Title
Conformational change and catalysis in thymidylate synthase

Author
Fauman, Eric Benjamin

Publication Date
1993

Peer reviewed|Thesis/dissertation

Conformational Change and Catalysis in Thymidylate Synthase

by

Eric Benjamin Fauman

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biochemistry and Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA

San Francisco

Preface

Like a good starting model for molecular replacement, my stay in San Francisco has been a most enjoyable phase. True, the city has a special magic, but it is all the great people I have worked with since 1987 that I will miss most once I leave.

Whatever crystallographic knowledge I take with me was gleaned from the many fine crystallographers I've been privileged to work with over the years. The first was Ramu, during my rotation my first months at UCSF. I am also pleased to acknowledge Kathy Perry, Bill Montfort, Janet Finer-Moore, Robert Stroud, Michael Shuster, Thomas Earnest, Partho Ghosh and Earl Rutenber for much much help over the years.

Scientific insight can not develop in a vacuum, but requires extensive discourse and interaction. The people I most often relied on to lend a scientific ear were Robert Fletterick, Janet Finer-Moore, Brian Shoichet, Jim Hurley, Chris Carreras, Dan Santi, Partho Ghosh, Sasha Kamb and Earl Rutenber.

You probably won't believe this, but I did do some wet chemistry while at UCSF. The people who guided me through the strange territory of gels, columns and restriction enzymes were Peter Hwang, Nancy Craig and Jo Davisson.

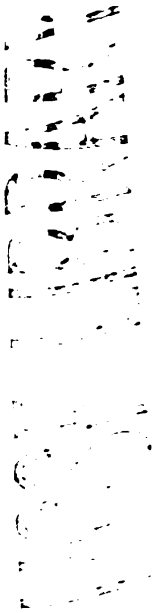
Being a graduate student, all my time was spent either at home or in lab. The people who most made lab a place I wanted to be were Susan Fong, Stephanie Mel, Earl Rutenber, and Paul Foster.

Although I didn't see them as much, I was fortunate to have fun roommates when I was home. Celia Schiffer, Kevin Turner, Ian Gould and Theresa Gamble were all house mates for a year or two. At 4.5 years, however, Wendy Cornell has been my house mate for longer than anyone

whom I wasn't related to! How is it possible we never even had a fight? And say hi to Maggie for me, who for two years was always waiting at the door for me when I came home.

Thanks go to Sue Adams for forcing me to have a thesis committee meeting (make sure you have them Paul, Bob, Bob, Chris and Louise!) and to the members of my thesis committee, Robert Fletterick, Dave Agard and Robert Stroud, for advice and counsel over the years.

And finally, never-ending thanks to my sanity and anchor, my wife, Cristina Lete. Throughout our 243 weeks, mostly apart but now nearly at an end, we stayed close through email and phone. See you soon. ILY.



Conformational Change and Catalysis in Thymidylate Synthase

Eric Benjamin Fauman

Abstract: Determination of the x-ray crystal structure of a protein is just the first step in relating that structure to biochemical, kinetic or genetic information. This thesis reports on several methods developed to compare crystal structures, primarily of the enzyme thymidylate synthase. Methods are presented to superimpose crystal structures so as to highlight conformational changes, and to quantify those conformational changes whether they be domain motions, motions of secondary structural units, or more subtle atomic shifts. Specific attention is given to the use of B factors in structure analysis and the assessment of uncertainty in crystallographic coordinates. These methods and others were used to analyze seven structures of thymidylate synthase, including a 1.83 Å structure of thymidylate synthase complexed with the reaction products, the highest resolution structure of thymidylate synthase reported to date. These analyses suggest that domain closure in thymidylate synthase may be electrostatically driven. The "capacitor model" states that the replacement of high dielectric water with the low dielectric ligands causes attraction between positive and negative domains on opposite sides of the active site. New information on the chemistry of the reaction comes from the 1.83 Å structure which reveals a crystallographically fixed water molecule which may serve to receive a proton from the substrate. In addition, this water may help the enzyme distinguish between the substrate and product nucleotides. An analogous water is proposed to exist in deoxyuridylate and deoxycytidylate hydroxymethylase, where it would be a reactant and become part of the hydroxymethyl addition.

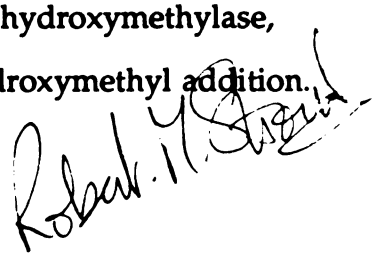


Table of Contents

List of Tables

List of Figures

Introduction

I entered the world of macromolecular X-ray crystallography because I enjoy spatial thinking, and I enjoyed the prospect of being able to understand enzymatic mechanism at an atomic level. As detailed in this thesis, my primary goal has been to understand the catalytic mechanism of thymidylate synthase, as essential enzyme in DNA synthesis.

At first, it seems straightforward to expect that once you have solved the crystal structure, the answers you have been seeking will be self-evident. However, as I quickly learned, having the structure is only the first step. The next step is finding a way to sort through the over 10,000 positional and thermal parameters which define that structure to find your answer. Chapter 1 is a discussion of the often ignored crystallographic parameter, the thermal or B factor.

