoms in each segment individually (labelled "Each"), in the whole structure ("4-148"), in the A, B, C and G helices only ("A-B-C-G"), in the B, C and G helices only ("B-C-G") and in the distal part of the molecule only ("distal"). The proximal and distal regions of the molecule are defined with repect to the mean heme plane.

	Each	4-148	A-B-C-G	B-C-G	distal
	Å	Å	Å	Å	Å
4-148	0.29	0.29	0.30	0.33	0.31
distal	0.29	0.30	0.31	0.29	0.29
proximal	0.23	0.28	0.28	0.32	0.36
A	0.18	0.21	0.21	0.31	0.23
AB	0.05	0.46	0.50	0.60	0.47
B	0.17	0.20	0.18	0.18	0.19
C	0.14	0.20	0.20	0.20	0.21
CD	0.22	0.35	0.33	0.39	0.33
D	0.15	0.25	0.24	0.24	0.23
E	0.19	0.22	0.25	0.28	0.23
EF	0.17	0.26	0.29	0.39	0.34
F	0.17	0.40	0.46	0.48	0.49
FG	0.27	0.38	0.42	0.43	0.47
G	0.17	0.24	0.23	0.20	0.23
GH	0.33	0.74	0.71	0.62	0.73
H	0.22	0.23	0.26	0.32	0.31

FIGURE LEGENDS

Figure 1

Electron density for Arg 45 CD3. The density shown is from a difference Fourier (F_o-F_c) map calculated without any contribution from Arg 45 CD3. The final refined conformations for the residue are also shown in the figure. 1a. Electron density at 4σ (0.4e/Å^3) above the average, where σ is the standard deviation of the map. 1b. Electron density at 2σ (0.2e/Å^3) above the average.

Figure 2

The heme group, and deviations from planarity. The nomenclature used for the atoms of the heme are shown. The deviations of atoms which deviate by more than 0.1Å from the plane of the 24 central porphrin atoms (not including the iron) in CO-Mb are indicated. The figures in parentheses are the deviations for deoxy-myoglobin. The dihedral angle ϕ used to describe the CO conformations are shown in the figure.

Figure 3

Electron density for CO. 3a. Density at 5.5σ (0.55e/Å^3). 3b. Density at 3.0σ (0.3e/Å^3). 3c. Density at 3.0σ, orthogonal view. The two final conformations for the CO are shown in the Figures.

Figure 4

(a) Overlap of calculated and observed density for the CO ligand as a function of ϕ and θ. The observed density was calculated using a F_o-F_c synthesis, ommitting the CO atoms from the calculation. To determine which values of ϕ and θ best fit this density, atomic coordinates for the ligand were generated for values of ϕ between 0° and 360° and θ between 90° and 180°, using an Fe-C bond-length of 1.92Å and a C-O bond length of 1.17Å. Electron density was calculated from these coordinates, using the formula given by Ten Eyck (1977). For each conformation, the overlap was estimated by summing the product of the observed and calculated density over all the grid points in a 10ÅX10ÅX10Å box around the ligand carbon atom. The contours in the Figure are at different values of the overlap, expressed relative to the minimum overlap. Dashed lines: Density overlap at 1.25, 1.50, 1.75 and 2.0 times the minimum overlap. Dot-Dashed lines: Density overlap at 2.25, 2.5 and 2.75 times the minimum overlap. Solid lines: Density overlap at 3.0, 3.25, 3.50, 3.75, 4.0, 4.25, 4.50 and 5.0 times the minimum overlap. The two crosses mark the final refined conformations (C and D, see text).

(b) θ - ϕ energy map for CO in the CO-Mb protein structure.

bending and rotation. CO structures were generated as described above and van der waals interaction energies between the ligand and the protein were calculated using the program CHARMM (Brooks et al.,1983). There are ten energy contours between -5 Kcal/mole and 8.5 Kcal/mole (solid lines) and 10 between 10 Kcal/mole and 23.5 kcal/mole (dashed line).

(c) $\theta - \phi$ energy map for CO in the deoxy-Mb protein structure. CO structures were generated as above, except that the center of the porphyrin ring (rather than the Fe atom) was used as the origin. The protein structure used was that of Takano (1977). Energy contours are as in Fig 4(b), above.

Fig. 5 (a,b,c) Energy maps for oxygen in a plane 2.5Å above the heme plane. There are 10 energy levels from -5.0 Kcal/mole to 30.0 kcal/mole. 5a. Deoxy-myoglobin energy map. 5b. CO-myoglobin energy map. In Fig. 5a. and 5b., the positions of the four refined conformations of the oxygen atom are marked with an X. The sidechains of His 64 E7 and Val 68 E11 are also shown. 5c. CO-myoglobin energy map with χ_1 and χ_2 of His 64 E7 changed to 280° and 275°, respectively, corresponding to the new minimum in Fig. 6b. The ring of the distal histidine is now out of the range of the map. The backbone atoms of the histidine, which are more that 9Å above the plane of the heme, are shown for reference.

Figure 6

The heme group, CO, the distal histidine and the proximal histidine. Comparison of structures in deoxy- (thin line) and CO-myoglobin (thick line). The structures were superimposed using all the backbone atoms.

Figure 7

$\chi_1 - \chi_2$ energy map for His 64 E7. Total energy of the residue is calculated as a function of χ_1 and χ_2 with the rest of the protein kept fixed. There are 10 contour levels from -10.0 to 100.0 Kcal/mole. 6a. Arg 45 CD3 is in conformation A. The X-ray conformation on His 64 E 7 is marked with an X. 6b. Arg 45 CD3 is in conformation B. The conformation of His 64 E7 that was used in calculating the energy map in Fig. 4d is marked with an X.

Figure 8

Two views of the two conformations for ARG 45 CD3.

Figure 9

Backbone deviations between deoxy- and CO-myoglobin. The two molecules are superimposed on all the backbone atoms. The deviations are averaged over the backbone N, C, C_α atoms.

Figure 10

Histograms of shifts between CO- and deoxy-myoglobin. The two molecules are superimposed on the B, C and G helices and the component of the shift in the direction of the heme group (in CO-myoglobin) is calculated for every atom on the proximal side of the heme group. A negative shift indicates a motion towards the heme group on going from deoxy to CO-myoglobin. 9a. Backbone atoms only. 9b. All atoms.

Figure 11

C_αΔ-distance matrices at the 0.25Å level. (a) CO-Mb vs. deoxy-Mb (Takano, 1977b) (b) CO-Mb vs. deoxy-Mb (Phillips, 1981) (d) CO-Mb vs. met-Mb (Kuriyan, 1985).

FIGURE 1 (a)

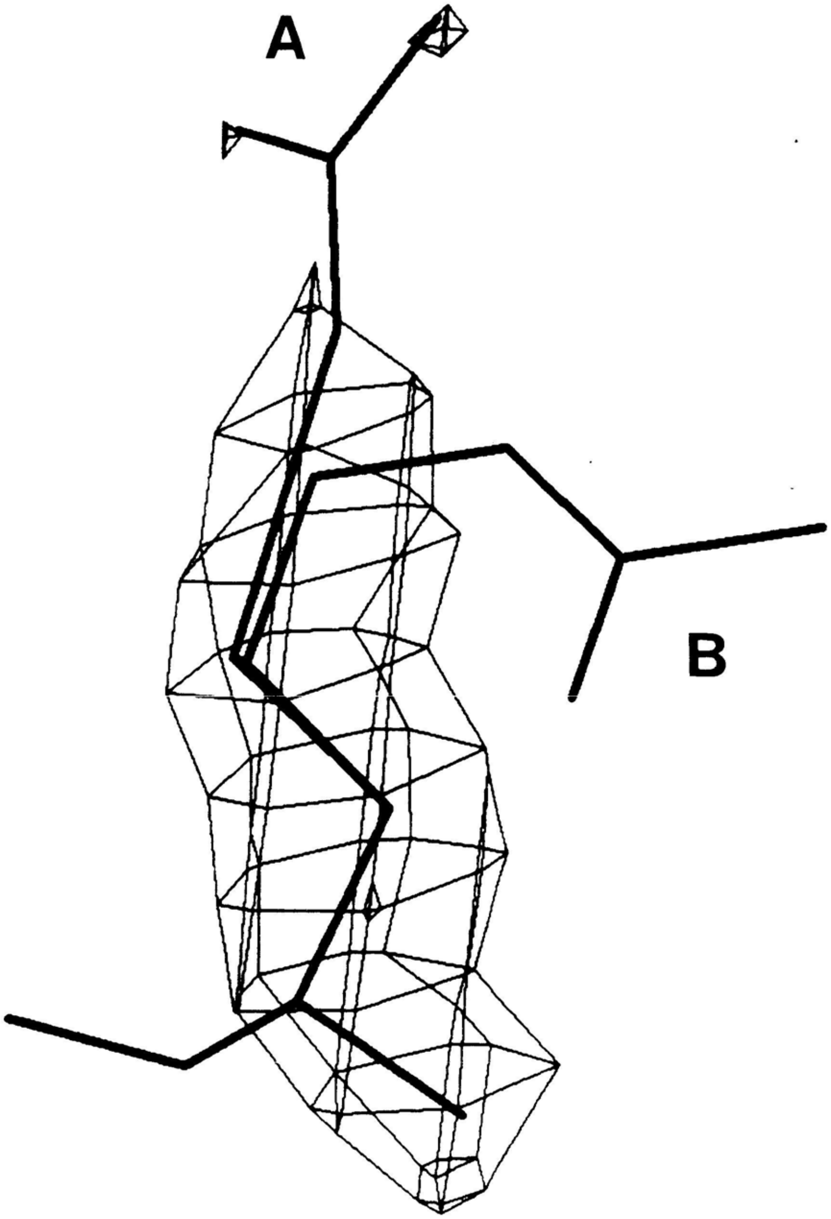


FIGURE 1(b)

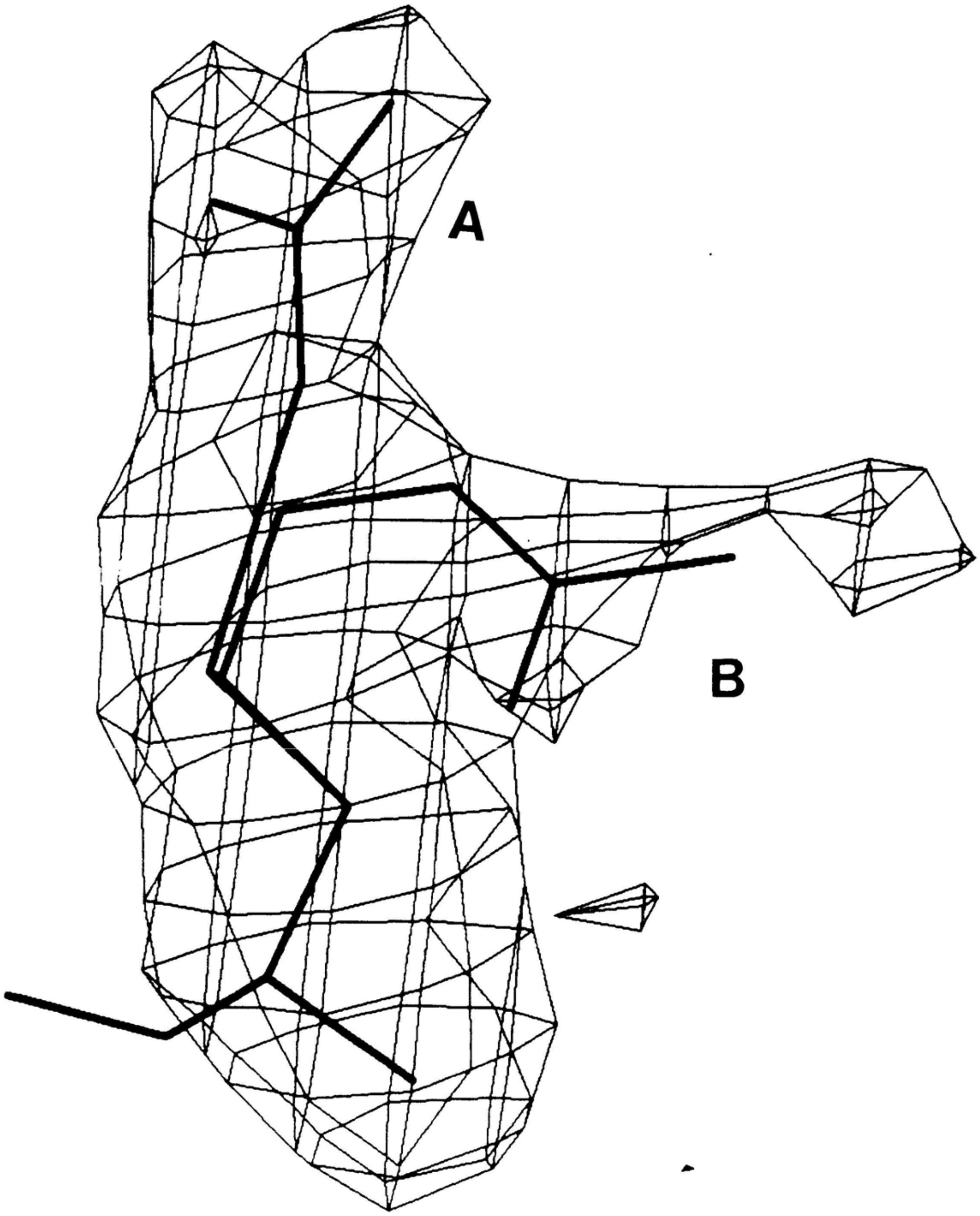


FIGURE 2

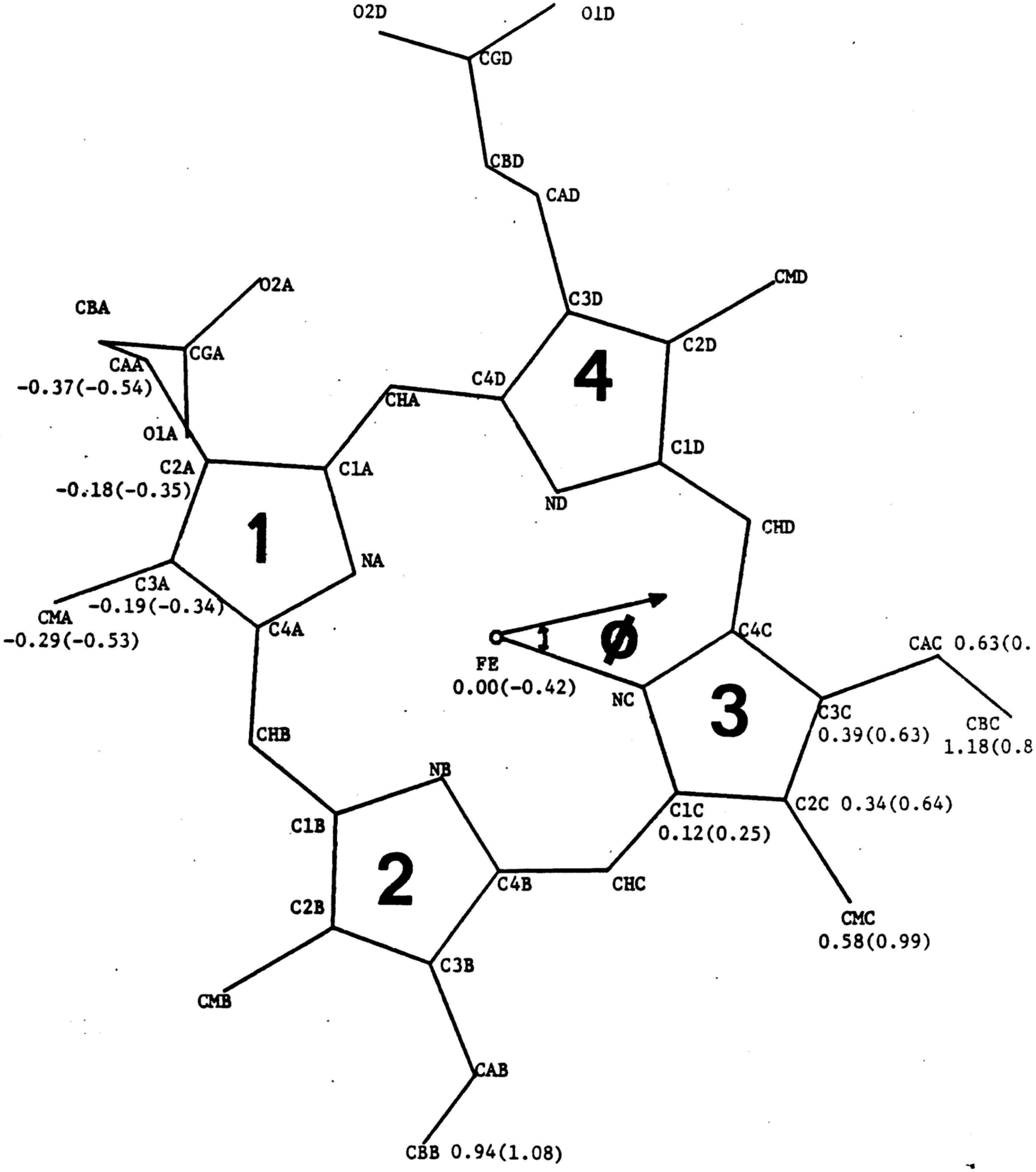

O2D
O1D
CGD
CBD
CAD
CMD
C3D
C2D
4
C4D
CHA
C1D
ND
CHD
O2A
CBA
CGA
CAA
-0.37(-0.54)
O1A
C2A
-0.18(-0.35)
C1A
1
NA
C3A
-0.19(-0.34)
CMA
-0.29(-0.53)
C4A
Ø
FE
0.00(-0.42)
NC
C4C
3
CAC 0.63(0.
C3C
0.39(0.63)
CBC
1.18(0.8
CHB
NB
C1B
C1C
0.12(0.25)
C2C 0.34(0.64)
2
C4B
CHC
C2B
C3B
CMC
0.58(0.99)
CMB
CAB
CBB 0.94(1.08)

FIGURE 3 (a)

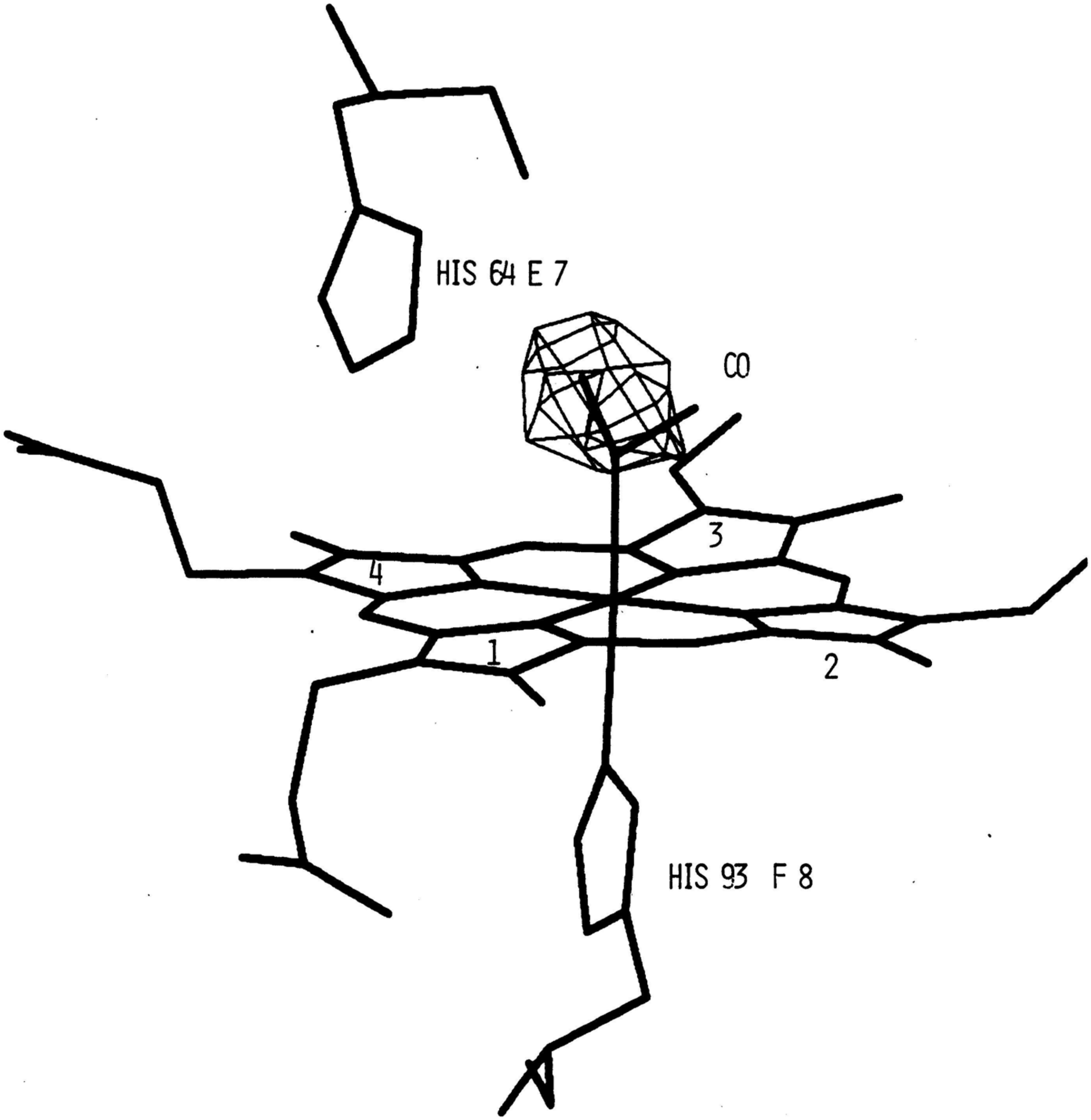


FIGURE 3 (b)