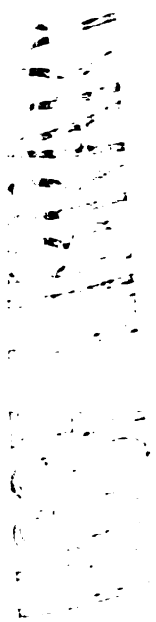
Chapter 2 describes the methods which I have developed to analyze protein structures, and more importantly, to compare pairs of protein structures to pick out the significant differences. When I joined the thymidylate synthase project (1987) there was one thymidylate synthase crystal structure. As I write this (1993) there are well over twenty different thymidylate synthase structures (different species, different mutants, different ligands), and the ability to quickly discern the relevant differences between structures is increasingly important.

Chapter 3 shows how the methods of Chapter 2 have been used to learn important aspects about the function of thymidylate synthase, how it accepts random mutations in evolution and how it uses conformational change to carefully align substrates and individual water molecules necessary for catalysis.

An understanding of conformation changes is important not only for answering mechanistic questions, but also for design of *de novo* inhibitors to known targets. The case of the HIV-I protease, in the final chapter, shows dramatically how the assumption of a rigid protein, commonplace in rational drug design, can sometimes be quite incorrect.

Chapter 1.

B factors



A. Definition and observations

At its simplest, a crystal structure is of a list of atom positions. With each atom is an associated probability density function (p.d.f.), which describes the observed variation of atomic position from the mean position. This variation arises from spatial averaging, from molecule to molecule in the crystal, and temporal averaging as the atoms vibrate during data collection. In protein crystallography, the p.d.f. most often takes the form of an isotropic three-dimensional Gaussian function. The Gaussian form can be derived as the sum of classical harmonic oscillators, each oscillator weighted by its Boltzmann weighting factor. Going from an isotropic to an anisotropic p.d.f. requires as additional five parameters per atom, which would yield more parameters than can be safely refined in a protein crystal structure.

The isotropic Gaussian p.d.f. is defined by a single parameter, the standard deviation or root mean square (rms) value of the Gaussian. Taking u to be the distance (along the crystallographic **s** vector) an atom is displaced from its equilibrium position, the atomic B factor is defined as:

$$B = 8\pi^2\langle u^2\rangle$$

It is important to note that u in the above equation is a one dimensional variable, that is, the distance from the mean atomic position along any one axis.

Atomic B factors typically have values between 2.0 $Å^2$ and 40.0 $Å^2$. The B factor is synonymous with the thermal or temperature factor. In contrast, the Debye-Waller factor is defined as:

$\exp(-2B\sin\theta^2/\lambda^2)$, where

B is the B factor,

θ is the Bragg scattering angle, and

λ is the wavelength of the radiation used.

The Debye-Waller factor has a value between 0 and 1 and describes the degree to which the intensity of a reflection is lessened by virtue of the atomic B factors.

An overall B factor for the protein can be calculated from a Wilson plot. Intensity of the crystallographically obtained reflections decreases with the increasing resolution of the reflection both because the electrons in an atom are not confined to a single point and because of the thermal smearing specified by the p.d.f. By dividing out the effects of atom size, a Wilson plot allows direct measurement of the decrease in reflection intensity due to the B factor.

Atomic B factors are usually simply refined so as to minimize the difference between the observed and calculated structure factors. Because atomic bonds are quite rigid, it is sensible that the thermal smearing or B factor of one atom be similar to that of a covalently bonded neighbor, and this is often a restraint imposed during B factor refinement.

In contrast to this rather weak restraint on B factors, the atomic positions are subject to a very large number of restraints from knowledge of bond lengths, bond angles and dihedral angles. The greater freedom afforded the B factors is what allows B factors to artificially lower the crystallographic R-factor, and is why B factors are sometimes called "trash-cans" in refinement.

I have observed that the average of all atomic B factors in a crystal structure is often quite close to the Wilson plot derived overall B factor. I have also found it useful to generate a Wilson plot from the calculated structure factors to make sure the B factors in the structure are well-behaved.

I have also observed that a plot of the distribution of B factors in a structure (see function BCOUNT in the program GEM in the appendix) can be fit by a Maxwellian distribution. This seems very consistent, yet I have never seen it discussed. I have thought of three possible explanations for this phenomenon: 1) Since the B factor is proportional to a distance squared, it is proportional to energy, and the distribution of B factors reflects a Boltzmann distribution of vibrational energy among the atoms in the structure. 2) The B factor values are randomly chosen from a Gaussian with a specific mean and standard deviation, and the refinement program doesn't allow B factors below 2, so it just appears Maxwellian. 3) The B factor values are often harmonically restrained to be within 2 $Å^2$ of their neighbor's value. This generates a random walk in B factor values going along the protein chain. This restraint, along with the restriction that no B factor be less than 2 Å may by itself be able to generate the observed Maxwellian distribution.

Petsko, et al., in their paper on cryocrystallography showed graphs of B factor distribution as a function of temperature, and the graphs show the behavior expected for the distribution of velocities in an ideal gas as a function of temperature. However, they make no note of the physical significance of these data.

B. B factors and errors

As noted in the preceding section, because of the relatively weak restraints on B factors, B factors tend to absorb errors in the structure. For example, if a sidechain is built where there should be no atoms, the refined B factors for that sidechain will go up so as to spread out and minimize the electrons at that point. In other words, misplaced atoms are more likely to be assigned large B factors.

On the other hand, atoms with large B factors are more likely to be misplaced. For example, if a residue is extremely floppy, it is very difficult to model that as a single sidechain, or to discern the electron density for that sidechain at all. Even if a sidechain is accurately modeled as a single conformation with isotropic vibrations, the larger the atomic B factors, the harder it is determine the atomic positions. With infinite and perfect data, this would not be true, but errors in the observed structure factors (Fo) and the limited resolution of the data introduce noise into the electron density maps. A larger B factor means a smaller peak in the electron density map, and hence it becomes harder to pick the center of the peak.

This phenomenon can be modeled by fitting a Gaussian curve to noisy data. I have done this and observed that indeed the error in calculating the center of the peak increases as the peak height approaches the level of the added noise.

The following paper studies the relationship between errors and B factor values. Here I show that not only can we can we observe this effect, but it is stable enough that we can generate a formula which will predict the uncertainty in an atom's position based upon its B factor.

Errors from B factors: An empirical approach to analysis of accuracy in x-ray protein structures by reference to atomic B factors.

Eric B. Fauman and Robert M. Stroud

Department of Biochemistry and Biophysics

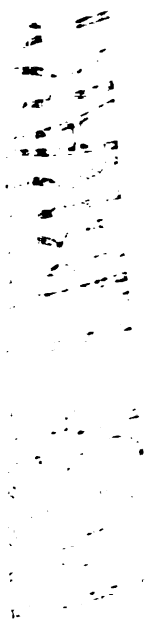
School of Medicine

University of California, San Francisco

San Francisco, CA 94143-0448

In a crystallographic structure determination, each atom has a refined position as well as an atomic thermal B factor. The B factor is related to the second moment of the atomic position through the equation $B = 8\pi^2\langle u^2\rangle$, where u is the displacement of the atom from its equilibrium position in any one dimension. It has long been recognized that an atom's B factor is related to the accuracy of its position, namely, that the median positions for atoms with lower B factors are better determined. We present here a quantitative relation between an atom's B factor and the accuracy of its position obtained through a statistical analysis of existing macromolecular crystal structures. Eighteen structures from the Protein Data Bank with multiple molecules in the asymmetric unit were used, and the positional differences between the molecules were parameterized on the basis of atomic B factor. After testing many potential indicators of structure accuracy, a factor containing the ratio of the number of reflections to the number of atoms used in refinement was

found to best correlate with the magnitude of the positional differences in different structures. These results are embodied in a six parameter equation which can be applied to other refined structures to estimate the uncertainty in position of each atom. The accuracy of an atomic coordinate within a macromolecular crystal structure is further found to correlate with the connectivity of the atom. However, the accuracy shows little or no correlation with atomic number, most likely because the connectivity is a dominant effect in macromolecular structures.



Nomenclature

$\Delta x, \Delta y, \Delta z$	Difference in position for a single atom in a pair of structures along the x, y or z axis, respectively.
χ	Generic one dimensional axis. Equivalent to average over all possible coordinate axes.
σ_χ	one dimensional standard deviation of the Gaussian portion of positional differences in a pair of structures.
Δr	$\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}$
σ_r	standard deviation of the Maxwellian distribution of positional differences in a pair of structures. $\sigma_r = \sqrt{3}\sigma_\chi$.
$\Delta\mathcal{R}$	Distance between the observed position of an atom and its "true" position.
$\sigma_\mathcal{R}$	standard deviation of the Maxwellian distribution of ΔR. $\sigma_\mathcal{R} = \frac{\sqrt{2}}{2}\sigma_r$.
ATOM	number of independently refined atom positions in the asymmetric unit
REFL	number of independent reflections used in refinement
$\mathcal{E}_\chi$(B,ATOM/REFL)	empirically derived estimate of σ_χ as a function of B factor and the ratio of ATOM/REFL for a given structures.
N.E.S.(subset)	Normalized Error Score, defined as the deviation from $\mathcal{E}_\chi$(B,ATOM/REFL) for a selected subset of atoms.
N.E.S.$_i$(subset)	Normalized Error Score calculated using only structure i of the 18 structures used in the analysis.
σ_{NES}(subset)	Standard deviation of the 18 values for N.E.S.$_i$(subset).

Introduction

Knowledge of the level of accuracy in macromolecular crystal structures is important in evaluating crystallographic results and in identifying significant structural differences between pairs of crystal structures. Normally, obtaining statistics on a complex physical measurement requires multiple sampling of the desired observation. A few examples exist of structures determined by more than one group (Chambers & Stroud, 1979; Clore & Gronenborn, 1991) or in different crystal forms (Kossiakoff, Randal, Guenot & Eigenbrot, 1992). Also, closely related protein structures can be compared (Chothia & Lesk, 1986) to analyze levels of errors.