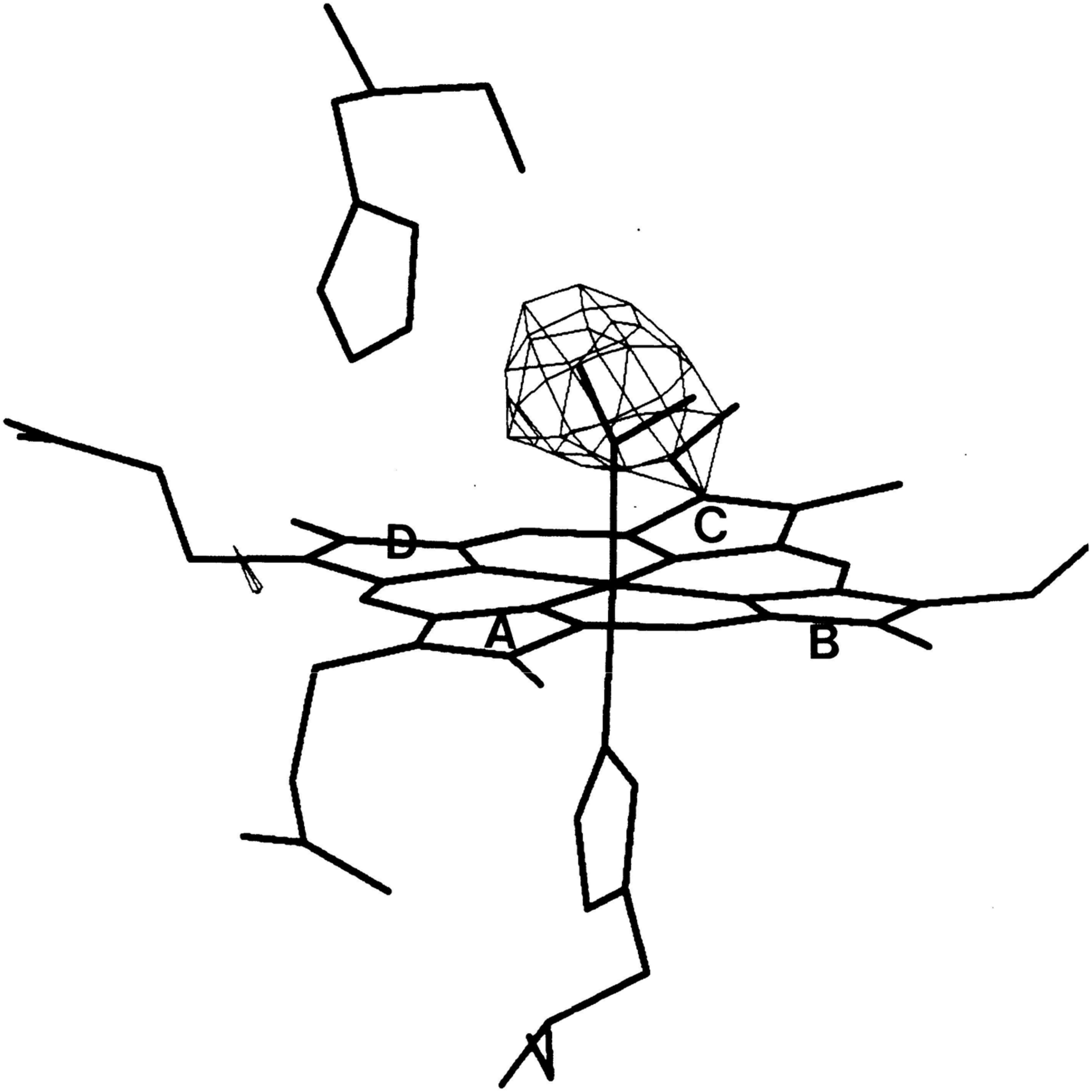


FIGURE 3 (c)

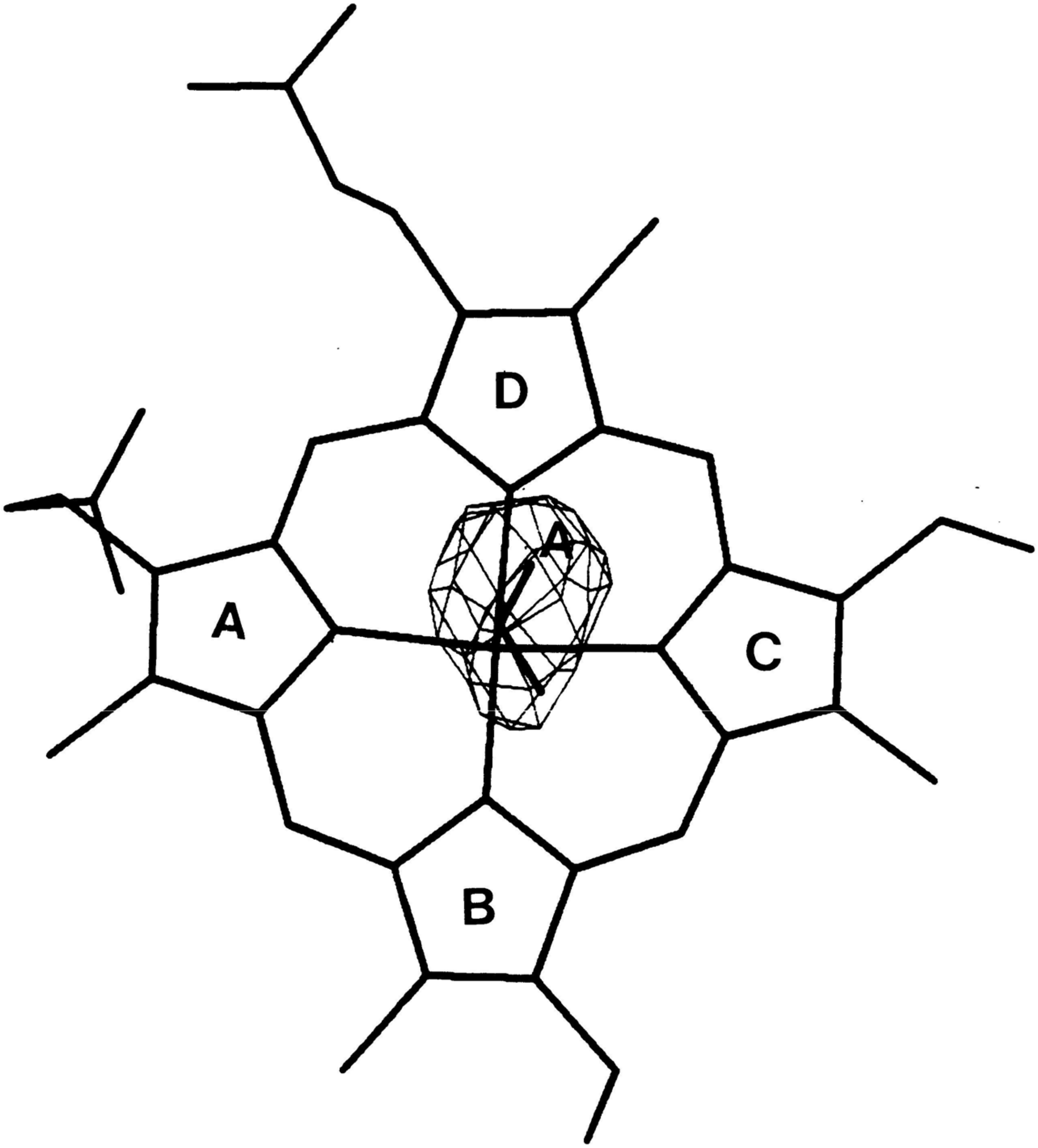
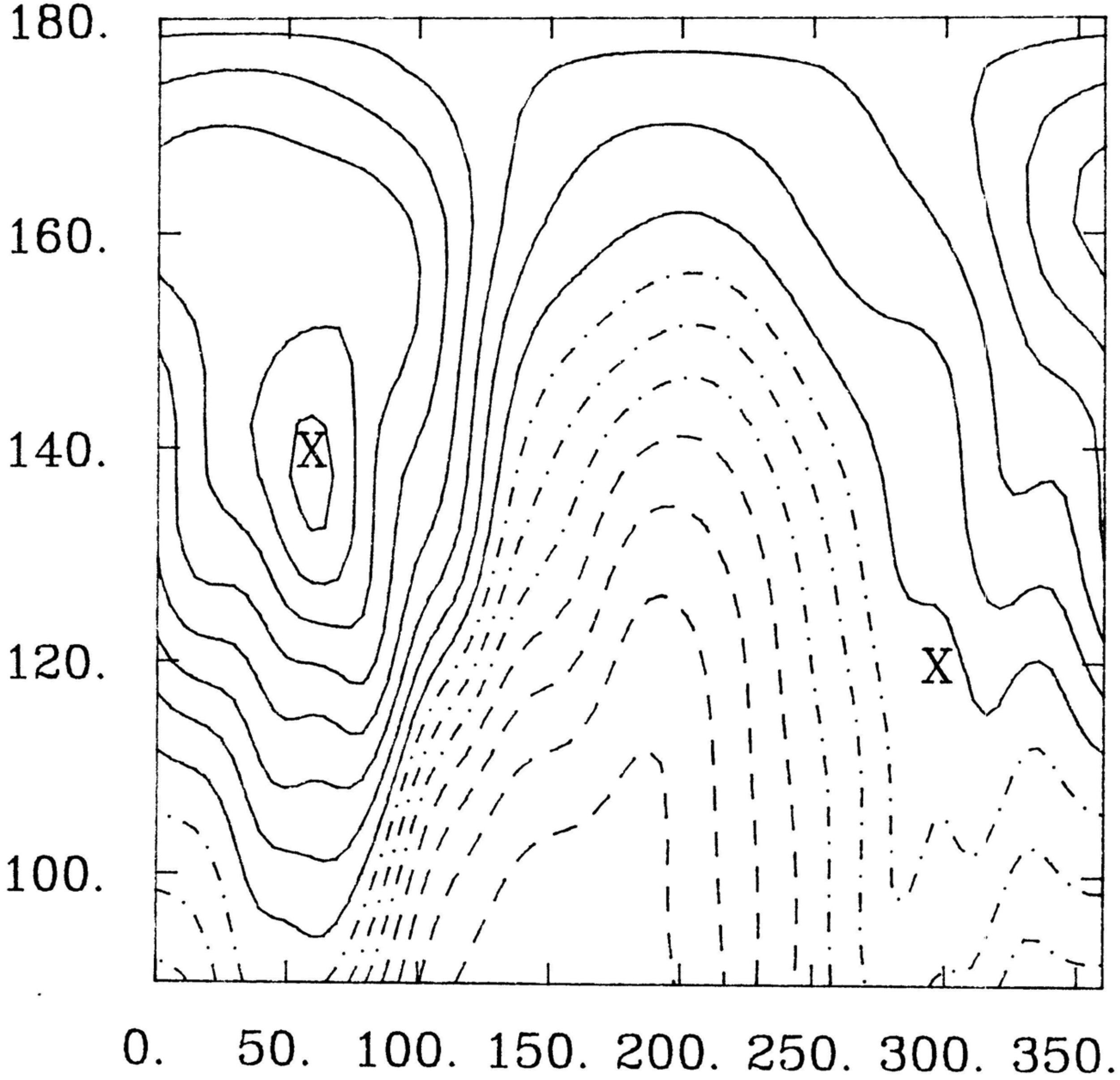


FIGURE 4(a)



CO DIFFERENCE—MAP DENSITY OVERLAP

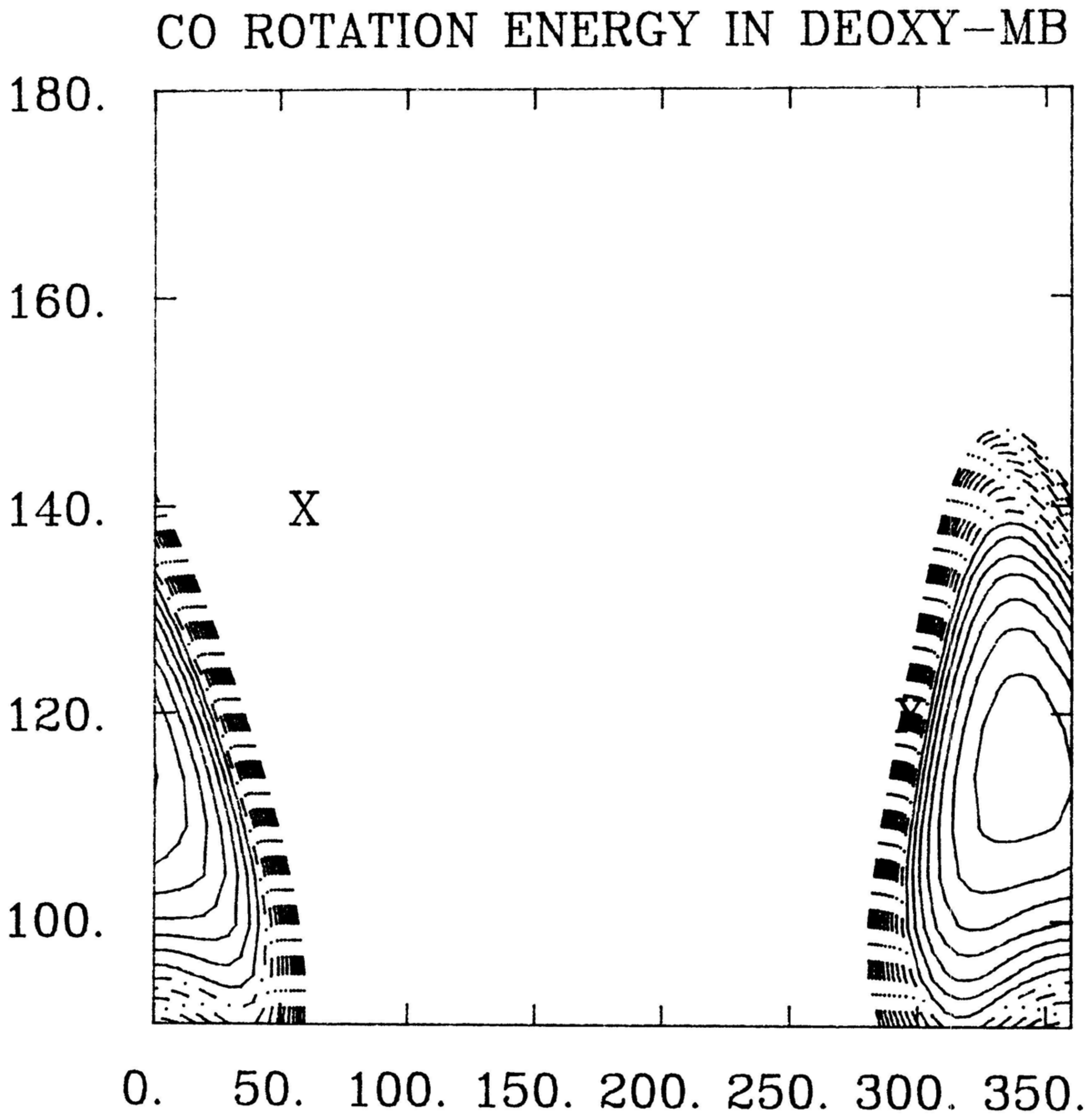

CO ROTATION ENERGY IN DEOXY-MB
180.
160.
140.
120.
100.
0.
50.
100.
150.
200.
250.
300.
350.
X
X

FIGURE 4(b)

FIGURE 4(c)

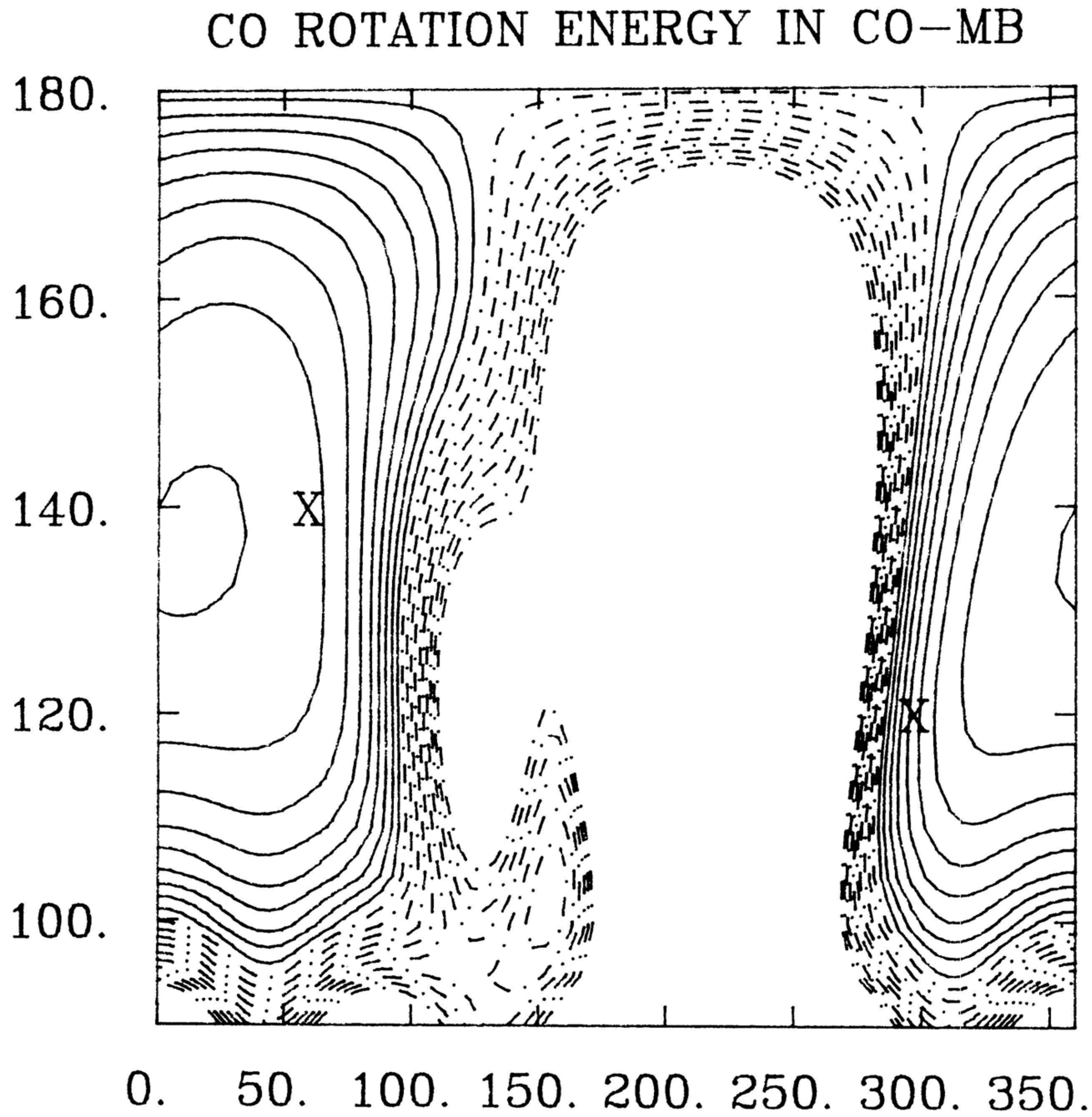

CO ROTATION ENERGY IN CO—MB
180.
160.
140.
120.
100.
X
X
0.
50.
100.
150.
200.
250.
300.
350.

FIGURE 5(a)

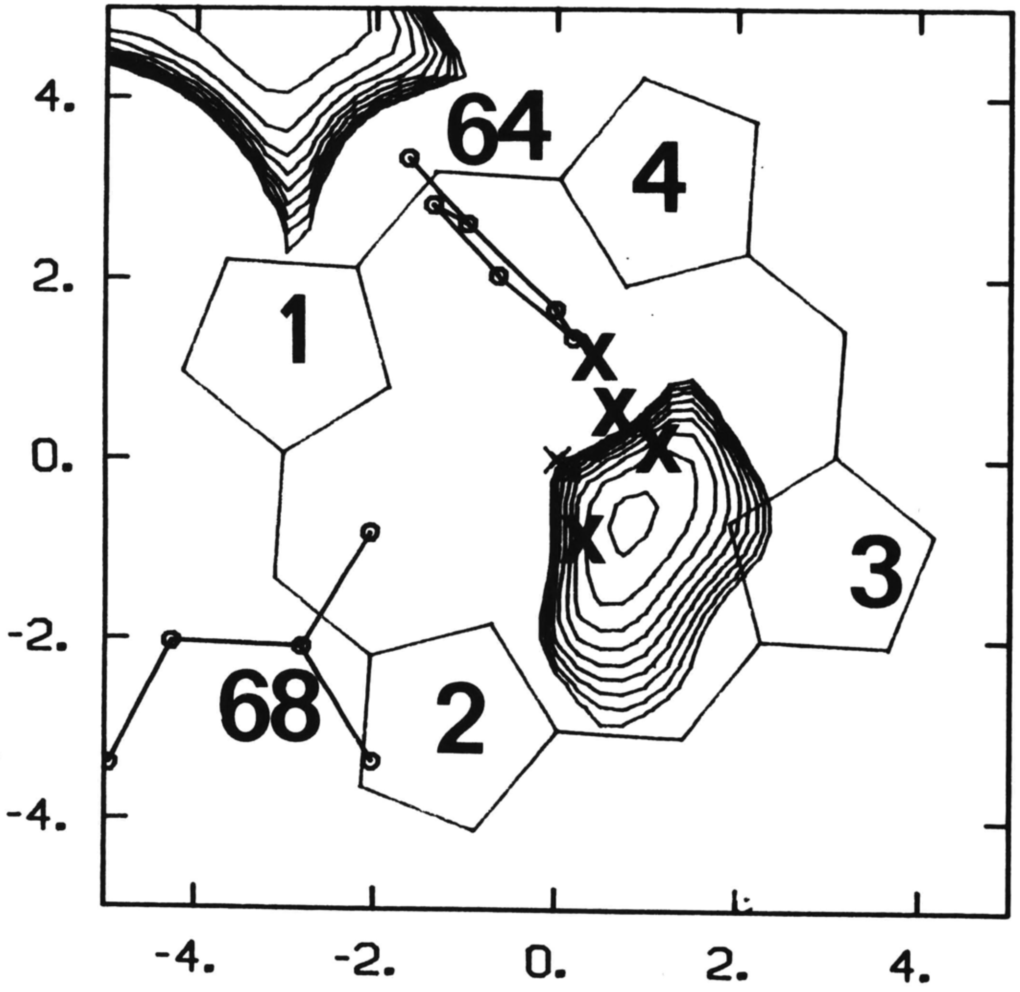


FIGURE 5(b)