Structures which have been solved with more than one molecule in the asymmetric unit represent a particularly rich source of information on accuracy in crystallography since within each structure variables such as crystallizing conditions, primary sequence, crystal habit, data collection strategy, resolution of the intensities and refinement methodology are the same for both molecules. We report here on the analysis of eighteen cases of structures with multiple, independently refined, molecules in the asymmetric unit.

The random errors, derived from observed differences in the structures, were evaluated and parameterized first with respect to B factor. After testing a number of potential indices of model quality, a factor containing the ratio of parameters to observations used in the refinement was found to correlate best with the level of error. The resulting empirical error formula can be applied to any refined macromolecular structure to obtain an estimate for the error of each atom. With the B factor plus an overall value

for model quality accounted for, the effect of other atomic attributes were evaluated for their contributions to positional uncertainty.

Preparation of structures used in analysis

Structures

Eighteen structures that contain multiple molecules in the asymmetric unit were found by searching the Brookhaven Protein Data Bank (PDB)(Bernstein, Koetzle, Williams, Meyer, Brice, Rodgers, Kennard, Shimanouchi & Tasumi, 1977) all PDB files containing the word "ASYMMETRIC" (Table I). Each structure accepted for the study contains at least two molecules in the asymmetric unit. In no case was non-crystallographic symmetry used throughout refinement. In some cases the crystallographers used one or more of the following techniques: the same starting structure was used for independent molecules; non-crystallographic symmetry was applied at the earliest stages of refinement; dihedral angles were averaged over the independent molecules early in refinement; the atomic positions were averaged early in refinement; changes to be made to one molecule were checked against the electron density maps for the other independent molecule(s). These techniques will all tend to reduce the observed positional differences.

For those structures which contain four independent molecules in the asymmetric unit (1HBS, 1HMQ, 1GD1, 2PFK), six separate comparisons were performed and then pooled and averaged to represent a single data point for that structure.

In preparing the structures, atoms were excluded if they had no refined B factor or a low occupancy. Eighty-seven atoms were eliminated because they or their analogous atom in the other molecule of the asymmetric unit

had a B factor of zero. Also, for atoms with multiple occupancies, only the greater occupied site was used. These criteria eliminated 45 atoms. In all, 70,120 atoms were included in the analysis representing 28,720 "multiply observed" atoms.

Overlap

Before evaluating differences in positions the independent molecules in each asymmetric unit were overlapped by minimizing the root mean square (rms) deviation of a set of core Cα's. The core of Cα's was selected by scanning a difference distance matrix looking for the largest set of Cα's which simultaneously fulfills the following two criteria: 1) every Cα in the core moves less than 0.5 Å relative to every other Cα in the core, and 2) the set is connected in a mathematical sense, with each Cα in the core being within 10 Å of at least one other Cα in the core. The core so selected varies from 15 to 70% of all Cα's in the protein, depending on the overall level of difference between the proteins being compared. The core was selected by the program NewDome, previously described (Montfort, Perry, Fauman, Finer-Moore, Maley, Hardy, Maley & Stroud, 1990; Perry, Fauman, Finer-Moore, Montfort, Maley, Maley & Stroud, 1990).

Selecting rotamers

The side chains of Phe, Tyr, Asp, Glu and Arg are two-fold symmetric about the last free dihedral angle and thus the labeling of certain atoms in these side chains is arbitrary with respect to the chemistry of the amino acids[1].

[1]There is a convention for unambiguously labeling these residues; the atoms should be named so as to give the lower dihedral angle value for the appropriate atoms. However, as noted by Morris et al. (1992), this convention is rarely if ever used by macromolecular crystallographers.

It was necessary therefore to specifically analyze these side chains after overlap to see if the rms deviation between the structures could be decreased by interchanging the two-fold symmetry related atoms. In addition, the side chains of His, Asn, and Gln were analyzed since crystallographically these side-chains are also two-fold symmetric in all but the very best structures. This affected 120 residues out of a total of 4,500 in the analysis. The overall rms deviation between pairs of structures was decreased by as much as 0.5%.

Generation of the empirical error curve, $\mathcal{E}_\chi(B)$

Extracting errors

There are real differences between multiple molecules in the asymmetric unit due to differences in the packing environment of the independent molecules. To examine differences which have random distribution we extracted all those differences which followed a normal or Gaussian distribution. It is an underlying premise that the normally distributed differences between these pairs of structures are due to the errors in the structures. Crystal contacts and other systematic differences were screened by fitting a Gaussian to the observed distribution of one dimensional differences.

In the first step, we examined each structure for any possible B factor dependence on the level of errors. In each structure comparison a scatter plot was constructed of Δr (the difference in atomic position in the pair of structures) versus the mean B factor assigned to the atom in the two structures (Figure 1). Running bins in B factor were constructed with a width of +/- 2 $Å^2$ and an increment of 0.1 $Å^2$. The distribution of Δr values within a single B factor bin can be displayed as a histogram, as in Figure 2. This histogram has the form of a Maxwellian distribution:

$$P(\Delta R) = \frac{\Delta R^2}{\sigma_R^3} \sqrt{\frac{54}{\pi}} e^{-\frac{3}{2}\frac{\Delta R^2}{\sigma_R^2}} \quad (1)$$

where $P(\Delta r)$ is the probability of obtaining a given value of Δr, and σ_r is the three dimensional standard deviation of Δr.

In order to work with a Gaussian distribution, rather than a Maxwellian distribution, each value of Δr was replaced with its one-dimensional probability distribution. That is, a given value of Δr has a certain probability of having an Δx component with a given value (derived in Appendix I). The probability of a specific value of Δx, is:

$$P(\Delta X) = \begin{cases} \frac{1}{\Delta R}; & 0 \le \Delta X < \Delta R \\ 0; & \Delta R \le \Delta X \end{cases} \quad (2)$$

We denote the generic one-dimensional axis by the letter χ, which can be thought of as the X axis rotated over all possible orientations. Thus, this conversion from three dimensions to one dimension simultaneously provides the advantage dealing in a one-dimensional variable (where Gaussian (Figure 3) rather than Maxwellian (Figure 2) distributions hold) while avoiding the arbitrariness of any given orientation of the coordinate basis vectors.

For each bin in B factor (e.g. 10 Å^2 to 14Å^2 as in Figures 2 and 3) a histogram was constructed of frequency versus one-dimensional difference in position (Figure 3). The standard deviation of the Δr values, σ_r, can be estimated by the rms (root mean square) value of the Δr. Since σ_χ, the one-

dimensional standard deviation, is related to σ_r by $\sigma_r = \sqrt{3}\sigma_\chi$, the abscissa of the histogram was binned in divisions of 1/17 of the rms value of the Δr values at the given B factor, or roughly 1/10 of the expected σ_χ.

A Gaussian distribution of the form:

$$P(\Delta\chi) = Ae^{-\frac{1}{2}\left(\frac{\Delta\chi}{\sigma\chi}\right)^2} \quad (2)$$

was then fit to the histogram through non-linear, unweighted least-squares minimization (Figure 3). This extracts the true normal distribution component of the differences in atomic positions. The standard deviation of this distribution is termed $\sigma_\chi(B)$, i.e. the one-dimensional standard deviation in atomic position at the given B factor. The scatter plot of ΔR versus B factor (Figure 1) is thus replaced with a curve of $\sigma_\chi(B)$ versus B factor (Figure 4).

Consistent with the hypothesis that $\sigma_\chi(B)$ reflects the errors in the crystal structure is the observation that the extracted differences diminish as the resolution increases. This is true, for example, in the case of azurin (see below), for which two structures, a medium resolution and a high resolution structure, were used. The differences between the two molecules decreased with the addition of the high resolution data. Other measures of crystal structure accuracy, such as dihedral angle quality and energy of hydrogen bonds, also improve with increasing resolution (Morris, MacArthur, Hutchinson & Thornton, 1992).

Errors are correlated with atom B factor

To generate a smooth B factor dependence, the graph of $\sigma_\chi(B)$ versus B factor was fit to an exponential of the form

$$\sigma\chi(B) = a + b * e^{(B/c)} \quad (3)$$

where a, b, and c are the refinable parameters (Figure 4). This functional form for the B factor dependence was selected over the parabolic form previously used ($\sigma_\chi(B) = a+b*B+c*B^2$ (Perry, Fauman, Finer-Moore, Montfort, Maley, Maley & Stroud, 1990)) because 1) the exponential used here is monotonically increasing, and 2) the exponential fit to the $\sigma_\chi(B)$ curves generally resulted in a smaller least squares value than the corresponding parabolic equation fit to the same curve. The values of a, b and c for each protein were refined by non-linear least squares minimization to data from all atoms with B factors between 0 Å^2 and 40 Å^2. A B factor cutoff of 40 Å^2 was imposed since in most cases there were not enough data points ($n < 100$) to obtain reliable estimates of $\sigma_\chi(B)$ for values of greater than 40 Å^2. This is not a great limitation, since 90% of the 70,000 atoms used in the analysis had B factors less than 40 Å^2. In some cases, proteins did not have enough atoms for the curve to extend all the way to 40 Å^2. In these cases, the exponential was fit to the reduced range and the limitation on range was noted. In addition, any limitation on the range for very small B factors was also noted for each structure. The B factor limitations for each structure are apparent in Figure 7, below. In this manner, each plot of $\sigma_\chi(B)$ *vs.* B factor was represented by the three parameters a, b and c.

Previous publications have shown that positional errors increase with increasing B factor (Cruickshank, 1949; Chambers & Stroud, 1979; Bott & Frane, 1990; Perry, Fauman, Finer-Moore, Montfort, Maley, Maley & Stroud, 1990), although the current report is the first to use the empirically derived three-parameter exponential form: $a+b*e^{(B/c)}$. In particular, Cruickshank derived the following formula for the one-dimensional standard deviation of uncertainty in atomic position:

$$\sigma_x = \frac{\sigma(A_h)}{A_{hh}}, \text{ where} \qquad (4)$$

σ_x is the standard deviation in the x direction for an orthorhombic space group, $\sigma(A_h)$ is the standard deviation of the first derivative (with respect to x) of the electron density in an Fo map, and A_{hh} is the second derivative (with respect to x) of the electron density, or the curvature, at the atom center. Since an atom with a larger B factor will have a smaller curvature, the Cruickshank formula predicts atoms with larger B factors will have larger positional errors. Although no analytic expression is given for this B factor dependence, an error curve generated by the Cruickshank formula can be fit very well by a three-parameter exponential as used in this report.