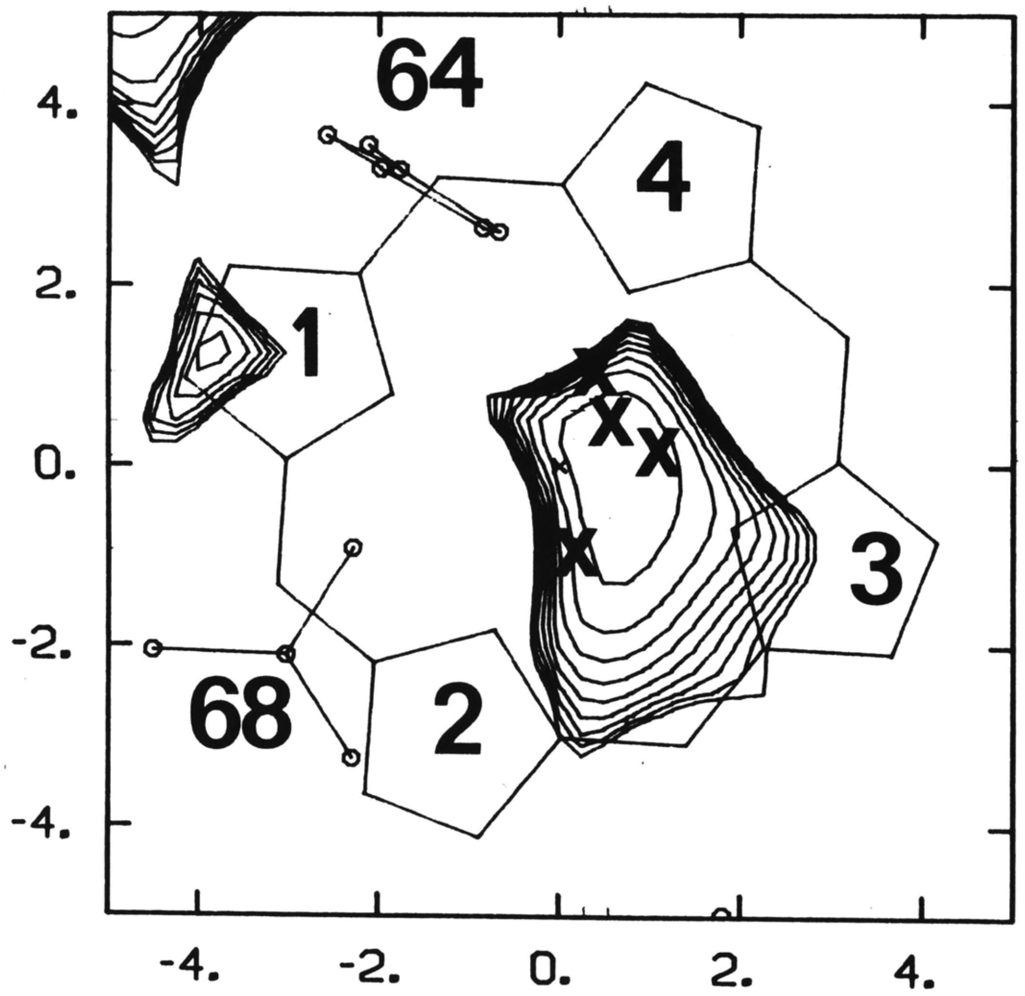


FIGURE 5(c)

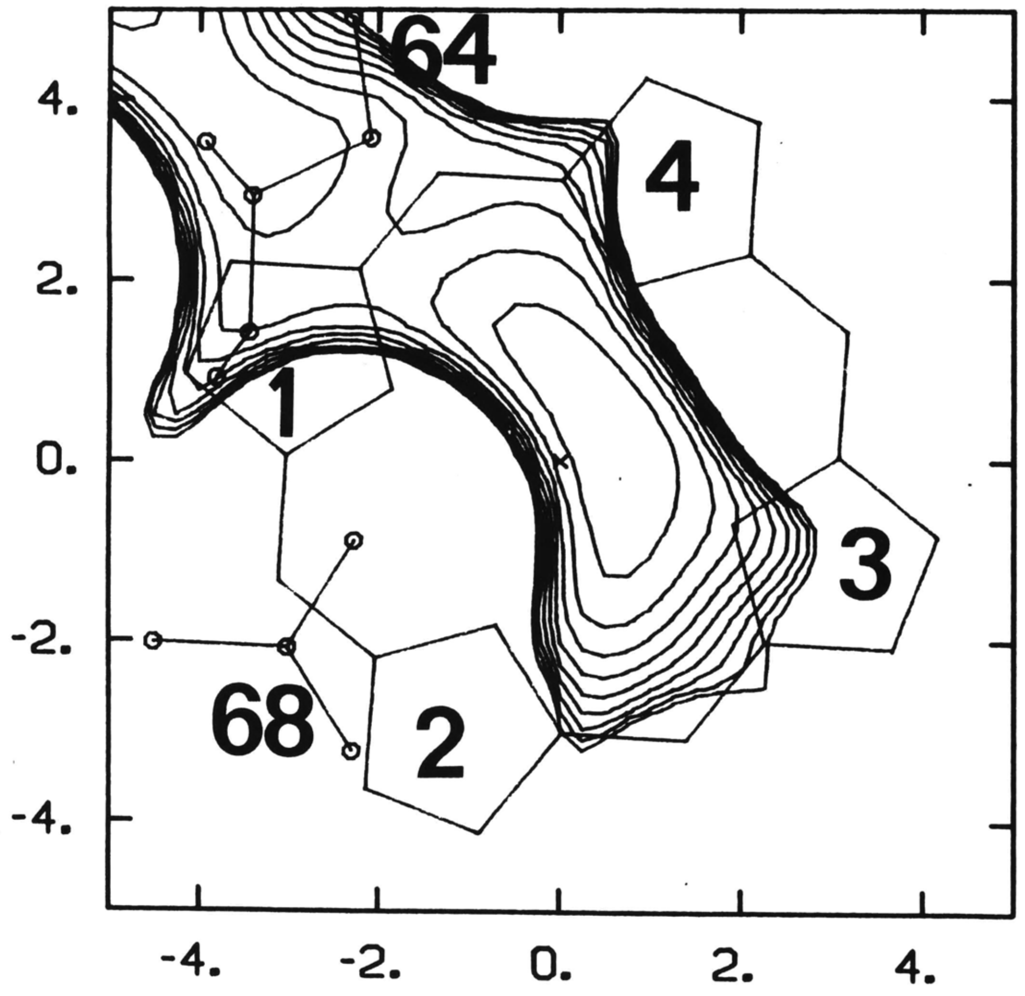


FIGURE 6

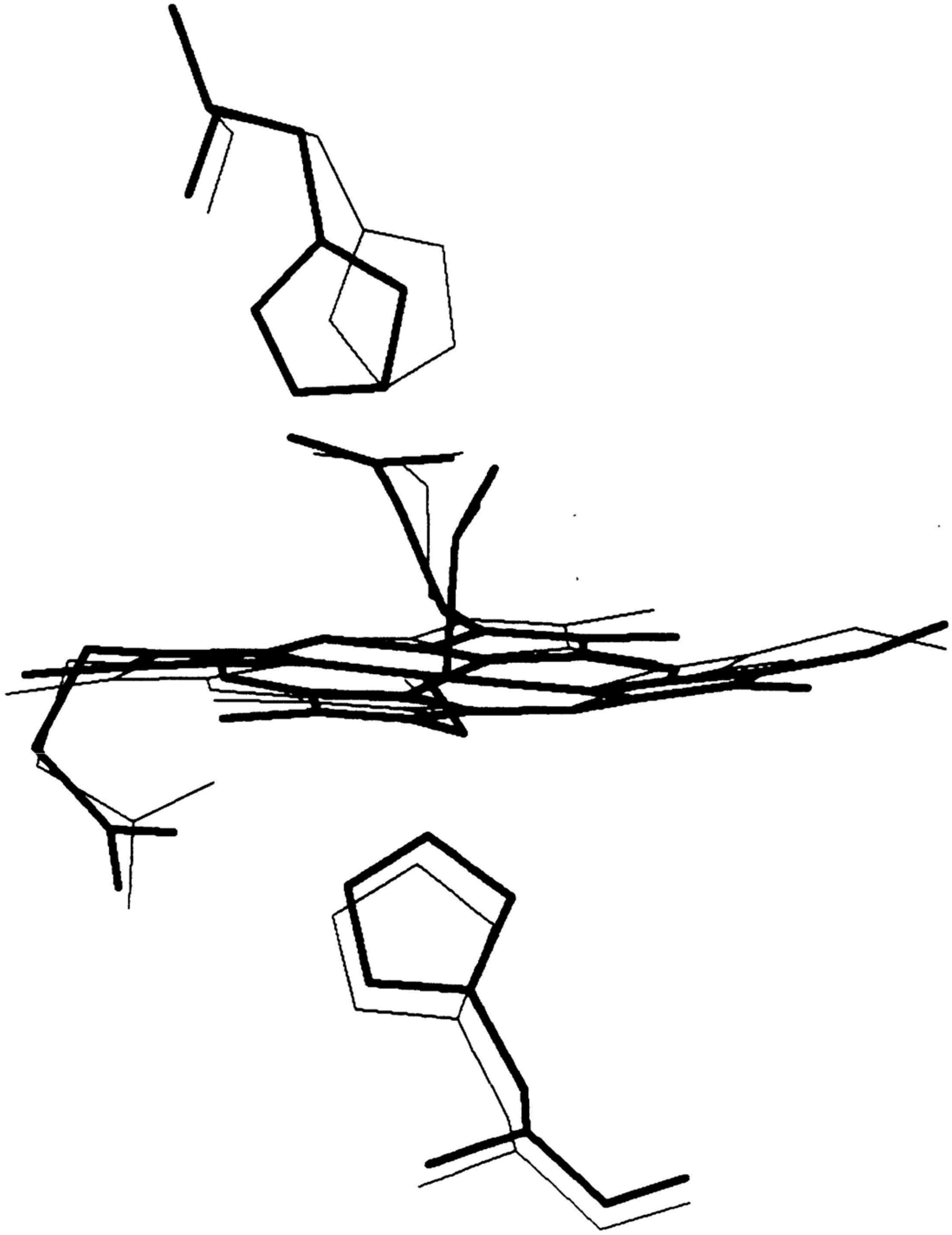


FIGURE 7(a)

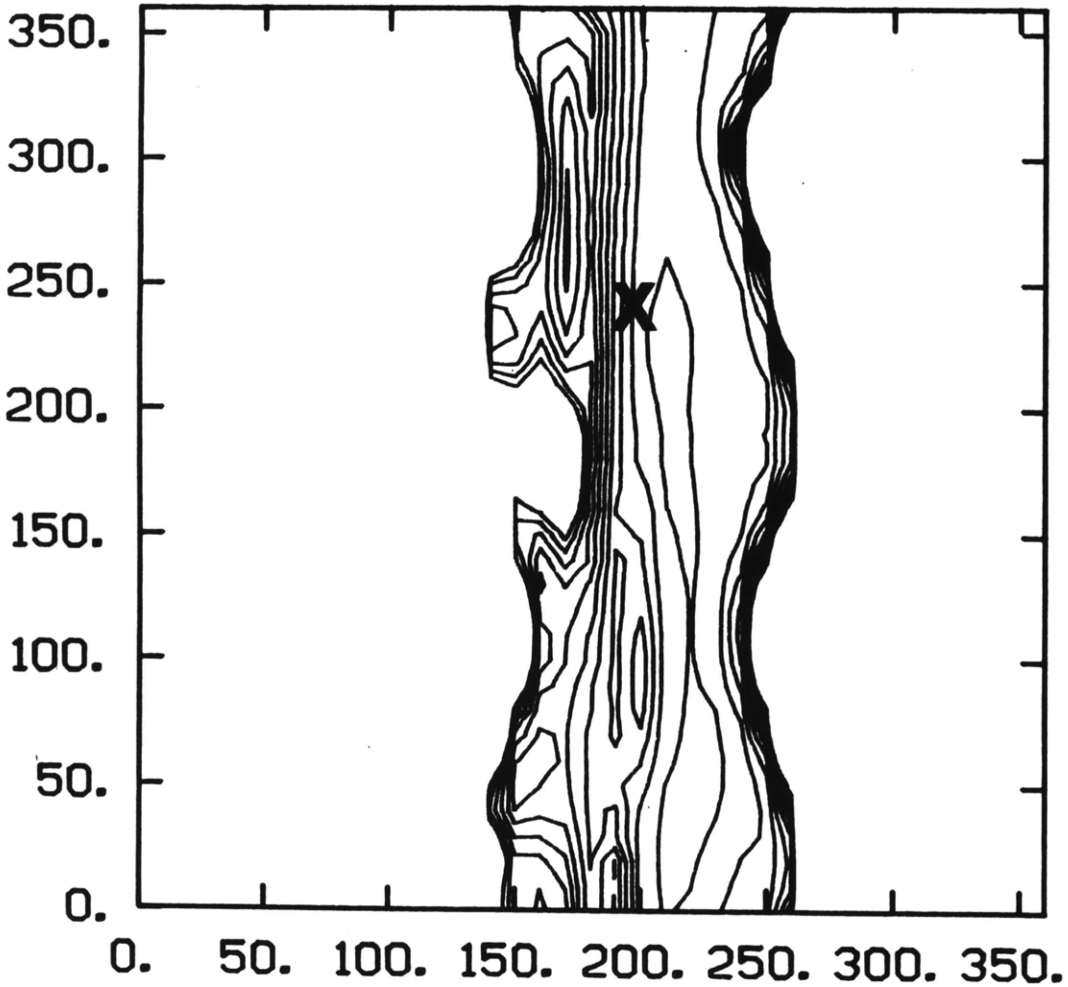


FIGURE 7(b)

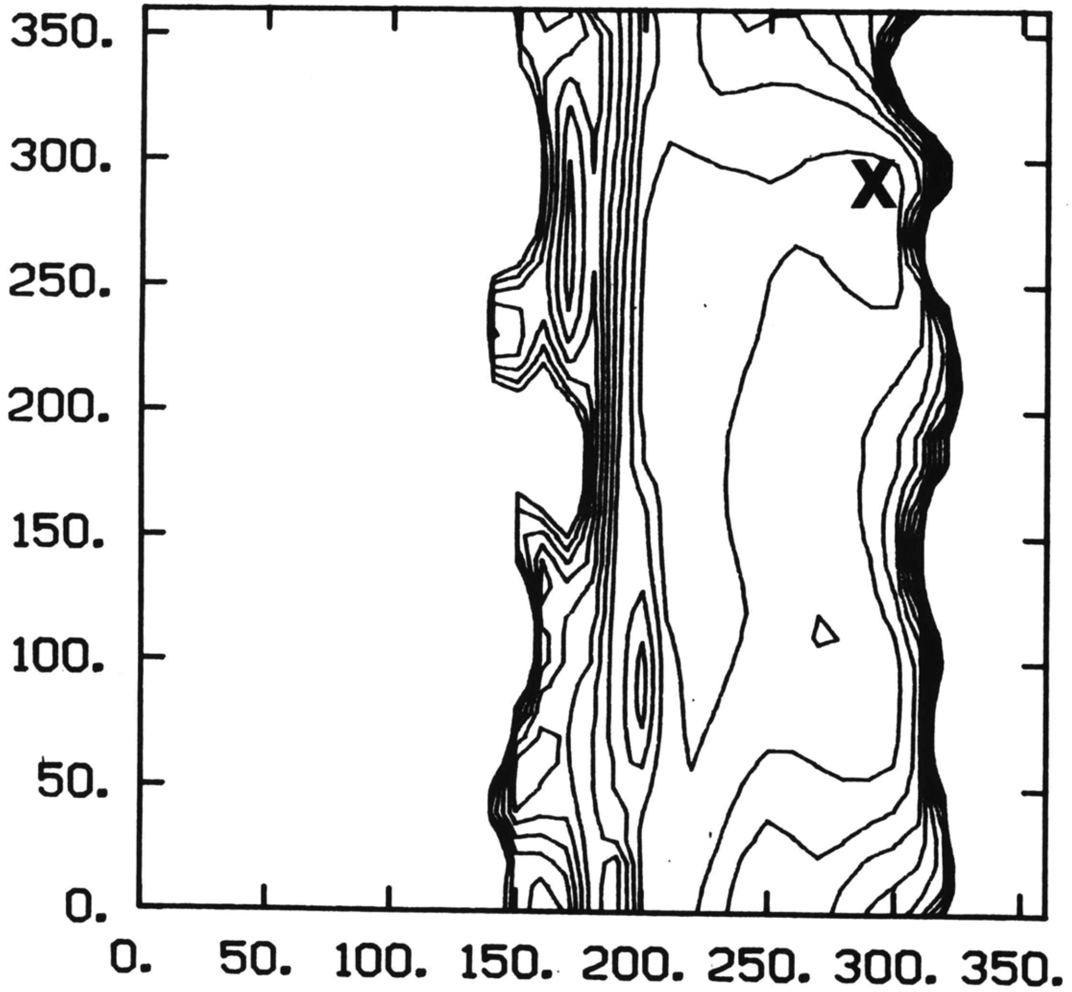


FIGURE 8

Stereo pairs of residues surrounding Arg 45 (CD3) in conformation A (Fig. 8a) and conformation B (Fig. 8b). Only one conformation of the CO (referred to as conformation C in the text) is shown for clarity.

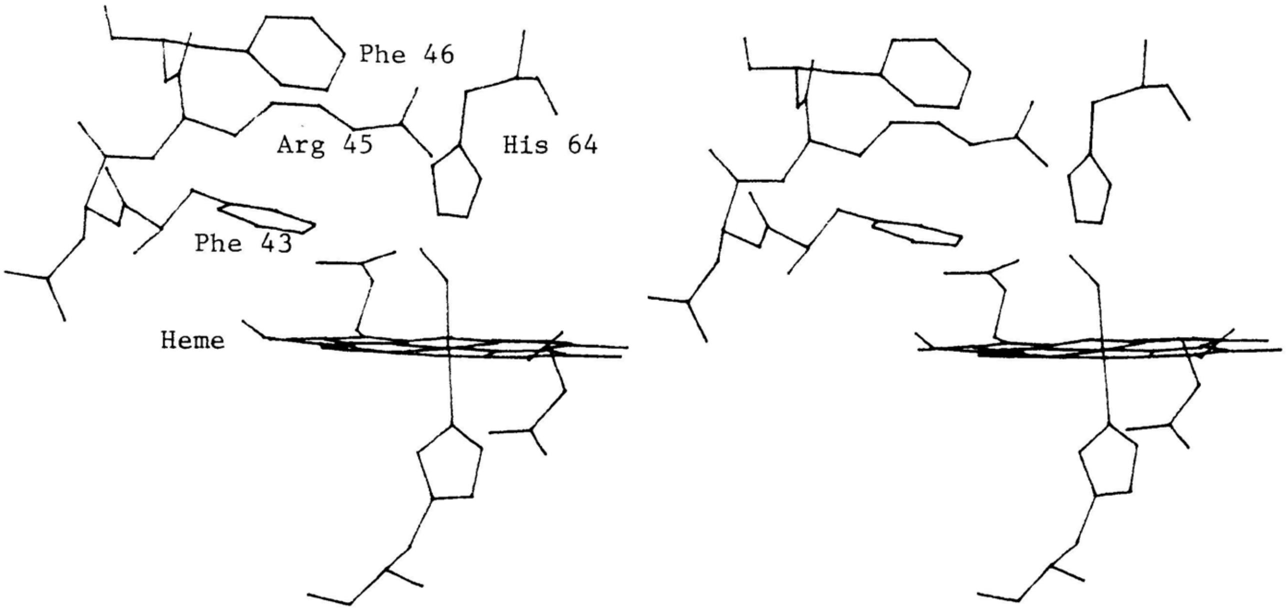


Figure 8(a)

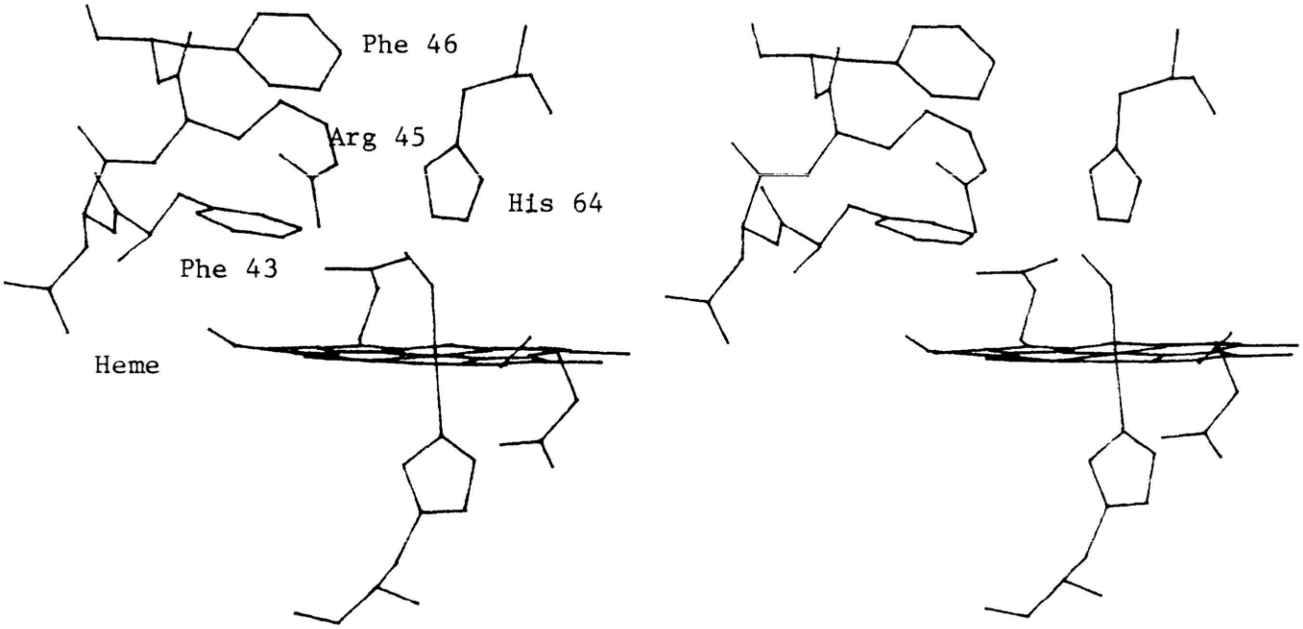


Figure 8(b)