Correlation of errors with measures of model quality

The dependence of errors on B factor for each structure is contained in the values of a, b and c for that structure. However, the values of a,b and c are different for each structure, the $\sigma_\chi(B)$ curves are all different (see Figure 7, below) and it is apparent that a single exponential curve does not suffice for all the structures. Other factors, such as resolution, which do not affect the atomic B factors, do affect the accuracy of the atomic positions.

In order to generate a family of exponential curves, we next chose to look for a parameter relating the different curves obtained for the different structures. To examine the dependence of errors on the quality of the model we plotted the value of $\sigma_\chi(B)$ for each structure *at a given B factor* versus ninety different potential indices of model quality (i.e. resolution, R factor

($R=\Sigma|(|F_o|-|F_c|)|\ /\ \Sigma|F_o|$), number of independent reflections, number of refined atom positions and functions of these). For each index of model quality we evaluated the correlation coefficient (r) from a linear least squares fit to the plot. The least squares lines are described by a Slope(B) and an Intercept(B), which are then functions of B factor.

Due to the limitations noted above, the σ_χ(B) *vs.* B factor curves for all 18 structures exist simultaneously only in the range 14 Å^2 to 31 Å^2 which accounts for 64% of the atoms used. The index of quality with the highest average correlation coefficient in this range is $e^{(-2ATOM/REFL)}$ where REFL is the total number of reflections used in refinement and ATOM is the total number of atoms in the asymmetric unit subject to refinement (Figure 5). This index yielded an average correlation coefficient of 0.89, and ranged from a maximum of 0.94 at a B factor of 14 Å^2 to a minimum of 0.76 at a B factor of 31 Å^2. For 18 data points, a correlation coefficient of 0.6 is statistically significant at the 99% confidence limit. That is, a random collection of 18 points has only a 1% chance of yielding a correlation coefficent greater than 0.6.

In this study, the errors in a crystal structure were more closely correlated with the ratio of the number of atoms to the number of reflections than to the resolution of the structure (average correlation coefficient of 0.79). As demonstrated in Appendix II, however, this ratio is proportional to the cube of the resolution, multiplied by the protein fraction in the unit cell.

The dependence of positional accuracy on the ratio ATOM/REFL can be understood in terms of the overdeterminancy of the crystal structure refinement, that is, the ratio of observations to parameters. In an unrestrained crystallographic refinement, the observations are the independent reflections and the free parameters are the positions and B-

factors of the atoms in the asymmetric unit. In macromolecular crystallography, the number of atoms involved prohibits completely unrestrained refinement, so restraints and constraints are applied. The presence of these restraints and constraints makes it impossible to calculate the precise overdeterminancy in macromolecular refinement, but it is still related to the number of atoms and number of reflections. The overdeterminancy will also be related to the exact number and nature of the restraints and constraints employed in the refinement.

The presence of the solvent fraction in the relationship between resolution and ATOM/REFL suggests that given the same protein in two different space groups, the one with the higher solvent content would yield the more accurate structure. This is because a higher solvent content means a larger unit cell, and consequently more reflections at a given resolution. However, in practice, crystals with a higher solvent content tend to diffract to a lower maximum resolution.

Interestingly, the R factor of the structure did not correlate significantly with the level of errors observed (average correlation coefficient of 0.57, and below 0.60 in all B factor bins). This could be due in part to the different conventions used for reporting R factor. For example some crystallographers apply a 2 sigma cutoff (i.e. $F/\sigma_F > 2.0$), which will give a lower R factor than if no cutoff is used. Also, all the structures used were final reported structures, and the R factor is most useful is evaluating the progress of crystallographic refinement. This means that the results obtained in this study are only applicable to other structures which have been completely refined, and not structures still in refinement.

Calculation of expected errors correlated with B factor

Slope(B) and Intercept(B) are derived by linear least squares analysis. The $\sigma_\chi(B)$ for each structure are three parameter exponentials (Eq. 3), and as a result, Slope(B) and Intercept(B), which are related to the $\sigma_\chi(B)$ in a linear matter, can also be described by three parameter exponential functions.

A plot of the Slope(B) *vs.* B factor fits exactly to a three parameter exponential curve:

$$Slope(B) = k1 + k2 * e^{(B/k3)}, \text{ where:} \quad (5)$$

k1 = -0.687222,

k2 = -0.002238 and

k3 = 6.162167

Likewise, a plot of Intercept(B) *vs.* B factor yields an additional 3 parameters:

$$Intercept(B) = k4 + k5 * e^{(B/k6)}, \text{ where:} \quad (6)$$

k4 = 0.642091,

k5 = 0.008518 and

k6 = 7.880717

Thus, all the information relating the B factor of an atom to the accuracy of its position is contained in these six parameters.