FIGURE 9

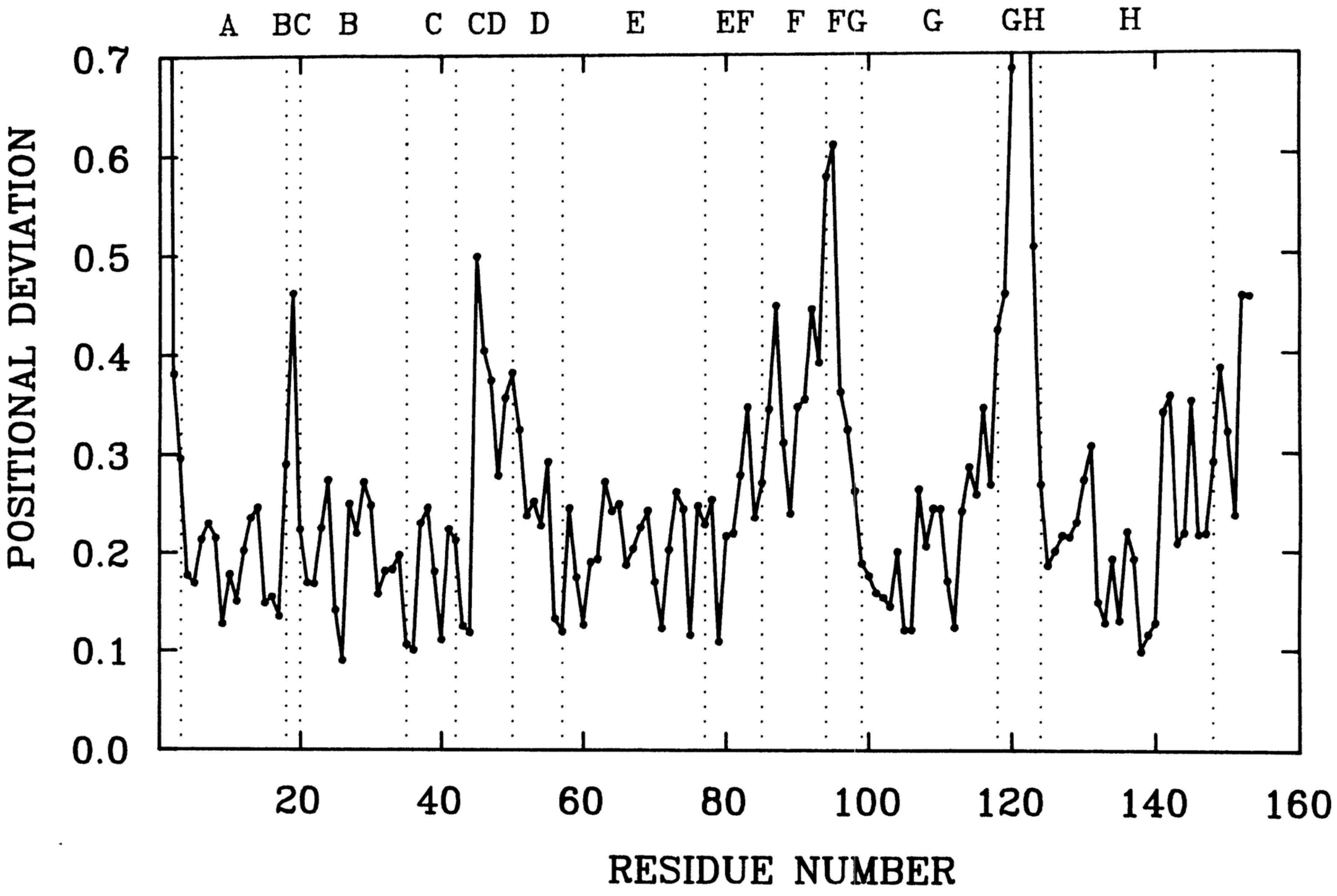

A
BC
B
C
CD
D
E
EF
F
FG
G
GH
H
POSITIONAL DEVIATION
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
RESIDUE NUMBER
20
40
60
80
100
120
140
160

FIGURE 10(a)

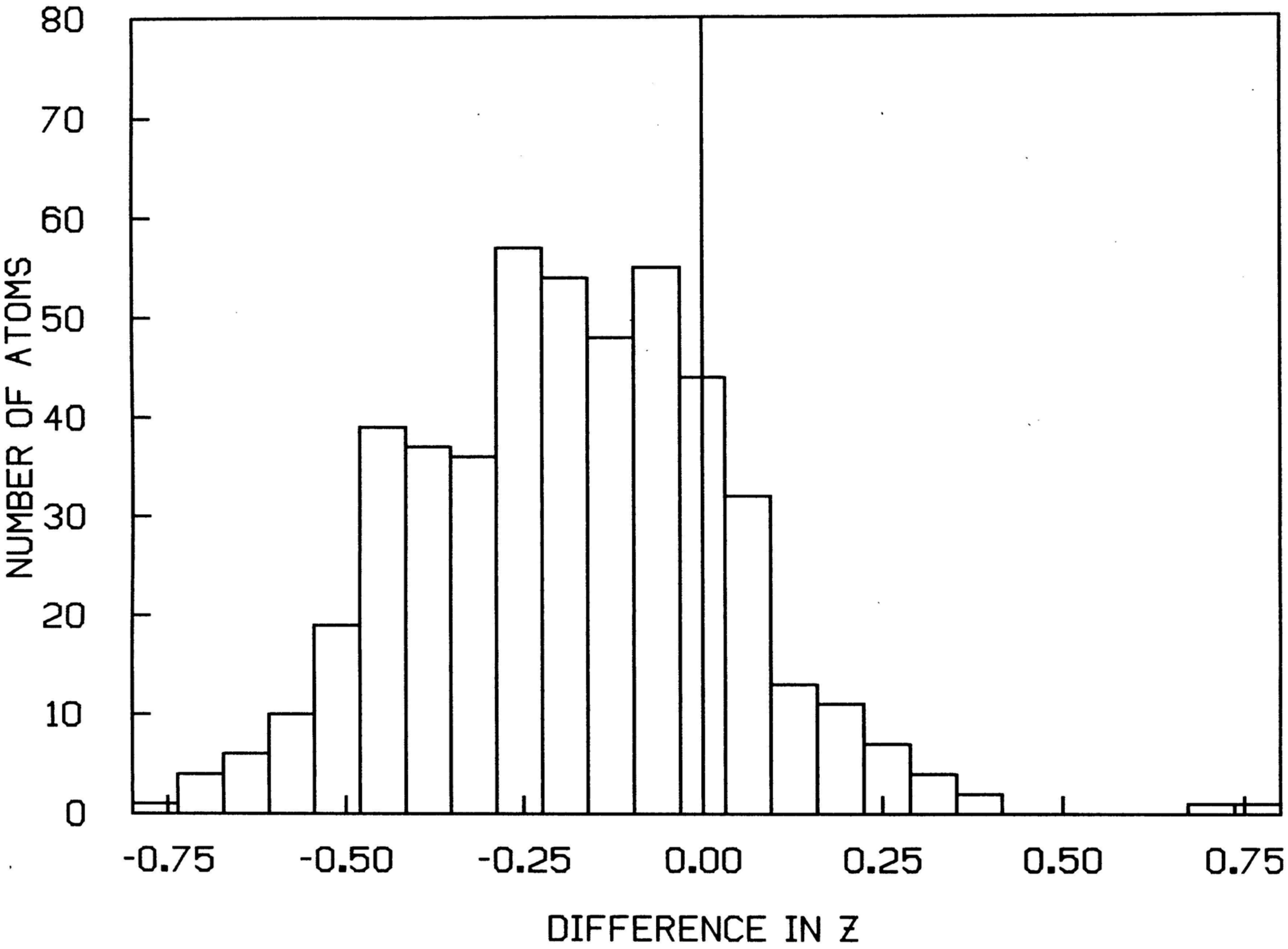
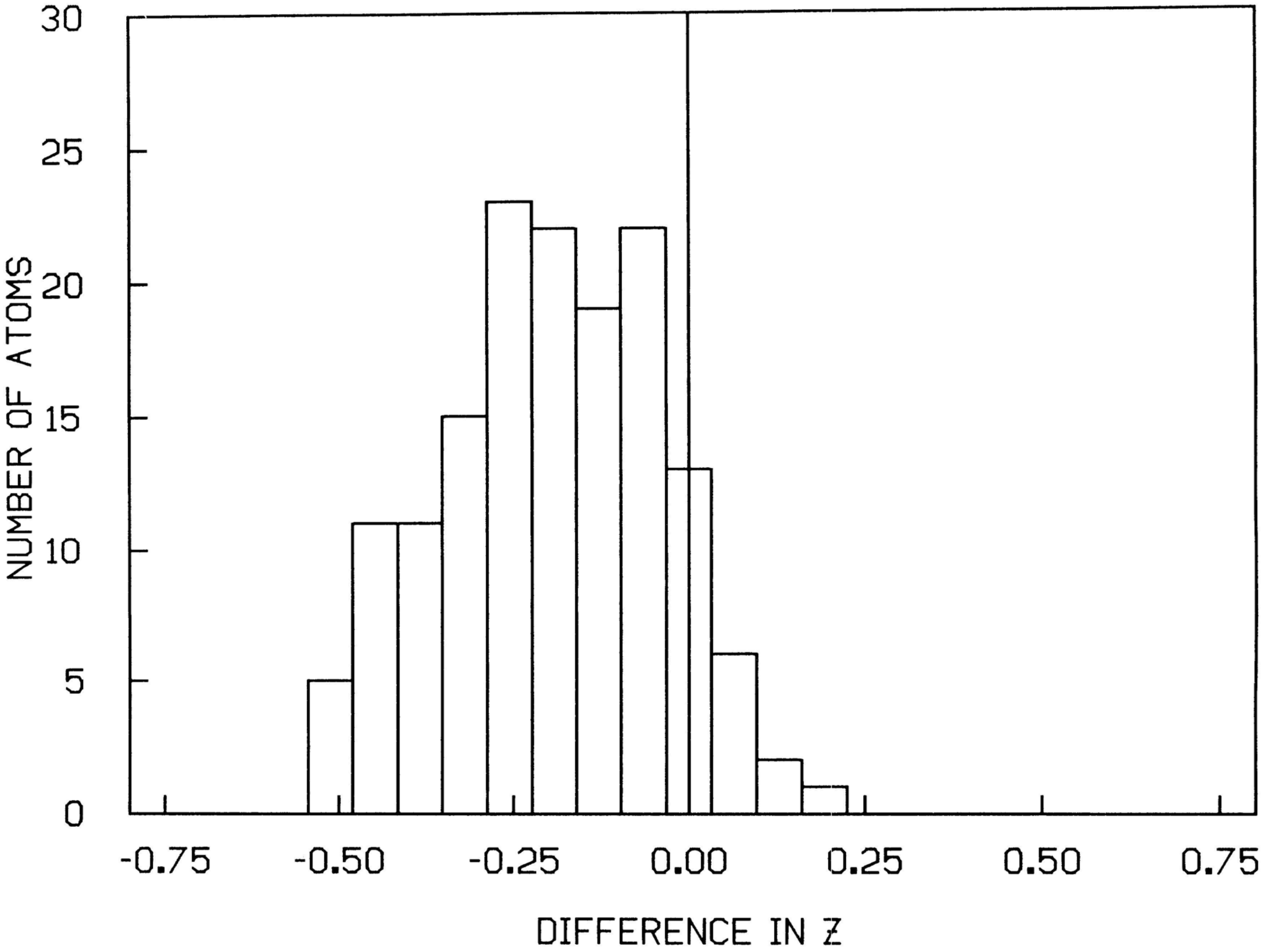
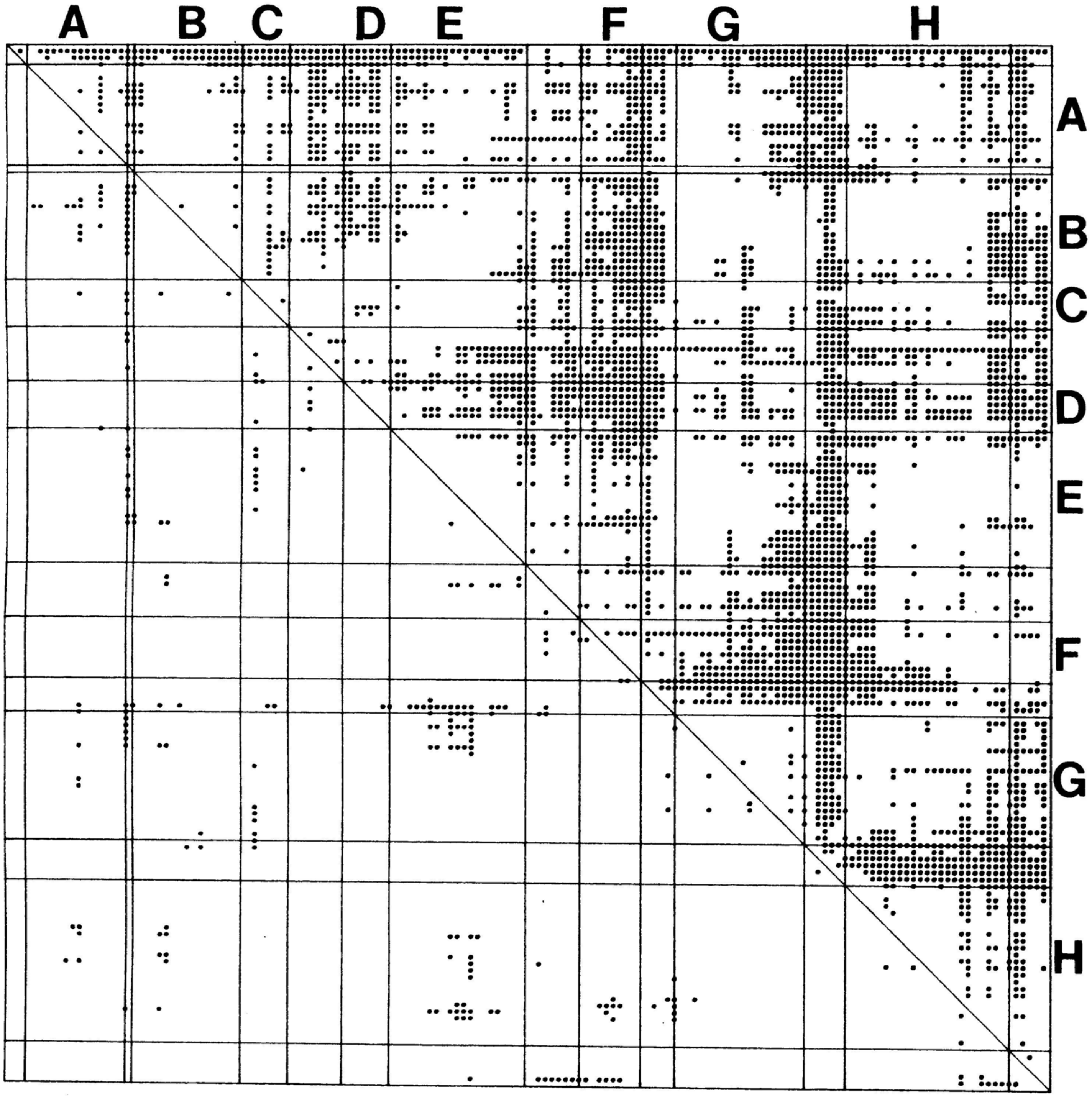

NUMBER OF ATOMS
80
70
60
50
40
30
20
10
0
-0.75
-0.50
-0.25
0.00
0.25
0.50
0.75
DIFFERENCE IN Z

FIGURE 10(b)


30
25
20
15
10
5
0
NUMBER OF ATOMS
-0.75
-0.50
-0.25
0.00
0.25
0.50
0.75
DIFFERENCE IN Z



CO/deoxy
Takano

FIGURE 11(a)

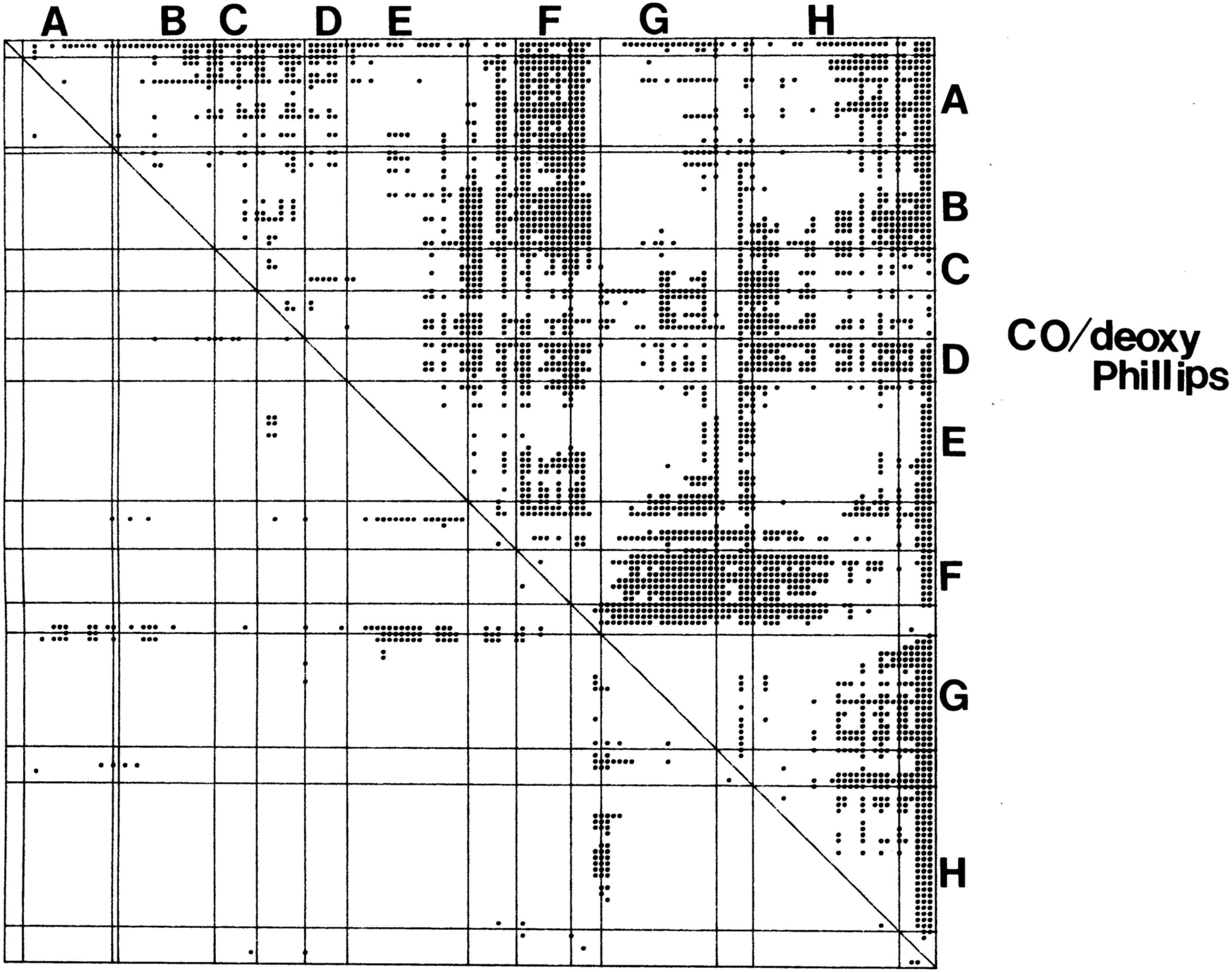
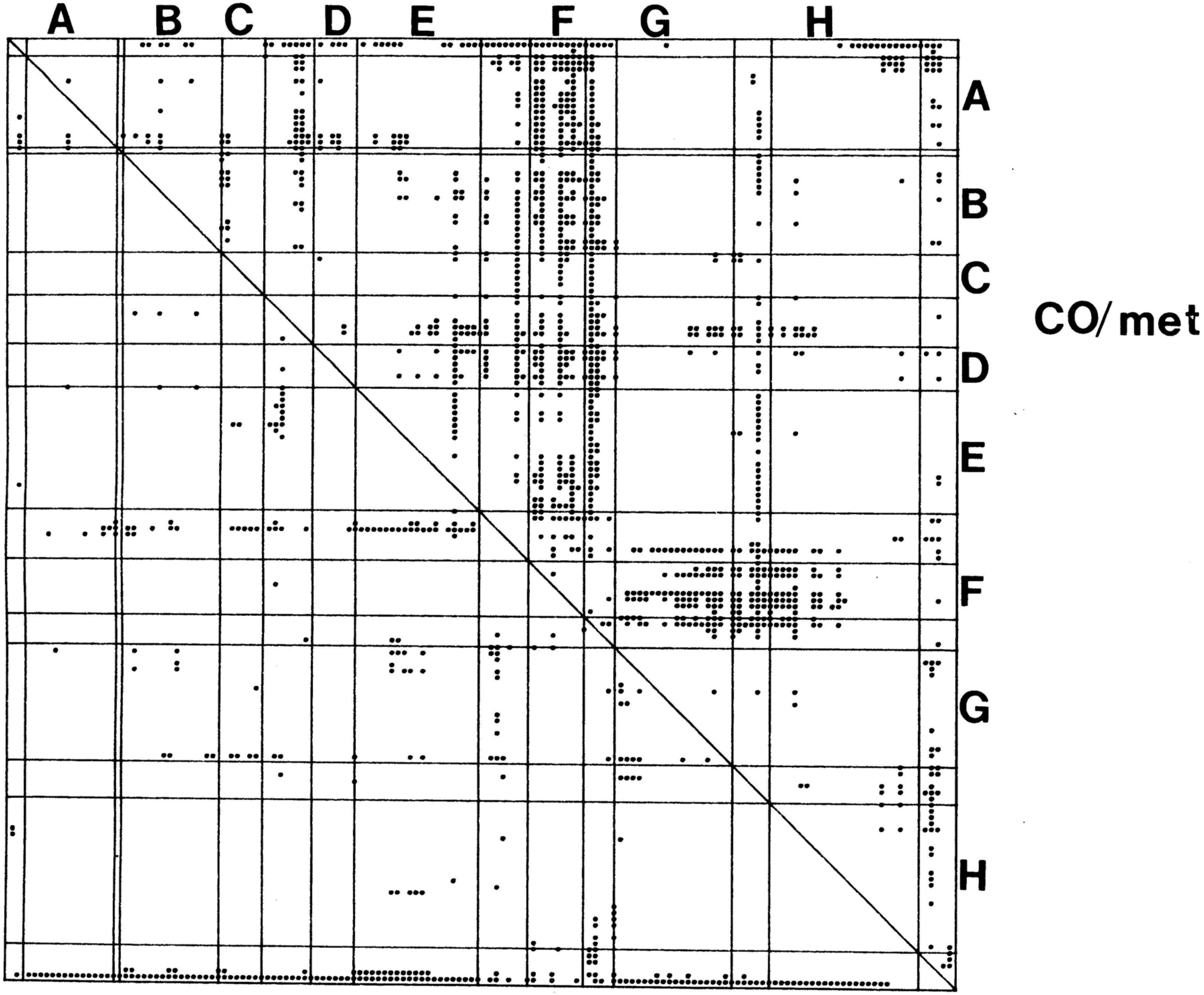


FIGURE 11(b)

Figure 11(c)



Chapter 5

The X-ray Structure and Refinement of CO-Myoglobin: Errors in the Structure and B-Factors

Abstract

The errors in the refined parameters for the 1.5Å X-ray structure of CO-myoglobin are estimated by combining energy minimization with least-squares refinement against the X-ray data. The minimization provides perturbed structures which are used to re-start X-ray refinement. The resulting structures have the same R-factor and stereochemical parameters as the original X-ray structure, but deviate from it by 0.13Å rms for the backbone atoms and 0.31Å rms for the sidechain atoms. The error in the B-factors is estimated to be 15%. This technique also results in structures that have a lower energy than those obtained by X-ray refinement alone. Comparison of the B-factors in different crystal structures of myoglobin shows that lattice disorder can lead to large (35% to 40%) differences in the B-factors. These differences cannot be accounted for by translational lattice disorder alone; a component which might be ascribed to rotational disorder causes the disorder contribution to change with the magnitude of the B-factor.