The expected error at each B factor, $\mathcal{E}_\chi(B)$, for a particular protein is then a function of the ratio of ATOM/REFL:

$$\mathcal{E}_\chi(B, ATOM/REFL) = Intercept(B) + Slope(B) * e^{(-2*ATOM/REFL)} \quad (7)$$

where Intercept(B) and Slope(B) are defined by equations 5 and 6.

For the analysis which follows, $\mathcal{E}_\chi$(B) for a single structure was recast as an exponential function of three variables:

$$\mathcal{E}_\chi(B) = p1 + p2 * e^{(B/p3)} \qquad (8)$$

as described in Appendix III.

This gives rise to a family of exponential curves of expected error *vs.* B factor each at a different value of ATOM/REFL (Figure 6). Figures 7a through 7r show the observed error curves, σ_χ(B), for the 18 structures along with the predicted error curves, $\mathcal{E}_\chi$(B), calculated according to the above equation.

Internal Control

Since the curves were constructed primarily from the 60% of atoms with B factors between 14 Å^2 and 31 Å^2, the resulting error curves were tested to see how well they predicted the differences between the structures used to derive them. To evaluate how well the function $\mathcal{E}_\chi$(B,ATOM/REFL) explained all differences in atomic positions between the pairs of structures, a Z-score was defined for each atom as:

$$Z - score = \frac{\Delta x}{\mathcal{E}_\chi(B)}, \frac{\Delta y}{\mathcal{E}_\chi(B)} \quad or \quad \frac{\Delta z}{\mathcal{E}_\chi(B)}, \text{ where} \qquad (9)$$

Δx, Δy and Δz are the differences in the position of an atom in the two molecules in the asymmetric unit, along each of the orthogonal axes.

If the error curves, $\varepsilon_\chi(B)$, really do reflect a normal distribution of differences, the distribution of Z-scores should be Gaussian with a standard deviation of 1.0. As shown in Table II, 13 of the 18 structures had an overall standard deviation within 20% of 1.0, and only one structure (2PFK) deviates from its $\varepsilon_\chi(B)$ curve by more than 50%. Thus the variation in positional uncertainty of most of the 29,280 atom pairs in the study is substantially contained within the 6 parameters (k_1 through k_6) used to construct ε_χ.

Other influences on accuracy of atomic positions

Normalized Error Score

The predicted error curve, $\varepsilon_\chi(B,ATOM/REFL)$ contains contributions from atomic B factor and the resolution of the data. To evaluate further atomic attributes which might influence positional accuracy, a normalized error score (N.E.S.) was defined as the standard deviation of a Gaussian fit to the Z-scores of a selected subset of atoms divided by the standard deviation of a Gaussian fit to the Z-scores for all the carbon atoms (which constitute over 64% of the atoms evaluated), e.g.:

$$N.E.S.(subset) = \frac{\sigma(Z-score(subset))}{\sigma(Z-score(carbon))} \qquad (10)$$

Thus, if $\varepsilon_\chi(B)$ correctly predicts the accuracy of a subset of atoms, the normalized error score for that subset of atoms should be close to 1.0. If the $\varepsilon_\chi(B)$ estimation is too large or too small, the N.E.S. will be less than or greater than 1.0, respectively.

A standard deviation for a normalized error score, σ_{NES}(subset), was calculated by first evaluating a separate normalized error score for each of the 18 structures separately, N.E.S.$_i$(subset), and then taking the standard deviation of these 18 values. This standard deviation indicates how consistent a particular normalized error score is over the eighteen structures used in the study.

The N.E.S. should already account for errors associated with B factor and resolution. As shown in Figure 8a, there is no variation in N.E.S. for atoms of different B factors, within the error given by σ_{NES}, which confirms that ε_χ(B) has accounted for variations due to B factors, even in the bins below 14 Å^2 and above 31 Å^2, which include atoms which were not used in constructing ε_χ(B).

No correlation with atomic number

Figure 8b shows that there is also little or no difference in N.E.S. due to atomic number. That is, carbons, nitrogens and oxygens are all positioned with equal accuracy on average. It is difficult to draw any conclusions about sulfurs, since there are so few in any given structure (between 5 and 14). This is reflected in the large error bar (σ_{NES}) for sulfur in Figure 8b. The N.E.S. for sulfur however indicates that its accuracy is close to that of the other atoms.

In contrast with the results presented here, however, the Cruickshank formula predicts that errors in position are inversely related to the number of electrons in the given atom type. This is because the curvature used in the Cruickshank equation will be greater for an atom with more electrons at a given B factor. The apparent lack of a dependence on atomic number here is probably due to the restraints and constraints applied in macromolecular

crystallography, which ensures that the accuracy of one atom is highly related to the accuracy of its covalently bound neighbors (see below).

Correlation with connectivity

On the other hand, the connectivity of an atom is strongly correlated with accuracy of its position (Figure 8c). Atoms of the mainchain (C, N, Cα and O) have lower than expected errors (N.E.S. < 1.0). The mainchain atoms have lower positional errors than side chain atoms with 3 non-hydrogen neighbors, which in turn have lower errors than atoms with 2 non-hydrogen neighbors. Side chain atoms of only 1 non-hydrogen neighbor have the most uncertainty of all, with a N.E.S. 50% greater than that for mainchain atoms. Note that this is after B-factor correlated effects have been accounted for. Thus, on average, a side chain atom with a B factor of 15 $Å^2$ has a 50% greater positional uncertainty than a main chain atom with a B factor of 15 $Å^2$.