(I) Introduction

In this chapter the reproducibility of the refined coordinates and temperature factors (B-factors) of CO-Mb[1] at 1.5Å are examined by a method which involves energy minimization of the refined structure followed by least-squares refinement against the X-ray data. In addition, the deviations in the atomic temperature factors for the protein atoms in different crystal structures of myoglobin are studied by comparing the B-factors in CO-Mb with those in oxy-Mb (Phillips, 1980) and in four different structures of met-Mb (Frauenfelder et al., 1979, Kuriyan, this thesis).

While the backbone B-factors in the 300K structure of met-Mb (Frauenfelder et al., 1979) and the 260K structure of oxy-Mb (Phillips, 1980) are very similar in magnitude, those of CO-Mb show large deviations from the met- and oxy-Mb values in some regions of the protein. The similarity of the met- and oxy-Mb backbone B-factors and their non-uniform deviations from those in the CO-Mb structure suggests that the differences could be due to changes in the internal dynamics of the protein on CO binding. However, the atomic B-factors include a component that is due to static disorder in the crystal and not to the motion of atoms in a particular molecule (Frauenfelder et al., 1979). This component depends on the quality of the crystal and is presumably affected by parameters such as the pH, the ionic strength and the degree of mechanical shock that the crystal is subject to during mounting.

[1] Abbreviations used: Mb, myoglobin; CO-Mb, carbon-monoxy (Fe II) myoglobin; oxy-Mb, O_2 (Fe II) myoglobin; met-Mb, OH_2 (Fe III) myoglobin; rms, root-mean-square.

In order to study the variation of the atomic B-factors from crystal to crystal, three new data sets to 2.0Å resolution were collected for met-Mb and models refined against them. The resulting B-factors are identical to within experimental error, but significantly different from those in the original met-Mb structure of Frauenfelder et al. (1979) and the oxy-Mb structure of Phillips (1980).

An attempt was made to separate the effects of ligand binding from those of lattice disorder by collecting data on the met and CO forms of Mb on the same crystal; it was assumed that the effects of lattice disorder would be the same in both data sets. This assumption proved to be invalid, but these experiments did demonstrate conclusively that the differences noticed earlier between CO-Mb and met- or oxy-Mb are not due to changes in the ligation state of the molecule. The backbone B-factors in the new met-Mb structures are identical, within experimental error, to the original 1.5Å CO-Mb B-factors, but deviate significantly from those in the new 2.0Å CO-Mb structure.

The results of the energy minimizations and re-refinements of the CO-Mb structure are discussed in Section II. Section II (a) describes the energy minimization method used and the refinements of the minimized structures. In Section II (b) the deviations between the coordinates and B-factors obtained by X-ray refinement alone and those obtained by combined energy-minimization and X-ray refinement are discussed. In Section III the atomic B-factors in different myoglobin structures are compared. The experimental details regarding the data collection are given in Section III(a). The refinement of models against these data sets is also described in this section. In Section III (b) the refined temperature factors obtained are compared with those of the 1.5Å CO-Mb structure (Chapter 4 of this thesis) and other myoglobin structures. Section IV summarizes the conclusions of this chapter.

II (a) Estimation of Errors in the Refined Parameters

In principle, an estimate of the errors in the refined coordinates and temperature factors can be obtained from the inverse of the least-squares normal matrix (Hendrickson and Konnert, 1980). However, in the Konnert-Hendrickson method, several approximations introduced into the normal matrix make it difficult to obtain a reliable estimate of the errors. These approximations are: (i) the neglect of most of the off-diagonal terms in the matrix, (ii) the use of restraints, which reduces the number of degrees of freedom of the system in an undetermined way and (iii) the use of an approximate weighting scheme for the structure factors (Hendrickson and Konnert, 1980).

A commonly used method to obtain an upper limit on the coordinate errors is that due to Luzzati (1953) (see, for example, Baker, 1980, Phillips, 1980, James and Sielecki, 1983, Honzatko et al., 1985). Luzzati (1953) assumed that the discrepancies between the observed and calculated structure factors are due entirely to errors in the coordinates of the atoms. Assuming that the molecule consists of a large number of identical atoms, and that the distribution of errors in the structure is Gaussian, he derived a relation between the overall error in the coordinates and the dependence of the R-factor on resolution (Luzatti, 1953). The assumptions used to derive this result are probably not valid for protein structures. The wide range of temperature factors found in protein molecules (Petsko and Ringe, 1984) causes the assumption of identical atoms to break down. The dependence of the R-factor on resolution can be ambiguous, since it depends on several adjustable parameters such as the relative weights assigned to the high and low resolution terms. Finally, the method results in only an overall estimate of the error; comparison of different structure determinations of trypsin has shown that the errors in the coordinates vary from region to region in the protein and are correlated with the atomic B-factors (Chambers and Stroud, 1979).

The approach we have taken to estimate the errors in the refined structure is to perturb it by energy minimization without reference to the X-ray data. The perturbed structure, with uniform temperature factors assigned to each atom, is then used as a starting model for least-squares refinement against the X-ray data. Refinement is continued until a structure is obtained with the same R-factor and restraint parameters as for a structure obtained using crystallographic refinement alone; comparison of the two structures yields an estimate of the reproducibility of the coordinates and temperature factors. Since the energy minimization is done without reference to the X-ray data, this is similar to introducing small random shifts in each atom and then continuing the refinement.

Experience has shown that the radius of convergence of refinement at 1.5Å resolution is between about 0.5Å and 1.0Å. If the atomic positions are shifted by more than this amount, refinement alone is unlikely to move the atoms back to their true positions and rebuilding of the structure on the basis of difference electron density maps will be necessitated. To avoid this the energy of the protein was minimized by a restrained method (Bruccoleri and Karplus, 1985). The resulting atomic shifts were small enough that examination of difference electron density maps indicated that no rebuilding of the structure was required prior to least-squares refinement.

The sidechain of Arg 45 (CD3) is apparently disordered in CO-Mb (Chapter 4 of this thesis). One conformation is more buried than the other and it might be expected that atoms around this residue would adjust their positions depending on the conformation adopted by Arg 45 (CD3). These adjustments, if they occur, are too small to be clearly discerned in the electron density maps, but they might be reflected in a larger uncertainty in the coordinates of the surrounding atoms. Two energy minimizations were done on the X-ray CO-Mb structure, one with Arg 45 (CD3) in conformation A (minimization A) and one in conformation B (minimization B). In both minimizations only water molecules within 6.5Å of the Arg 45 (CD3) sidechain were included and only one of the two conformations for each of the seven disordered residues and the CO ligand was included. The minimized structures were then separately refined against the X-ray data.

The initial structure used for the energy minimization was an intermediate one in the refinement of CO-Mb (Chapter 4 of this thesis), with rather poor stereochemistry (see Table 2, below). Hydrogen atoms on polar groups, which were not included in the X-ray model, were built by the program CHARMM (Brooks et al., 1983), using an algorithm developed by Brunger and Karplus (1985). The minimizations were done using the program CHARMM and the ABNR algorithm (Brooks et al., 1983). For the first 120 steps harmonic restraints of 50.0 Kcal/mole/Å were used for the waters and Arg 45 (CD3); restraints of 20.0 Kcal/mole/Å were placed on all the other protein atoms (Bruccoleri and Karplus, 1985). The reference structure for the restraints was initially the X-ray structure; it was reset to the current structure every 40 steps. After 120 steps the restraints were reduced to 20.0 Kcal/mole and 10.0 Kcal/mole for the two classes of atoms, respectively. Minimization A was continued for 250 steps and minimization B for 370 steps. The initial and final energies and deviations from the initial structure are given in Table 1, below.

TABLE 1 Energy minimization of CO-Mb

	Minimization A		Minimization B	
	Initial	Final	Initial	Final
Steps	-	250	-	370
Energies:				
(in Kcal/mole)				
Total	1403.7	-5535.0	275.5	-5712.0
Bonds	1899.6	31.4	1886.5	30.8
Angles	1307.6	223.4	1307.5	231.2
Dihedrals	314.3	148.2	353.3	151.1
Van der Waals	2005.5	-694.3	832.2	-673.7
Electrostatics	-4359.5	-5341.4	-4340.8	-5521.0
rms deriv.	64.1	0.21	324.5	0.11
rms deviations				
(in Å)				
backbone	0.0	0.19	0.0	0.24
sidechains	0.0	0.43	0.0	0.56

The initial energies of the two structures are different because of differences in the conformation of Arg 45 (CD3) and the number of water molecules included. The larger initial values of the van der Waals energy and the rms gradient in minimization A are mostly due to close contacts between the hydrogen atom of a water molecule and the terminal nitrogen of Arg 45 (CD3).

Difference electron density maps were calculated using phases derived from the two energy minimized structures. These maps indicated that even the charged surface sidechains, which moved the most during the minimization, did not need to be rebuilt before least-squares refinement against the X-ray data. Three parallel, but independent, least-squares refinements were started at this point.

In the first refinement the initial model was the refined structure of CO-Mb prior to energy minimization. As described in Chapter 4 of this thesis, the model includes 136 water molecules, one sulphate ion and two alternate conformations for seven residues. The initial R-factor was 16.5% with poor stereochemistry (see Table 2, below). Increasing the weights on the restraints, and continuing the refinement for 12 cycles resulted in a structure with an R-factor of 16.7% and better stereochemistry, which will be referred to as CO-Mb(X1). The refinement was then continued with higher weights on the restraints and four cycles resulted in a structure with very good stereochemistry, and an R-factor of 18.7% (see Table 2). This structure will be referred to as the CO-Mb(X2) structure.

The initial model for the second refinement was the structure obtained in minimization A. The water molecules and alternate conformations which were not included in the energy miminization were taken from the structure prior to energy minimization. The intial R-factor was 26%. 8 cycles of least-squares refinement reduced the R-factor to 16.7%, which is the same as that for CO-Mb(X1). The stereochemical parameters are also comparable to those for CO-Mb(X1). This structure will be referred to as CO-Mb (A1).

The structure obtained in minimization B was used for the third refinement. 11 cycles resulted in a structure with an R-factor of 16.7%, which will be referred to as CO-Mb(B1). The stereochemical parameters are once again comparable to those for CO-Mb(X1) and CO-Mb(A1). To examine the effect of tighter restraints on the structure the refinement was continued for four more cycles to obtain a structure, CO-Mb(B2), with stereochemical parameters similar to CO-Mb(X2). Surprisingly, the R-factor for CO-Mb(B2) (18.0%) is significantly lower than for CO-Mb(X2) (18.7%). This result is of interest, since the energy-minimization/refinement method has resulted in a "better" structure (lower R-factor with comparable stereochemistry) than straight crystallographic refinement. The final refined structure of CO-Mb used for the analysis in Chapter 4 was obtained from CO-Mb(B2) by continuing the refinement with looser weights on the stereochemistry[1] it will be referred to as CO-Mb(B3).

The final R-factors and the rms deviations of bond, angle and planar 1-4 distances and planes from ideality are given in Table 2, below.

[1] I forgot to mention this in the Chapter 4 on CO-Mb refinement; this will be corrected before the final draft.

Table 2 Overall Statistics for the Refinements

Structure	R-factor	r.m.s. delta of bonds	 angles	 1-4 dist	 planes
		Å	Å	Å	Å
Initial X-ray structure.	0.165	0.051	0.070	0.079	0.027
CO-Mb(X1)	0.167	0.044	0.063	0.072	0.023
CO-Mb(A1)	0.167	0.047	0.065	0.073	0.023
CO-Mb(B1)	0.167	0.044	0.064	0.070	0.023
CO-Mb(X2)	0.187	0.025	0.043	0.050	0.013
CO-Mb(B2)	0.180	0.025	0.043	0.051	0.013
CO-Mb(B3)	0.171	0.030	0.047	0.057	0.016
TARGET stereochemical standard deviations:		0.030	0.040	0.052	0.025

II (b) Comparison of the various refined structures of CO-Mb.

To summarize the results of the last section, three refinements were done, starting from the same initial coordinates. In one, the least-squares refinement against the X-ray data was continued in the normal way, resulting in two structures, CO-Mb(X1) (loose restraints) and CO-Mb(X2) (tight restraints). In the other two refinements, two energy minimized structures (with Arg 45 in conformations A and B, respectively), were used as the initial models, to obtain CO-Mb (A1) and CO-Mb (B1) (loose restraints), and CO-Mb (B2) (tight restraints).

In this section the overall deviations between the structures are calculated, and it is shown that the deviations increase with B-factor (Section IIb (i)). The deviations of atoms around Arg 45 (CD3) are also examined (Section IIb (ii)). The energies, as evaluated using the CHARMM program (Brooks et al., 1983), of two X-ray structures with loose restraints are compared in order to determine whether the energy-minimization/refinement procedure results in a more stable structure (Section IIb (iii)). Finally, the B-factors of the five structures are compared and their errors estimated (Section IIb (iv)).

IIb (i) Positional deviations

Table 3 compares the refined positions of atoms in the following six structures: the initial X-ray structure (which was the starting point for the energy-minimizations and refinements) and the five refined structures (CO-Mb X1, A1 and B1, with loose restraints, and CO-Mb X2 and B2, with tight restraints). The structures were oriented by least-square superpositioning of the backbone atoms (C,N,C_α) before the deviations were calculated. The first entry in each box is the rms deviation of backbone atoms, in Å, and the second entry is the rms deviation of the sidechain atoms, in Å.

Table 3: Positional Deviations Between Refined Structures

	X1	X2	A1	B1	B2
Initial	0.037 0.056	0.068 0.089	0.123 0.240	0.137 0.311	0.133 0.309
X1		0.069 0.108	0.122 0.240	0.135 0.309	0.130 0.307
X2			0.118 0.240	0.132 0.309	0.122 0.304
A1				0.078 0.175	0.076 0.177
B1					0.036 0.046

The structures obtained from the A and B minimizations are closer to each other than the structures obtained from purely crystallographic refinement. Increasing the restraints brings the X and B structures closer together only slightly; i.e., the shifts required to improve the stereochemistry are smaller than the differences between the X and B structures. In Fig. 1 the maximum backbone and sidechain deviations between CO-Mb(X1) and CO-Mb(B1) for each residue are plotted. The backbone deviations show only a small variation over the structure, the maximum deviations being less than 0.5Å. The sidechain deviations are larger and show more variation. In Fig. 1 the sidechains are separated into three classes: buried (with average sidechain accessible area (Lee and Richards, 1971) less than 3.0Å^2 for a water sized probe), partially buried (with average accessible surface areas between 3.0Å^2 and 15.0Å^2) and exposed (with average accessible surface areas greater than 15.0Å^2). Among the residues with deviations greater than 0.5Å, there are many that are buried or partially buried (Fig. 1), showing that the large deviations between the structures are not confined to the completely exposed surface sidechains.

Fig. 2 shows a histogram of the deviations in position between CO-Mb X1 and B1. The distribution is peaked at around 0.15Å and there are few atoms with deviations greater than 0.4Å. The deviations between the two CO-Mb structures are approximately three times smaller those observed between CO-Mb and deoxy-Mb (Chapter 4 of this thesis).

The deviations between the structures can be used to estimate the overall standard deviations, σ_r, in the coordinates. If all the atoms in a particular class, such as the backbone atoms, are assumed to obey the same Gaussian error distribution, then $\sigma_r^2 = \frac{1}{2} \Delta^2$ where Δ^2 is the mean square deviation between the atoms (Appendix to Chapter 3 of this thesis). Based on the deviations between the X1 and B1 structures, for example, σ_r is 0.1Å for the backbone and 0.2Å for the sidechains. It would be interesting to see whether unrestrained minimization of the structure (followed by rebuilding with reference to difference electron density maps, if necessary) and subsequent least-squares X-ray refinement would lead to a significantly larger estimate of the errors.

The deviations between the various structures are largest for atoms with large B-factors. Table 4 gives the average positional deviation and the standard deviation in the positional deviation for atoms with various magnitudes of B-factors.

Table 4: Positional deviation between CO-Mb X1 and B1 vs. B-Factor

The average deviation in a particular range of B-factor for the two loosely restrained structures CO-Mb (X1) and CO-Mb (X2) are given below. The errors appear to be independent of B-factor for small B-factors (less than 8Å^2).

B-factor range	Average Deviation in Å	Standard Deviation in Å	Number of atoms
0 - 2.0	0.14	0.11	44
2.0 - 4.0	0.12	0.07	125
4.0 - 6.0	0.13	0.13	230
6.0 - 8.0	0.14	0.08	201
8.0 - 10.0	0.18	0.11	167
10.0 - 12.0	0.19	0.12	134
12.0 - 14.0	0.20	0.17	102
14.0 - 16.0	0.20	0.12	90
16.0 - 18.0	0.29	0.21	55
> 20.0	0.55	0.66	112

IIb (ii) Deviations of atoms around Arg 45 (CD3)

In the X1 refinement the arginine sidechain was modelled with both conformations (A and B) while the A1 and B1 refinements were done with the sidechain in conformations A or B alone, respectively. This distinction is important because the Konnert-Hendrickson refinement program imposes non-bonded contact restraints which prevent atoms from approaching each other too closely. In the A1 refinement the sidechain was not in conformation B, thus allowing surrounding atoms to move towards the vacant space. The A1 and B1 structures are compared in Table 5 below. Both the X1 and the B1 refinements had the sidechain in conformation B; the non-bonded contact restraints imposed by the refinement are similar for the surrounding atoms in both refinements, and differences in the structures are due to the effects of the energy minimization. The X1 and B1 structures are also compared in Table 5, below.

Table 5: Deviations of atoms within 6.0 Å of Arg 45 CD3

The structures being compared were superimposed on the backbone atoms before the coordinate deviations were calculated. The table includes deviations only for those atoms that are within 6.0Å of the centroid of the guanidinium ring of Arg 45 (CD3) in the B1 structure and which deviate by more than 0.2Å in either of the two comparisons. The B-factors of the atoms in the X1 structure are also given.

Table 5

		Distance	Deviation	Deviation	
		from Arg 45	X1/B1	A1/B1	
Residue	Atom	guanidium			B-Factor
		Å	Å	Å	$Å^2$
Phe 43 (CD1)	C_β	4.35	0.28	0.05	7.2
Arg 45 (CD3)	C_β	2.57	0.15	0.26	9.4
	C_γ	2.62	1.16	0.26	9.8
	C_δ	1.68	1.56	0.52	10.4
	N_ε	0.69	0.77	1.78	10.0
	C_ζ	1.07	0.91	3.82	10.0
	NH1	1.78	0.73	4.94	9.6
	NH2	2.24	1.05	4.65	8.4
Phe 46 (CD4)	$C_{\delta 2}$	3.44	0.20	0.24	5.3
	$C_{\varepsilon 1}$	3.87	0.20	0.11	5.5
	$C_{\varepsilon 2}$	2.97	0.32	0.37	5.0
	C_ζ	3.33	0.21	0.27	6.0
Asp 60 (E3)	$O_{\delta 1}$	4.98	0.32	0.06	8.2
His 64 (E7)	$C_{\delta 2}$	4.62	0.27	0.07	10.7
Heme	CAD	5.26	0.27	0.06	5.9
	CBD	4.45	0.51	0.91	9.4
	CGD	3.91	0.28	0.30	10.8
	O1D	3.73	0.17	0.38	10.6
	O2D	4.31	0.35	0.20	12.2

The average sidechain B-factor in the X1 structure is 11.5Å^2. All the atoms in Table 5, except the O2D atom of the heme, have B-factors lower than this. Nevertheless, the deviations in Table 5 are among the largest seen in the structure (Fig. 2). The large change in the structure of the B conformation of Arg 45 (in the X1/B1 comparison above) shows that there are errors in the modelling of this residue which are probably associated with the lack of disorder in the model for the backbone and C_β atoms. Significant shifts in position are also seen in the two phenyl residues that interact with conformation B of Arg 45 (CD3), in Asp 60 (E3), which is H-bonded to conformation A of the arginine, in the distal histidine, and in the heme propionic acid sidechain. This group is H-bonded to both conformations of Arg 45 (CD3) and apparently changes its conformation to follow that of the arginine.

IIb (iii) Energies of the refined structures

Protein structures obtained by X-ray diffraction are used as the starting point in a number of different simulation techniques such as energy minimization, molecular dynamics, normal mode calculations and Monte Carlo calculations. In most cases the first step of the simulation is an energy minimization of the X-ray structure to reduce the large initial forces due to errors in the structure as well as in the empirical potential energy function; this results in shifts of about 0.25Å to 0.5Å in the backbone atom positions. In this section the energies of two equivalent X-ray structures (X1 and B1) of CO-Mb are calculated using the program CHARMM (Brooks et al., 1983) and it is shown that the structure obtained by energy minimization followed by X-ray refinement is more stable than the original X-ray structure. This is perhaps to be expected, since the energy minimization was done with reference to the CHARMM potential, but it is of interest to quantify the extent to which the energies differ.

Table 6, below lists the energies and rms gradients of the energy (the average force on each atom) for the X1 and B1 structures. The electrostatics and van der Waals terms in the energy function require that the polar hydrogen atoms be present explicitly in the structure. The positions of these hydrogens were calculated from the X-ray structures by the program CHARMM (Brooks et al., 1983) and the energies given for these two terms include the hydrogen atom contributions. For the other energy terms in Table 6, the hydrogens are excluded since their positions were not refined.

Table 6: Energies of the Refined Structures

Energy Term	X1	B1
	Kcal/Mole	Kcal/mole
Total	-674.9	-1507.1
Hydrogens Excluded:		
Bonds	1463.2	1509.0
Angles	1120.2	1013.2
Dihedral	273.6	250.1
Impropers	194.2	188.9
Hydrogens Included:		
van der Waals	521.5	-125.7
Electrostatics	-4247.6	-4342.6
rms derivatives: (Kcal/mole/Å)		
Non-bonded	44.0	22.6
Others	47.0	47.0

The B1 structure (energy-minimization/X-ray refinement) is more stable than the X1 structure (X-ray refinement alone) by 830 Kcal/mole (approximately 0.7 Kcal/mole per atom). Most of the stabilization comes from lower van der Waals and electrostatic energies (including H-bonding contributions). The bonds are the most tightly restrained parameters in the X-ray refinement and the restraint values are slightly different from .the CHARMM equilibrium values (Kuriyan et al., 1985); the bond energies are actually higher in the B1 structure than in the X1 structure. All the other energy terms are lower in the B1 structure.

The rms gradient of the energy for the non-bonded terms is lower by 50% in the B1 structure, while that for the other energy terms is approximately the same in both structures. The two X-ray structures have not been minimized to see whether the shifts in atomic positions required to get a small rms energy gradient are significantly different in the two structures, but this would be of interest.

IIb (iv) Errors in the Temperature Factors

The B-factors in the various refined structures are compared in Table 7. All the atoms were used in the comparisons and in each case four statistical parameters were calculated. The first entry in each box is the correlation coefficient (from a linear regression) between the B-factors of the two structures. The second entry is the fractional error, defined as:

$$\text{Fractional Error} = \frac{\Sigma\,|B_i-B_j|}{\Sigma\,B_i} \text{ X } 100.0$$

where B_i refers to the i^{th} row and B_j to the j^{th} column in Table 7. The third entry is the average absolute error $|B_i-B_j|$ and the fourth is the average error B_i-B_j. The last term indicates whether there is a constant offset in the B-factors being compared.

Table 7: B-factor Comparisons

	X1	X2	B1	B2	A1
Initial	0.998 3.3 0.290 -0.015	0.976 10.6 0.925 0.097	0.949 15.5 1.360 -0.715	0.947 14.8 1.303 -0.574	0.963 14.0 1.223 -0.695
X1		0.979 9.3 0.816 0.113	0.950 15.6 1.363 -0.699	0.952 13.7 1.207 -0.559	0.962 14.1 1.247 -0.680
X2			0.926 19.5 1.689 -0.812	0.954 13.7 1.191 -0.678	0.936 18.5 1.601 -0.793
B1				0.984 9.5 0.907 0.147	0.986 7.1 0.671 0.020
B2					0.968 12.4 1.158

Increasing the restraints is seen to have a large effect on the B-factors (see comparison of X1 and X2 or B1 and B2): the fractional error in the B-factors is 9%-10% between loosely restrained and tightly restrained structures. The backbone and sidechain B-factors have been compared and it can be shown that the backbone B-factors increase and the sidechain temperature factors decrease with tighter restraints. The program restrains the difference in temperature factors between atoms that are bonded to each other or to the same third atom (Konnert and Hendrickson, 1980). Yu et al. (1985) have shown that these restraints are in qualitative agreement with the results of molecular dynamics simulations, but are too restrictive. This is consistent with the increase in sidechain temperature factors observed on relaxing the restraints.

The correlation coefficients are high in all the comparisons, showing that there are no large, systematic, deviations in the B-factors between any of the structures. There are constant offsets in the B-factors between different structures, but these are small. The fractional error in the B-factors between two structures is largest for the X1 and B1 structures (16%) and decreases slightly with tighter restraints (14% between X2 and B2).

III Variation in the B-factors in Different Crystal Structures of Mb

In the first part of this section the experimental details regarding data collection and refinement are presented. In the second part the B-factors of the different Mb structures obtained in this work are compared with those of met-Mb (Frauenfelder et al.,1979) and oxy-Mb (Phillips, 1980). The errors in the coordinates of the new met-Mb structures are discussed in Chapter 3 of this thesis. They are similar in magnitude to the errors estimated for CO-Mb in the first part of this chapter.

IIIa Data Collection and Refinement

IIIa (i) Data Collection on Three met-Mb Crystals.

The crystallization procedure differed from the method of Kendrew and Parrish (1956) which was used to obtain crystals for the 1.5Å CO-Mb data set described in the last Chapter. Instead of using pH 6.0 phosphate buffer (Chapter 4), the crystals were obtained by adding 250 ml of a 48 mg/ml solution of Mb (Sigma Chemical Co., filtered through glass wool) to 750 ml of 75% satd. ammonium sulphate, without using buffer (Takano 1977a). Three data sets to 2.0Å resolution were collected on three crystals, two at 290K (room temperature) and one at 255K, on a Nicolet P3 diffractometer. The data collection and processing method is essentially the same as that described in Chapter 4 of this thesis, except that background-peak-background scans were used instead of Wyckoff scans. The empirical absorbtion curve (North et al., 1968) in each case was collected before the data set. The total data collection time, the extent of radiation damage (averaged over five check reflections) and the number of reflections above 2 standard deviations are given in

Table 8, below. All the data sets are to 2.0Å resolution and there are approximately 9000 unique reflections between 10.0Å and 2.0Å.

Table 8: Statistics on Data Collection

	Total Exposure (hours)	Radiation Damage (over total time)	Number of reflection (> 2σ)
met-Mb (1) 290K	39	13%	8613
met-Mb (2) 290K	111	27%	8700
met-Mb (3) 255K	37	10%	8304
CO-Mb (4) 255K	51	28%	8144

The counting times per reflection were different for each data set. Faster counting times result in less radiation damage but at the cost of decreased precision in the data.

IIIa (ii) Data collection on CO-Mb

The method used to convert met-Mb crystals to the CO form is the same as that described in Chapter 4 of this thesis. One of the crystals used for the 290K met data sets was converted to the CO form and used to collect a 2.0Å data set at room temperature. After data collection the crystal was dissolved in deoxygenated phosphate buffer (pH 6.0) and UV and visible spectra were run. The spectra indicated that a significant amount (~30%) of the protein was in the met form (Hanania et al., 1966). Refinement of the data led to the iron atom moving out of the plane of the heme, intermediate between the in-plane iron of CO-Mb and the iron in met-Mb, which is about 0.2Å out of the mean heme plane (Chapter 4 of this thesis). This confirmed that the contamination due to the met form was present during data collection and the data set was rejected.

The crystal used for the low temperature (255K) met data collection was then treated with dithionite and CO in the same way, except that equilibration with a CO atmosphere was allowed to proceed for two days rather than one. A test crystal of size comparable to the one required for data collection was removed from the reaction vessel and dissolved in buffer solution. UV and visible spectra showed that conversion to the CO form was not complete (Hanania et al., 1966). The entire procedure, including the addition of dithionite, was then repeated. Spectra run on a test crystal removed at this stage showed no trace of met contamination and the crystal required for data collection was removed under a glove bag and mounted as described in Chapter 4. After completion of data collection at 255K the crystal was dissolved in buffer. The UV and visible spectra showed no peaks due to met-Mb.

As shown below, the refined B-factors in this CO-Mb structure are significantly higher than in the met-Mb structure obtained from the same crystal. In the data reduction step for the CO-Mb data set no correction was made for the radiation damage already sustained by the crystal during the collection of the met data set. Burley et al. (1986) have shown that in ribonuclease radiation damage as high as 50% of the initial intensities results in no significant increase in the B-factors and so the neglect of the intial 10% damage due to the met data collection is not likely to be the main cause for this increase. A more likely explanation is that the mechanical shocks suffered by the crystal when breaking the capillary tube, transferring the crystal to the reaction vessel and finally remounting it in a new capillary tube were sufficient to increase the disorder in the crystal.

IIIa (iii) Refinement of the 2.0 Å met- and CO-Mb structures

The initial model used for the met-Mb refinements was the 300K structure of met-Mb (Frauenfelder et al., 1979). The refinement of the two 290K structures proceeded without much rebuilding and the final models in both cases include 75 water molecules and 2 sulphate ions. A larger number of solvent peaks were seen in difference electron density maps using the 255K data; the final model includes 117 water molecules and three sulphate ions. No attempt was made to model disordered residues with more than one conformation in any of the met-Mb refinements.

The initial model used for the CO-Mb refinement was the final 1.5Å refined structure (Chapter 4 of this thesis). No rebuilding of the sidechains was necessary and the same number of solvent molecules and sidechains with alternate conformations was included in the final model as in the 1.5Å structure. The final R-factors and stereochemical parameters for the four structures are given in Table 9 below:

Table 9: Overall Statistics for the Refinements

Structure	R-factor	rms deviation from ideality of bonds	 angles	 1-4 dist Å	 planes Å
met-Mb (1) 290K	0.185	0.024	0.045	0.047	0.011
met-Mb (2) 290K	0.182	0.026	0.048	0.052	0.012
met-Mb (3) 255K	0.182	0.026	0.040	0.052	0.012
CO-Mb 255K	0.189	0.030	0.049	0.058	0.013
TARGET stereochemical standard deviations:		0.030	0.040	0.052	0.025

IIIb Comparisons of B-factors

IIIb (i) Comparison of the B-factors from the 1.5 Å Structure of CO-Mb with those from oxy-Mb and met-Mb

The backbone B-factors for met-Mb at 300K (Frauenfelder et al., 1979) and oxy-Mb (Phillips, 1980) are similar, both in magnitude and in the variation from residue to residue. Fig. 3a compares the residue averaged backbone B-factors for the two structures. The average backbone B-factors are 11.5Å^2 for the 300K met-Mb structure and 11.9Å^2 for the 260K oxy-Mb structure. The sidechain B-factors are, however, significantly larger in oxy-Mb than in met-Mb (300K); the sidechain average B-factor is 17.1Å^2 in the former and 13.1Å^2 in the latter. The average backbone and sidechain B-factors are given in Table 10 below.

Table 10: Average B-factors for various Mb structures

	Average B-factor (in Å^2)		
Structure	Backbone	Sidechain	All Atoms
CO-Mb (260K) 1.5Å	8.7	10.0	9.5
CO-Mb (255K) 2.0Å	13.8	14.8	14.4
met-Mb (300K) 1.5Å (Frauenfelder, et al., 1979)	11.9	13.1	12.6
met-Mb (290K) 2.0Å	9.9	11.3	10.8
met-Mb (290K) 2.0Å	9.6	10.8	10.3
met-Mb (255K) 2.0Å	8.8	10.1	9.6

The met-Mb structure was refined by the Konnert-Hendrickson method with restraints on the B-factor variation (Frauenfelder et al., 1979), while the oxy-Mb structure was refined by the Jack-Levitt method, without such restraints (Phillips, 1980). This is reflected in the sharper variation of the B-factors from residue to residue in oxy-Mb (Fig. 3a). The correlation coefficient between the two sets of backbone B-factors is 0.76 and the fractional error (defined in Section IIb (iv)) is 21%. The agreement between the backbone B-factors for the two structures would be reduced if the oxy-Mb structure were refined with B-factor restraints, since this tends to increase the backbone B-factors, the increase being greatest for residues with very mobile sidechains. Thus, the agreement between the met-Mb (300K) and oxy-Mb B-factors is to some extent fortuitous.

On comparing the backbone B-factors of the 1.5Å structure of CO-Mb (260K) (Chapter 4) with met-Mb (300K) and oxy-Mb (260K) the fractional errors are 35% and 40%, respectively. In Fig 3b the backbone B-factors for the CO-Mb structure and the met-Mb structures are plotted as a function of residue number. The magnitude of the B-factors in the loops are similar in both structures, but the B-factors in helix regions are reduced in CO-Mb. Fig. 3c plots the difference in backbone B-factor (averaged over five-residue segments) between met-Mb and CO-Mb. The differences are not uniform over the structure (i.e., they cannot be removed by adding a constant offset to the B-factors in one structure) and they vary in magnitude, in the helix regions, from 2Å^2 to 5Å^2. In Section II of this chapter the overall error in the B-factors for CO-Mb was estimated to be about 15%, which corresponds to about 2Å^2 for the backbone atoms. The differences in Fig. 3c are larger than this.

IIIb (ii) B-factors from Refinements of the New Data.

It was originally thought that the differences in the B-factors described above could be due to changes in the internal dynamics of the protein on CO binding. However, comparison of the CO-Mb B-factors with those in the new met-Mb structures shows that this is not the case; the new met-Mb B-factors for the backbone are identical, within experimental error, to the CO-Mb backbone B-factors. Fig. 3d compares the backbone B-factors in the new 255K structure of met-Mb with those in the 1.5Å structure of CO-Mb (260K). The also agreement, with a correlation coefficient of 0.93 and a fractional error of 18%, is better than that between met-Mb (300K) (Frauenfelder et al., 1979) and oxy-Mb (260K) (Phillips, 1980). The B-factors of the three new met-Mb structures agree very well among themselves, with fractional errors for the backbone atoms ranging from 12% to 15% (comparable to the errors estimated in the CO-Mb B-factors).

While the B-factors in the 1.5Å structure of CO-Mb and the three new met-Mb structures agree well, the new 2.0Å CO-Mb structure has B-factors that are significantly higher. Fig. 3e compares the backbone B-factors for the 2.0Å 255K met-Mb structure and the 2.0Å CO-Mb structure. Both structures were refined against data collected on the same crystal, yet the average backbone B-factor for the met-Mb structure is 8.8Å^2 while that for the CO-Mb structure is 13.8Å^2.

To summarize these results, the differences between the 1.5Å 260K CO-Mb backbone B-factors and those in the 300K met-Mb structure arise not from the change in ligation state but from differences in the crystals. For example, Fig. 3f plots the backbone B-factors for the 300K met-Mb structure (Frauenfelder et al., 1979) and the 290K structure of met-Mb (this work). The difference between the two is similar to that shown in Fig. 3b, between CO-Mb and met-Mb.

IIIb (iii) Linear Regression Analyses between sets of B-factors.

Fig. 4a shows a scatter plot between the two sets of CO-Mb B-factors (260K and 1.5Å, 255K and 2.0Å). Though there is a great deal of scatter, a large source of the deviations between the two sets of B-factors is a constant offset of the temperature factors in one with respect to the other. Fig. 4b shows a scatter plot of the B-factors in the 1.5Å CO-Mb structure and the 300K met-Mb structure (Frauenfelder et al., 1979). In this case there is a constant offset as well as a non-unity slope in the best fit line through the points, i.e., the differences between the B-factors changes with the magnitude of the B-factors.

Examination of such scatter plots shows that the deviations between the B-factors between different crystal structures have two components. One is a linear deviation between the two sets of B-factors, which can be corrected for by the intercept and slope of the least-squares line through the points in the scatter plot. The other is the essentially random scatter of points about the least-squares line. The magnitude of this scatter does not seem to be significantly larger than the estimated error in each set of B-factors (Section II). Hence, the slope and intercept of the least-squares line can be thought of as the systematic deviations between the two sets of B-factors whereas the scatter about the least-squares line reflects the intrinsic error in the B-factors.

Linear regression analyses on the B-factors have been performed on various pairs of structures. For each pair being compared the following statistical parameters were calculated: r, the correlation coefficient; the fractional error $\frac{\langle|\varepsilon|\rangle}{B}$ (where ε is the deviation between the B-factors without any corrections); the rms error, ε_{rms}; the average error, $\langle\varepsilon\rangle$; the intercept of the least-squares line, α; the slope of the least-squares line, β; and σ_y, the scatter of points around the least-squares line. Table 11 gives these parameters for comparisons of the 1.5Å 260K CO-Mb structure with various other structures:

Table 11: Comparison of 1.5Å CO-Mb (260K) B-factors with other structures.

	r	$\langle \frac{\lvert\varepsilon\rvert}{B} \rangle$	ε_{rms}	$\langle\varepsilon\rangle$	α	β	σ_y
			(Å^2)	(Å^2)	(Å^2)		(Å^2)
CO-Mb(2) (255K)							
All atoms	0.89	52%	5.4	-4.9	6.4	0.9	2.2
backbone	0.90	59%	5.5	-5.0	6.3	0.9	2.1
met-Mb(1) (300K)							
All atoms	0.92	35%	3.8	-3.1	6.1	0.7	1.5
backbone	0.94	38%	3.7	-3.1	5.8	0.7	1.2
met-Mb(2) (290K)							
all atoms	0.93	20%	2.4	-1.2	1.3	1.0	2.1
backbone	0.96	17%	2.0	-1.2	1.3	1.0	1.5
met-Mb(3) (290K)							
all atoms	0.92	17%	2.2	-0.8	1.3	1.0	2.0
backbone	0.95	15%	1.7	-0.8	1.4	1.0	1.5
met-Mb(4) (255K)							
all atoms	0.89	20%	2.6	-0.1	0.25	1.0	2.6
backbone	0.93	18%	1.9	0.0	0.43	1.0	1.9
oxy-Mb (260K)							
all atoms	0.69	64%	8.9	-5.8	3.4	1.2	6.7
backbone	0.76	40%	4.7	-2.8	3.8	0.9	3.7

The slope of the least-squares line is unity for the three new met structures, as expected from the comparisons described earlier. The slope is 0.7 for the comparison with the 300K met-Mb data, but the scatter of points around the least-squares line is small. The scatter of points around the least-squares line for the comparison with oxy-Mb is large, because the B-factors were unrestrained in the refinement of the latter structure, resulting in a larger variation from atom to atom than in the other structures.

These results show that the scatter of points about the least-squares line is approximately the same for all structures refined by the Konnert-Hendrickson method; the magnitude of the scatter is similar to the error in B-factors deduced for CO-Mb by the energy-minimization/ refinement method in the last section. This also indicates that though a simple offset in the B-factors is not always enough to account for the variation of B-factors from crystal to crystal, a two parameter linear model seems adequate. A physical basis for this model might be obtained by decomposing the static disorder into two components, a translational one and a rotational one. The translational component affects all atoms equally, and could be the source of the constant offset in the B-factors. The effect of the rotational component increases with distance from the centroid of the molecule; so do the B-factors, and hence this could be the source of the deviation of the slope of the regression line from unity.

Hartmann et al. (1983) have estimated the magnitude of the disorder contribution to the B-factors by comparing the mean-square displacements of the iron in Mb as calculated from the X-ray B-factors and from Mossbauer measurements. They assume that the lattice disorder is purely translational and estimate that a correction of 3.6Å^2 should be subtracted from the B-factors. The results discussed above indicate that a rotational component to the disorder is required as well.

The results presented in this section clearly show that a change in the ligation state of the iron in Mb does not result in a detectable large scale change in the atomic temperature factors. There are, however, some specific changes that do take place. Though the lattice disorder makes comparisons of B-factors between two different structures difficult, it is possible to detect changes in the mobility of atoms by comparing their B-factors relative to other parts of the molecule. Such a comparison has been done for the distal histidine sidechain relative to the backbone, and it has been shown that the sidechain is more mobile (or disordered) relative to the backbone in CO-Mb than in met-Mb (Chapter 4 of this thesis).

SECTION IV Conclusions

(i) Refinement of perturbed structures

The use of restrained energy minimizations in crystallographic refinement has been shown to have two advantages. One is that it provides a means of perturbing the X-ray structure slightly; re-refinement of the perturbed structure results in a different X-ray structure which can be compared to the original one in order to estimate the errors in the positions and B-factors. The advantage of this method over merely introducing random shifts in the structure prior to re-refinement is that the resulting X-ray structure is energetically better than the original structure. This may be of some advantage if the X-ray structure is to be used as a starting point in simulations of protein dynamics. This method has been used to show that structures refined against the same 1.5Å data set for CO-Mb differ by 0.14Å rms in the backbone coordinates, 0.31Å in the sidechain coordinates and 15% in B-factors.

The other advantage of the method is that it provides a way of investigating the structural consequences of disorder in proteins. The results presented here are only preliminary, but they indicate that the disorder in the sidechain of Arg 45 (CD3) results in increased uncertainty in the positions of the neighboring atoms, especially those that interact with it through H-bonds which differ in the two conformations. Energy minimization provides a way of generating energetically feasible conformations of the surrounding residues which can then be used as starting points for X-ray refinement.

There are, however, some problems with the approach as it has been used in this chapter. Energy minimization results in small shifts in regions of the protein where the stereochemistry is good to begin with and the packing of atoms involves no bad contacts. These shifts might be smaller than the actual error in the coordinates, leading to an underestimation of the errors. Another problem is that only a small number (two or three) of independent structures can be generated by energy minimization of a single starting structures (these different structures can be obtained by using different minimization alogrithms). Therefore, only a small number of re-refined coordinates can be obtained and the statistics for individual atoms are poor.

Both these problems can be overcome by randomizing the X-ray structure in some way so as to obtain a large number of perturbed coordinates. These perturbed structures can then be refined against the X-ray data to obtain good estimates of the errors in the structure. It is important that the random shifts introduced into the original coordinates be large enough that realistic estimates of the errors are obtained, but small enough that the atoms are within the radius of convergence of the least-squares refinement. Also, the perturbed structures should not have very bad stereochemistry; if the bonds, angles and other restrained parameters are too far from their ideal values the refinement may not be able to improve the structure.

These perturbations can probably best be introduced by molecular dynamics simulations. To prevent the atoms from moving too far away from the original structure, but at the same time preventing too much bias towards it, square-well potentials can be introduced around the initial positions of the atoms. The width of the square well could be set to a value approximately that of the estimated radius of convergence of the refinement. Rapid refinement of 10-15 perturbed structures should be possible with the use of supercomputers, such as the CRAY, making this a feasible project for a protein the size of myoglobin. It is estimated that one least-squares cycle for myoglobin at 1.5Å will take one minute of central processor time on a CRAY-1S.

(ii) Variation in the B-factors

Comparison of B-factors from different crystal structures of myoglobin has demonstrated that lattice disorder can lead to deviations of 35% to 40% in the B-factors. Regression analysis of the various data sets has shown that the effect of disorder can be corrected for by a model with two parameters: the intercept and slope of the least-squares line. That the slope of the line is non-unity in many cases indicates that the contribution due to rotational disorder might be larger than assumed earlier (Frauenfelder et al., 1979).

Acknowledgements:

I would like to thank R.F.Tilton, jr. for providing some of the crystals used in this work and for helpful suggestions. I am grateful to D. Neidhart for running some of the CO-Mb spectra.

REFERENCES

Baker, (1980) J. Mol. Biol. 141 441-484.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M. (1983) J. Comp. Chem., 4, 187-217.

Bruccoleri, R.E. and Karplus, M. (1985) to be published.

Brunger, A.T. and Karplus, M. (1985) to be published.

Burley, S.K., Ringe. D. and Petsko, G.A. (1986) to be published.

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

Frauenfelder, H., Petsko, G.A., and Tsernoglou, D. (1979) Nature (London), 280, 558-563.

Hanania, G.I.H, Yeghiayan, A. and Cameron, B.F. (1966) Biochem. J. 98, 189-192.

Hendrickson, W.A. and Konnert, J. (1980) in "Computing in Crystallography", (ed. Diamond, R., Ramasheshan, S. and Venkatesan, K.) Indian Institute of Science, Bangalore, pages 13.01 - 13.23.

Hartmann, H., Parak, F., Steigemann, W., Petsko, G.A., Ponzi, D.R., and Frauenfelder, H. (1982) Proc. Natl. Acad. Sci. (USA) 79, 4967-4971.

Honzatko, R.B., Hendrickson, W.A. and Love, W.A. (1985) J. Mol. Biol. 184, 147-164.

James, M.N.G. and Sielecki, A. (1983) J. Mol. Biol. 163, 299-361.

Kendrew, J.C. and Parrish, R.G. (1956) Proc. Roy. Soc. ser. A, 238,305-324.

Konnert, J.H., Hendrickson, W.A., (1980) Acta Cryst. A36, 344-49.

Kuriyan, J., Petsko, G.A., Levy, R.M. and Karplus, M. (1985) J. Mol. Biol., in press.

Luzzati, P.V. (1953) Acta Cryst., 6, 142

North, A.C.T., Phillips, D.C. and Mathews, F.S. (1968) Acta Cryst. A24, 351-359.

Petsko,G.A. and Ringe, D. (1984) Ann. Rev. Biophys. Bioeng. 13, 331-71.

Phillips, S.E.V. (1980) J. Mol. Biol. 142 531-54.

Takano, T. (1977) J. Mol. Biol. 110, 537-68

Yu, H., Karplus, M. and Hendrickson, W. (1985) Acta Cryst., B41, 191-201.

FIGURE LEGENDS

Figure 1

Residue averaged deviations between the X1 and B1 structures after superimposing the backbone atoms. (see text). Solid line: backbone deviations (N,C,C_α). Dashed line: sidechain deviations. For the sidechains, the following symbols are used to indicate solvent accessibility: circles: buried residues (average accessible area less than 3.0Å^2); crosses: partially buried (average accessible area between 3.0Å^2 and 15.0Å^2); triangles: exposed (average accessible area more than 15.0 Å^2). The vertical dotted lines demarcate the helix and loop regions.

Figure 2

Histogram of positional deviations in Å between the X1 and B1 structures after superimposing the backbone atoms.

Figure 3

Backbone (N,C,C_α) atom B-factor comparisons. Fig. 3a: B-factors for oxy-Mb (Phillips, 1980) and met-Mb (Frauenfelder, 1979). Fig. 3b: B-factors for met-Mb (Frauenfelder et al., 1979) and CO-Mb (260K, 1.5 Å). Fig. 3c: difference in backbone B-factors between met-Mb (Frauenfelder et al., 1979) and CO-Mb (260K, 1.5 Å) averaged over 5 residue segments. Fig. 3d: B-factors for 255K met-Mb (2.0 Å) and 260K CO-Mb (1.5 Å). Fig. 3e: B-factors for 255K met-Mb and 255K CO-Mb, both refined against data on the same crystal. Fig. 3e: Fig. 3f: B-factors for 300K met-Mb (Frauenfelder et al., 1979) and met-Mb (290K).

Figure 4

Scatter plots for the B-factors. All the atoms are included. Fig. 4a: CO-Mb (255K, 2.0Å, "B-factor(1)") and CO-Mb (260K, 1.5Å, "B-factor (2)"). Fig. 4b: met-Mb (300K, Frauenfelder et al., 1979, "B-factor(1)") and CO-Mb (260K, 1.5Å, "B-factor (2)").

FIGURE 1

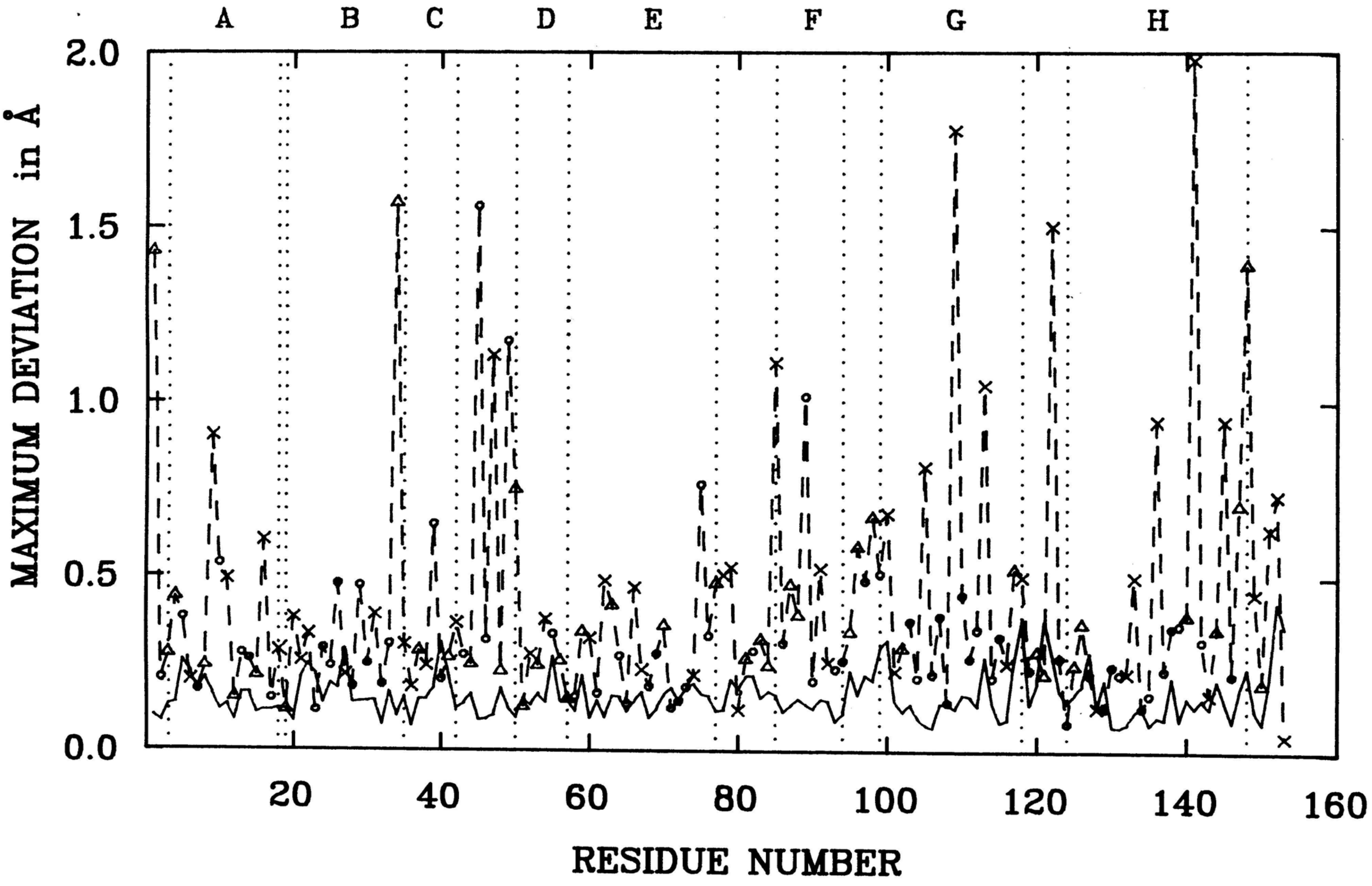

A
B
C
D
E
F
G
H
2.0
1.5
1.0
0.5
0.0
MAXIMUM DEVIATION in Å
20
40
60
80
100
120
140
160
RESIDUE NUMBER

FIGURE 2

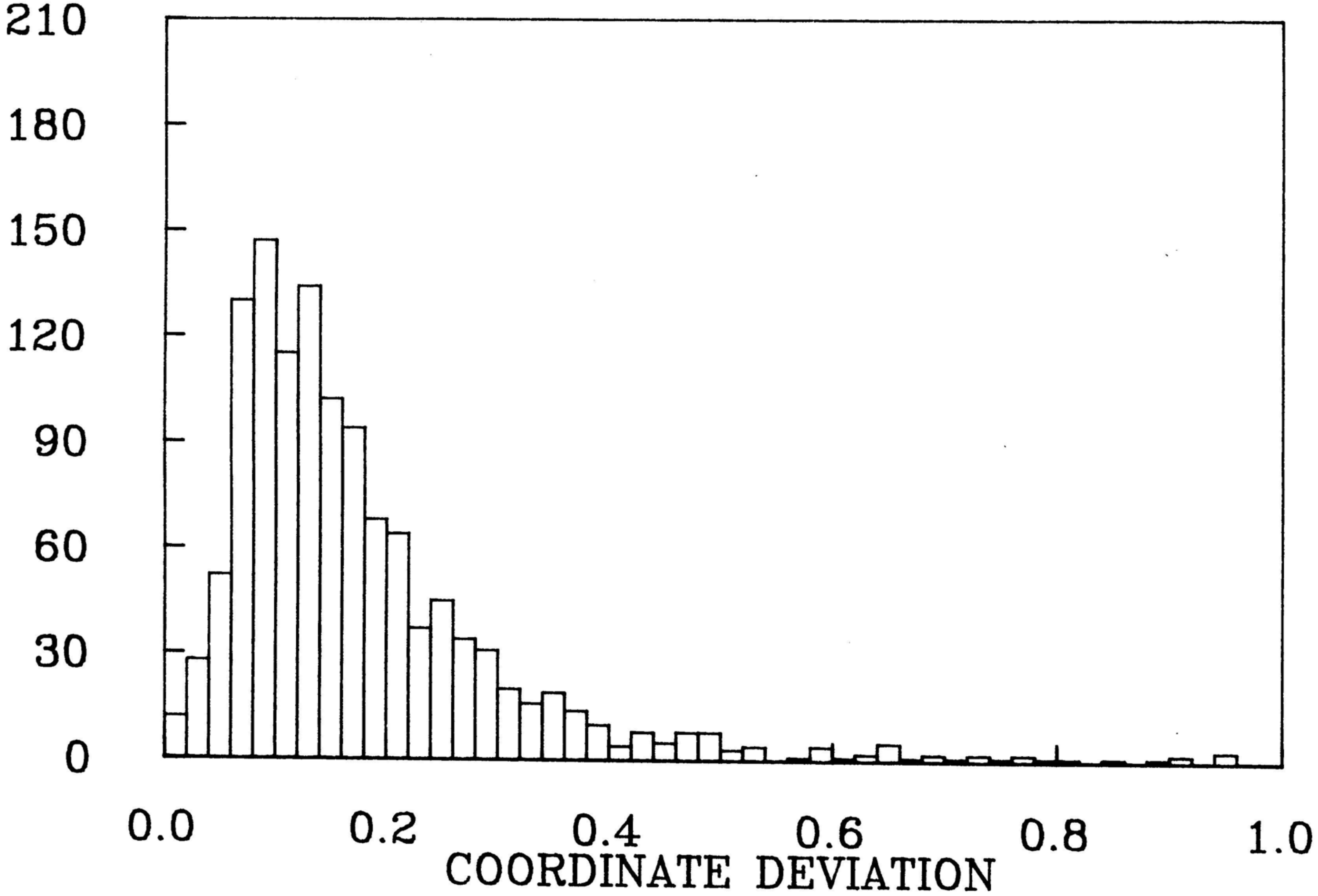

210
180
150
120
90
60
30
0
0.0
0.2
0.4
0.6
0.8
1.0
COORDINATE DEVIATION

FIGURE 3 (a)

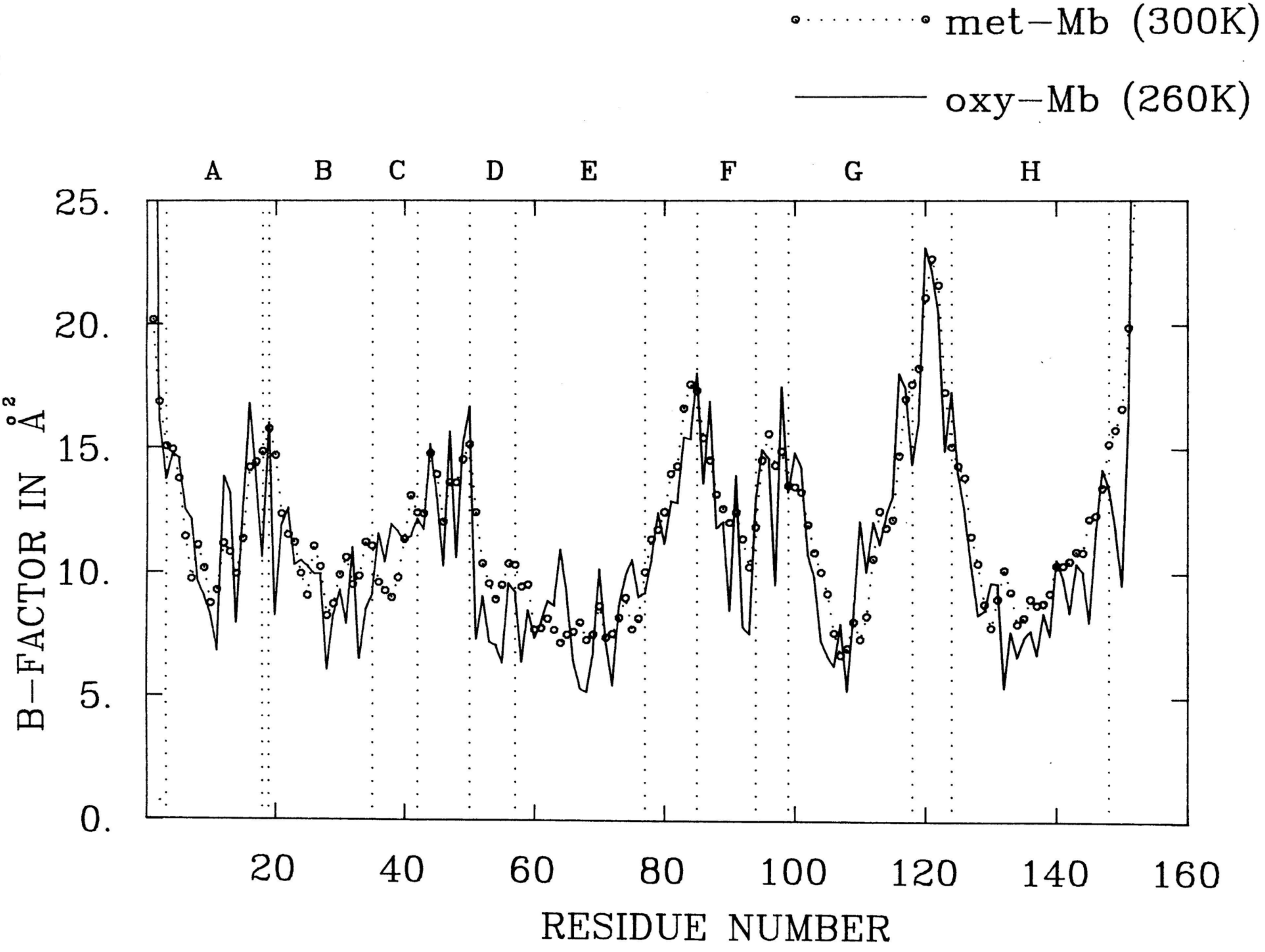

met−Mb (300K)
oxy−Mb (260K)
A
B
C
D
E
F
G
H
B−FACTOR IN Å^2
25.
20.
15.
10.
5.
0.
20
40
60
80
100
120
140
160
RESIDUE NUMBER

FIGURE 3 (b)

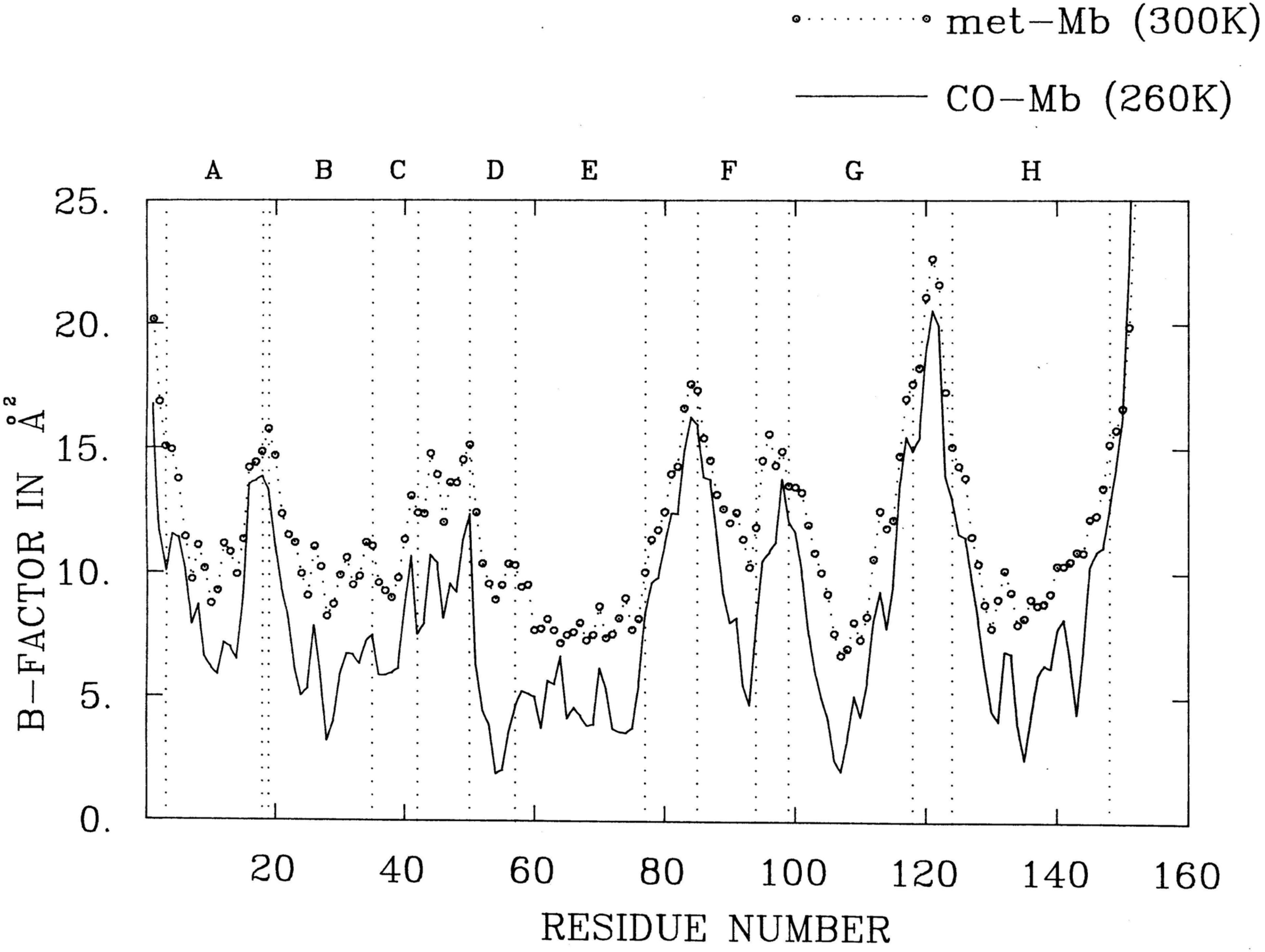

met–Mb (300K)
CO–Mb (260K)
A
B
C
D
E
F
G
H
B–FACTOR IN Å²
0.
5.
10.
15.
20.
25.
20
40
60
80
100
120
140
160
RESIDUE NUMBER

FIGURE 3 (c)

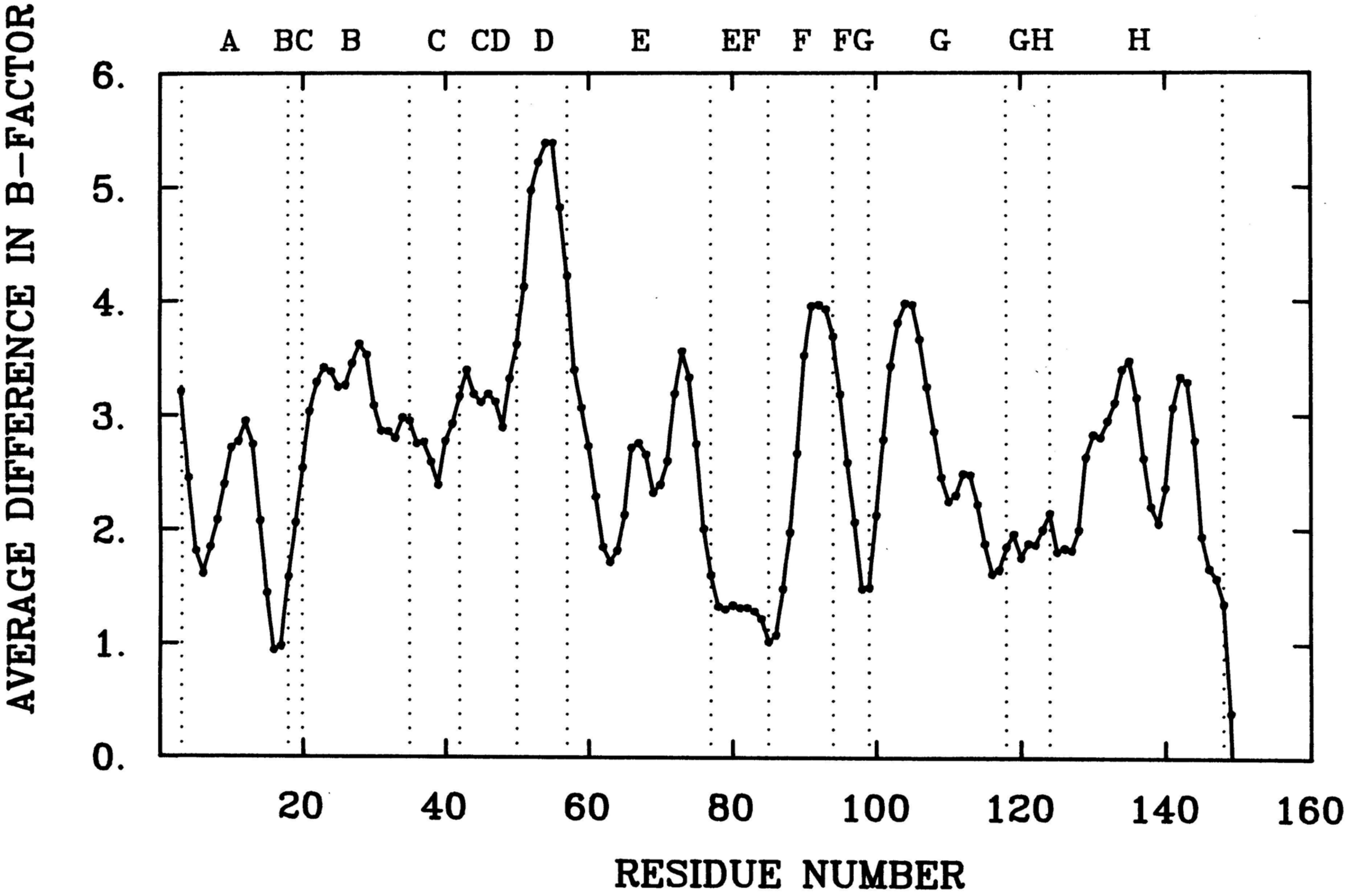

A
BC
B
C
CD
D
E
EF
F
FG
G
GH
H
AVERAGE DIFFERENCE IN B-FACTOR
6.
5.
4.
3.
2.
1.
0.
20
40
60
80
100
120
140
160
RESIDUE NUMBER

FIGURE 3 (d)

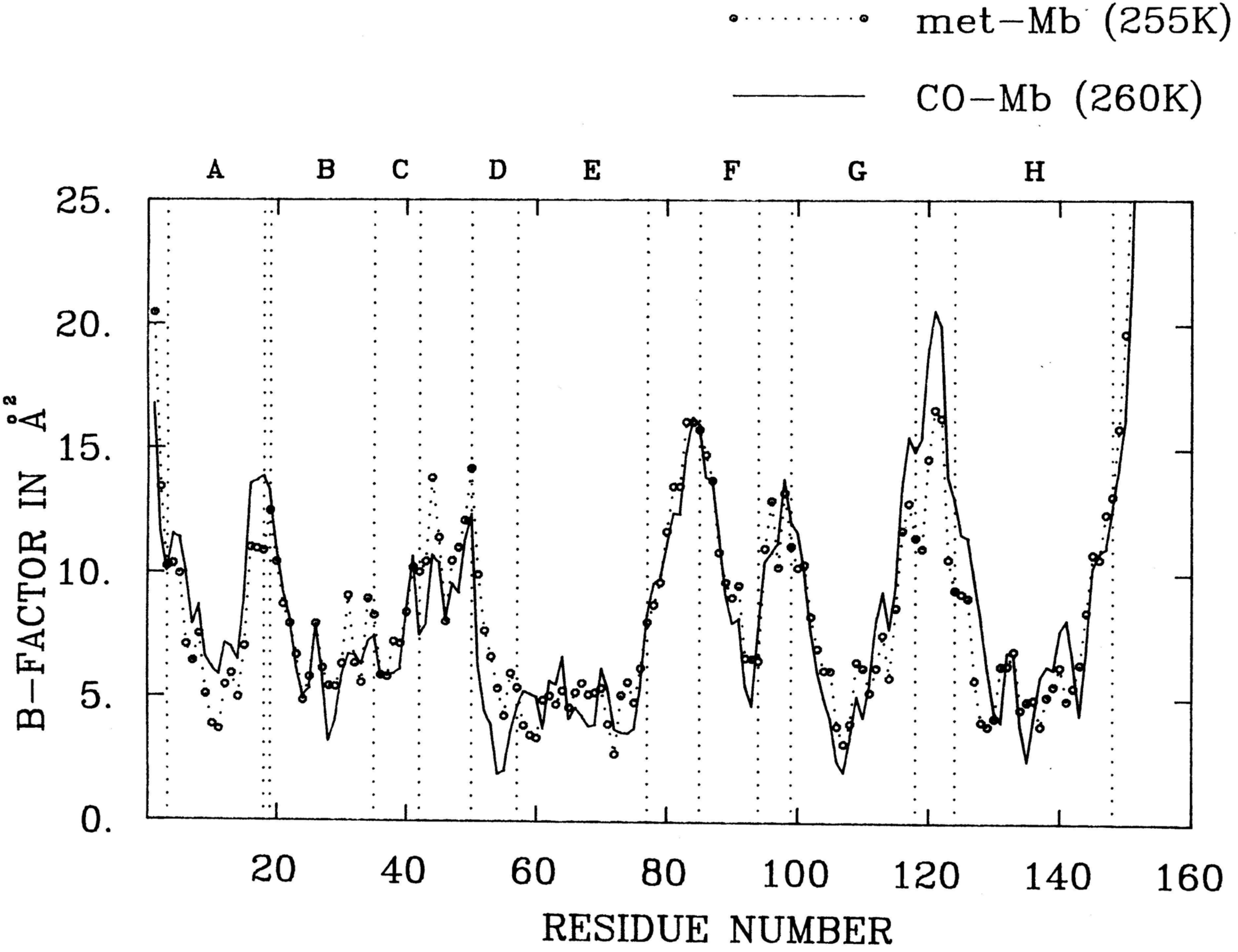

met−Mb (255K)
CO−Mb (260K)
A
B
C
D
E
F
G
H
25.
20.
15.
10.
5.
0.
B−FACTOR IN Å²
20
40
60
80
100
120
140
160
RESIDUE NUMBER

FIGURE 3(e)

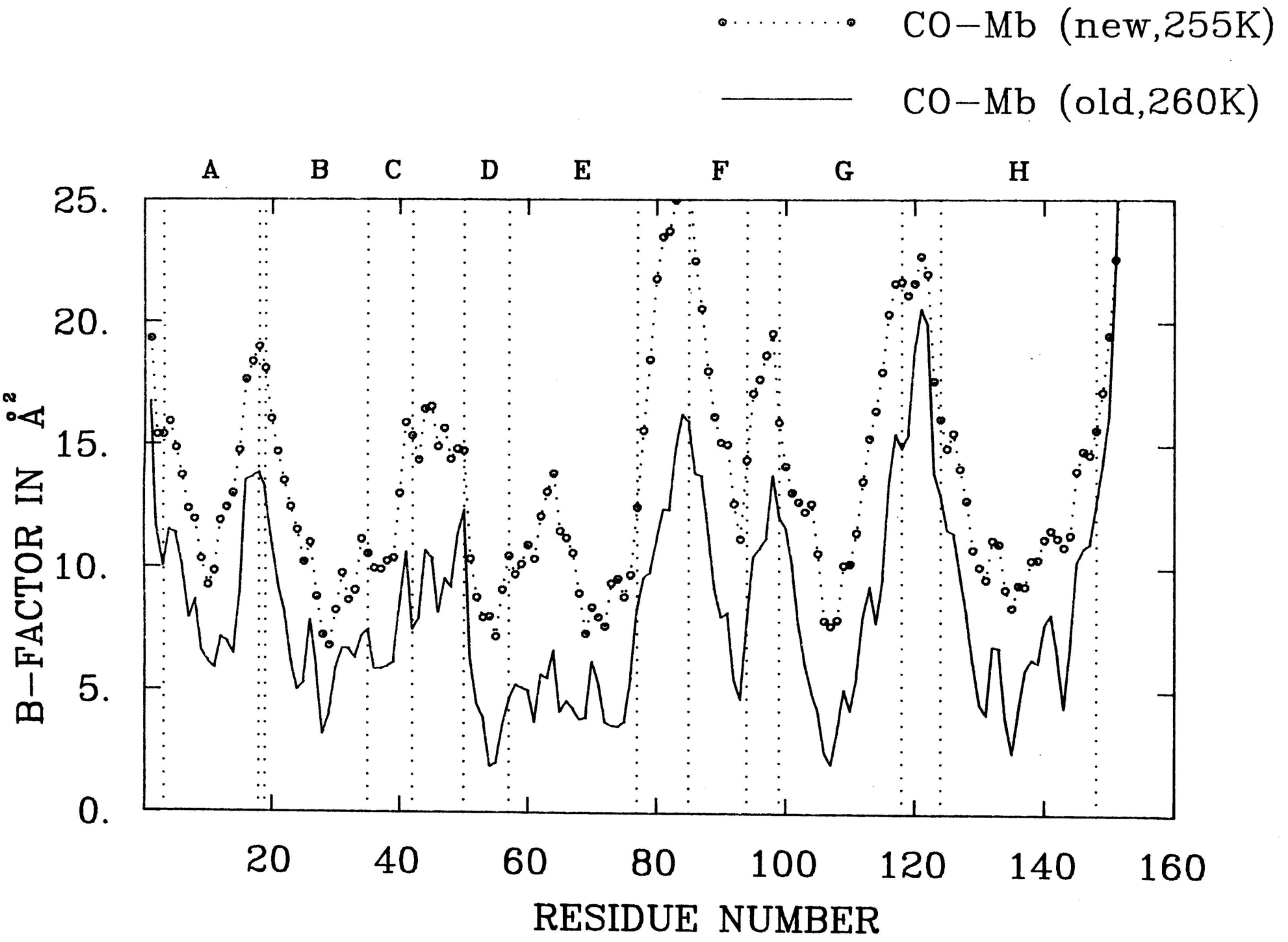

CO–Mb (new,255K)
CO–Mb (old,260K)
A
B
C
D
E
F
G
H
25.
20.
15.
10.
5.
0.
B–FACTOR IN Å^2
20
40
60
80
100
120
140
160
RESIDUE NUMBER

FIGURE 3(f)

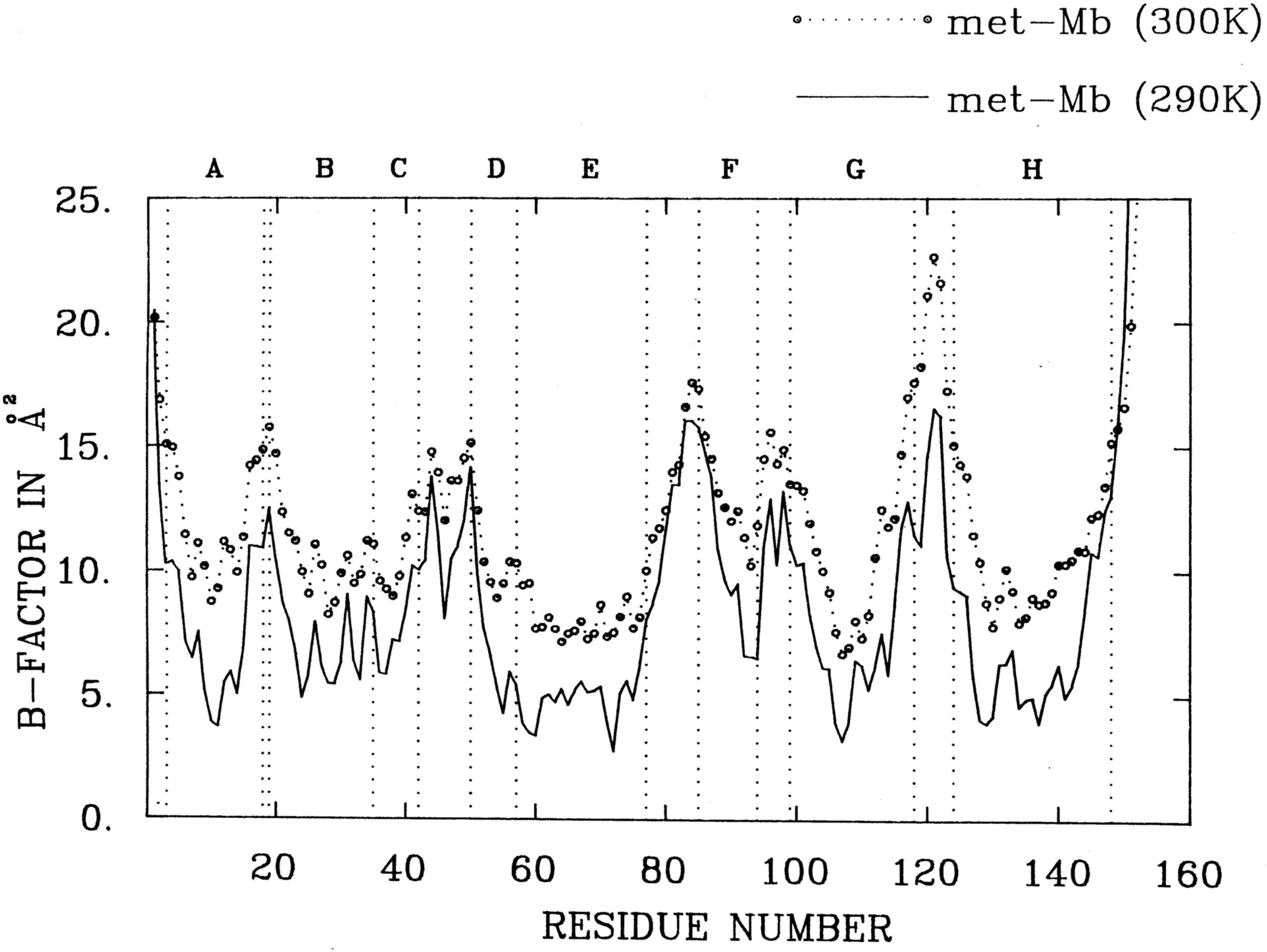
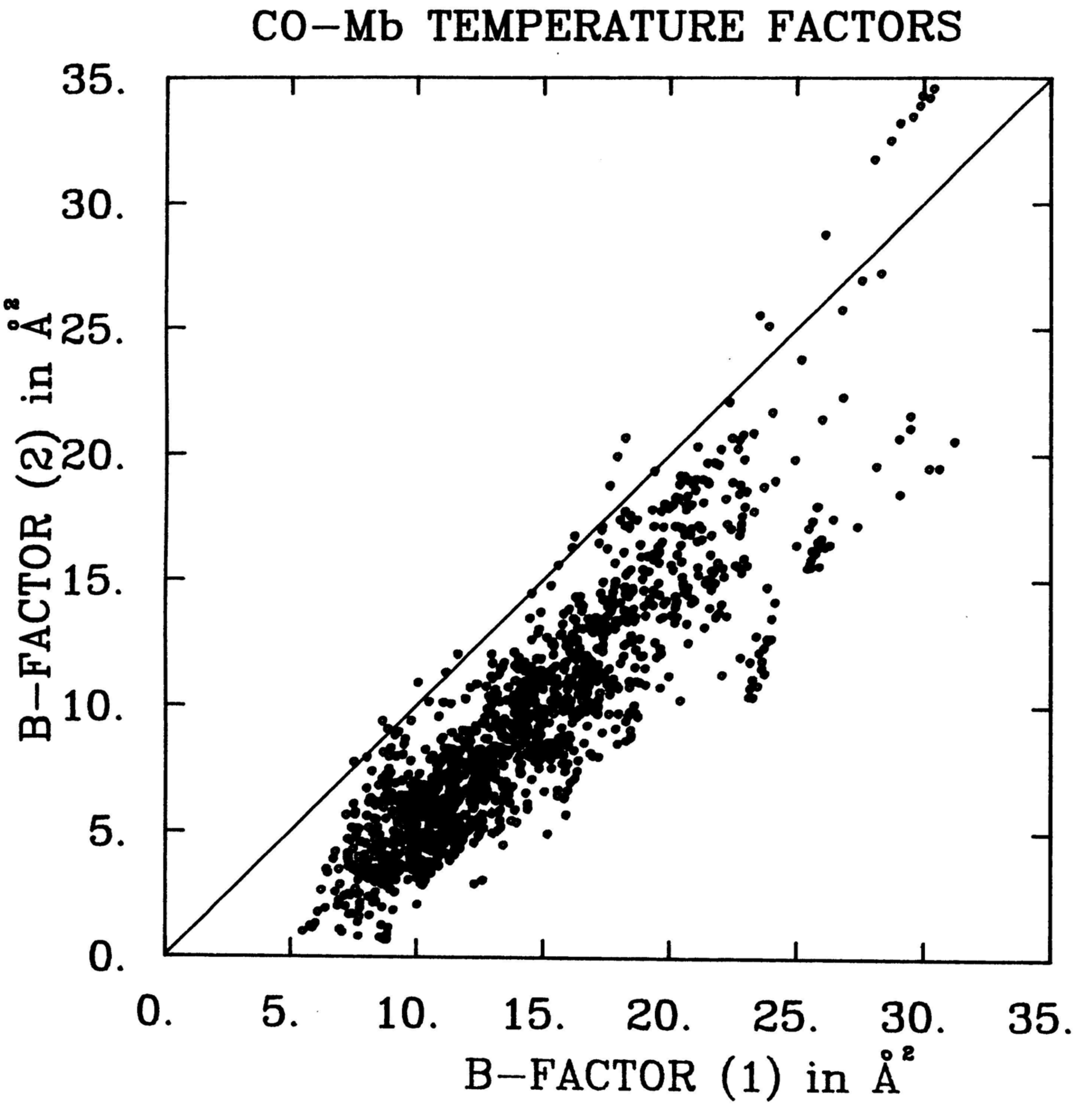

met−Mb (300K)
met−Mb (290K)
A
B
C
D
E
F
G
H
25.
20.
15.
10.
5.
0.
B−FACTOR IN Å²
20
40
60
80
100
120
140
160
RESIDUE NUMBER

FIGURE 4 (a)


CO–Mb TEMPERATURE FACTORS
B–FACTOR (2) in Å²
B–FACTOR (1) in Å²
0.
5.
10.
15.
20.
25.
30.
35.

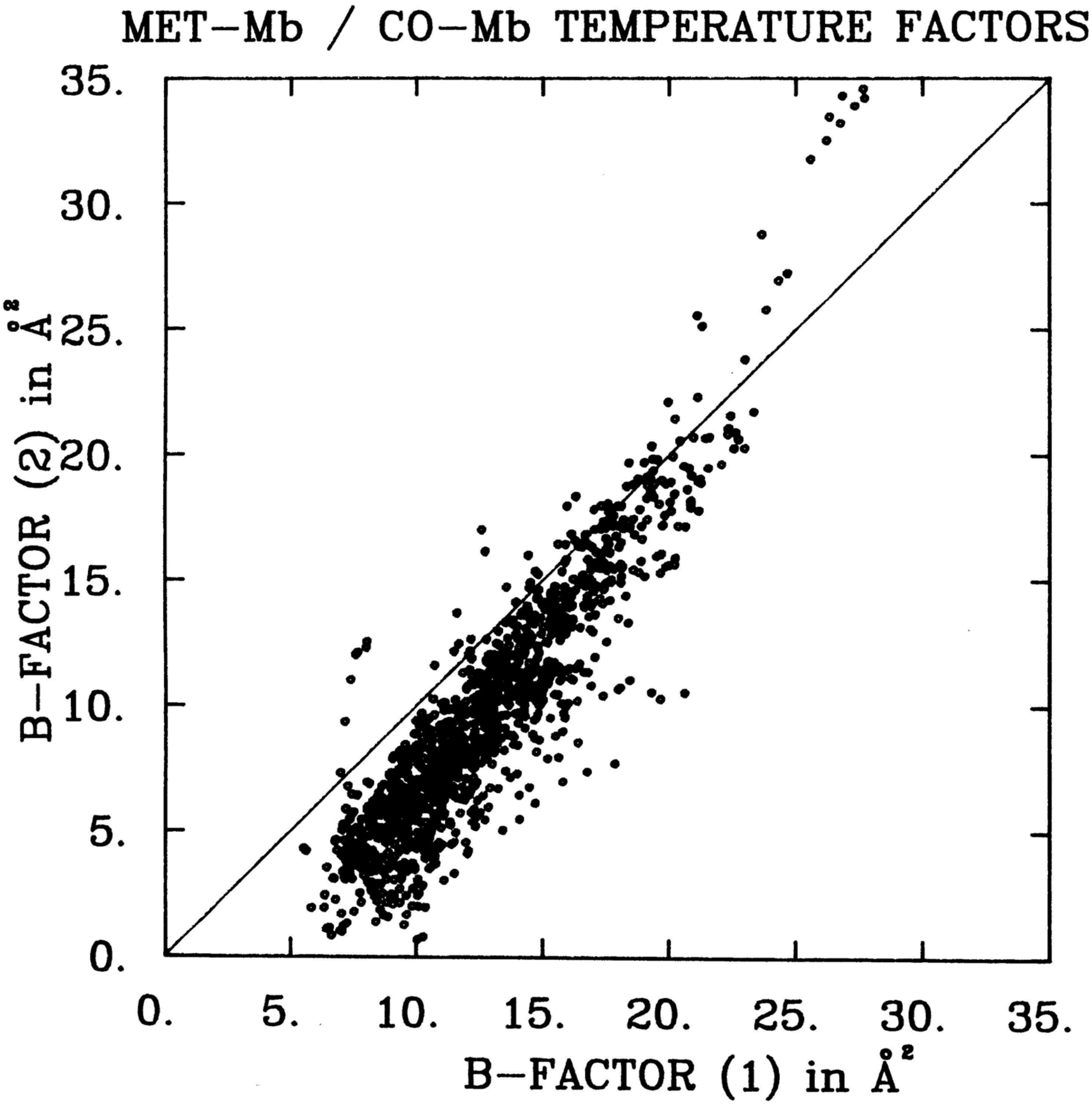

MET–Mb / CO–Mb TEMPERATURE FACTORS
B–FACTOR (2) in Å²
B–FACTOR (1) in Å²
0.
5.
10.
15.
20.
25.
30.
35.

FIGURE 4(b)

Acknowledgements

It is with great pleasure that I acknowledge the help and inspiration I have received from friends, relatives, teachers and colleagues. Without them this thesis would never have been completed.

I would like to thank Greg Petsko and Martin Karplus for encouragement, for providing wonderful research environments and for making a joint thesis possible. Working between two universities has been a very rewarding experience in itself, and working for these two particular advisors has been extremely stimulating and challenging.

I would like to thank all the people I have worked with at Harvard and MIT; they have made both research groups very exciting ones to have been a part of. At Harvard, I would particularly like to mention S. Swaminathan, Toshiko Ichiye, Wally Reiher, John Brady, Mike Cook, Bernie Brooks, Rich Pastor, Bob Bruccoleri, Axel Brunger, Jeremy Smith, Gary Hoffman, Bruce Tidor, Ann Giammona, Neena Summers, Carol Post, Lennart Nilsson, Charlie Brooks, Monte Pettitt, Hsiang-Ai Yu, Ron Elber, and our frequent visitor from York, Rod Hubbard, and also Bill Weis from the Wylie group.

At MIT I would like to thank Will Gilbert, Barbara Seaton (who proof read this thesis and is now at Harvard), Sherry Mowbray, Elias Lolis, Steven Burley, Bob Tilton, Rob Campbell, David Rose, John Dewan and David Neidhart. Bob Davenport provided much fun and excitement and also kept us going during the slower stretches. I am grateful to Mary Roberts for giving me a summer job in her research group at MIT before I started graduate school.

I would like to thank my parents, my sisters and Anna Mani for understanding, encouragement and support; Leslie and Freda Cook for helping me choose a graduate school and for encouragement and advice while I was at Juniata College; Bill and Janet Haines for their friendship and hospitality; Lily Harishchandra for providing relaxing breaks (with lots of tea) between bouts of thesis work; Dini Chandra for help, support and general comments on varied aspects of life; and, finally, my teachers at Juniata, especially Esther Doyle, Ruth Reed and Paul Schettler, for making my two years there enriching and memorable.

" Light thickens ... "

(Macbeth)