The more non-hydrogen neighbors a given atom has, the lower its error, independent of the B factor of the atom. This can be seen as an extension of relationship between observations/parameters and overall accuracy. The positions of neighboring atoms can be seen as additional observations affecting the given atom position. Likewise, the more neighbors, the fewer degrees of freedom, or parameters, are available for positioning the given atom.

Comparison to Luzzati

A widely used measure of the positional accuracy of crystal structures is the Luzzati plot (Luzzati, 1952). However, constructing a Luzzati plot requires access to the original structure factors (Fo's) and the Luzzati method assumes

all atoms have the same B factor. To the extent that atoms have a range of B factors, the Luzzati plot, since it emphasizes the high resolution data, represents the errors of only the atoms with the lowest B factors in the structure.

The Luzzati method produces a single overall value for the accuracy of a structure, $<\Delta\mathcal{R}>$, which is the average atomic displacement from the "true" structure.

To calculate an overall $<\Delta\mathcal{R}>$ for a structure from $\mathcal{E}_\chi(B)$, individual atomic $\Delta\mathcal{R}$s were calculated from the relationship:

$$\Delta\mathcal{R}_{mp} = \mathcal{E}_\chi(B,ATOM/REFL), \text{ where} \qquad (11)$$

$\Delta\mathcal{R}_{mp}$ is the most probable value for $\Delta\mathcal{R}$ based on the atom's B factor and the value of ATOM/REFL for the structure (Appendix IV).

Ten of the structures used in this study had Luzzati values reported for them. For comparison, in Table II, a value for $<\Delta\mathcal{R}>$ has been calculated from $\mathcal{E}_\chi(B)$ for those ten structures using all atoms with B factors less than 40 $Å^2$, by our method (Eq. 8). There is a rough correspondence between the values, with a correlation coefficient of 0.86 (which is statistically significant at the 99.9% level). However, most of this correlation is due to 1HBS, the lowest resolution structure in the study, since if this point is omitted, the remaining 9 structures yield a correlation coefficient of only 0.15.

Since the Luzzati method uses the observed structure factors, it is useful for evaluating the progress of refinement, which $\mathcal{E}_\chi(B)$ is not. However, the Luzzati method can not assign errors to individual atoms, as $\mathcal{E}_\chi(B)$ does. In addition, calculation of $\mathcal{E}_\chi(B)$ for a structure requires only the

number of atoms and the number of reflections used in refinement, which should be provided in any published report of a crystal structure.

Use of $\varepsilon_\chi(B)$

The function $\varepsilon_\chi(B)$ can be used to estimate the errors in refined macromolecular crystal structures. Since pairs of structures were used to derive $\varepsilon_\chi(B)$, the function represents the expected (one-dimensional) differences between two structures. The expected errors in any one structure are then $\frac{\varepsilon_\chi(B)}{\sqrt{2}}$. The expected random differences between two structures will then be:

$$\varepsilon^2_{total} = \frac{\varepsilon^2_{s1} + \varepsilon^2_{s2}}{2}, \text{ where} \qquad (12)$$

ε_{s1} is $\varepsilon_\chi(B)$ for the first structure and ε_{s2} is $\varepsilon_\chi(B)$ for the second structure. ε_{s1} and ε_{s2} will be different if the ratios, ATOM/REFL, are different for the two structures.

The results of such an analysis are presented in Figure 10, for the comparison of two independently solved structures of bovine trypsin (Chambers & Stroud, 1979). The expected errors are in general close to the observed differences between the structures, especially for B factors below 20 Å^2. For B factors above 20 Å^2, the observed differences exceed the predicted errors. This probably indicates that $\varepsilon_\chi(B)$ underestimates the true uncertainty in a crystal structure, since it was derived from pairs of molecules which were solved simultaneously and under identical conditions.

A word about B factors

Two assumptions about B factors in this study are: 1) that B factors are accurate, and 2) that they are refined in a consistent manner by all crystallographers. In the case of macromolecular crystallography, this is clearly debatable. For example, B factors, far more than the positional variables (X, Y and Z) are extremely sensitive to how the observed amplitudes (F_o's) are scaled, and to what resolution range is used in refinement. In addition, while atomic positions are restrained by known stereochemistry and van der Waals interactions, atomic B factors are typically restrained only minimally, for example through a standard deviation linking the B factors of bonded atoms.

However, to the extent that $\varepsilon_\chi(B)$ could be parameterized on B factors, the assumptions are justified. The remaining discrepancy between observed and predicted error levels exhibited in Figure 7, however, could be due to the breakdown of the above assumptions. For example, for 2PFK (Figure 7m), $\sigma_\chi(B)$ falls far below $\varepsilon_\chi(B)$. However, the B factors in 2PFK extend up to 100 $Å^2$ and have a mean of 40$Å^2$, and $\sigma_\chi(B)$ extends far to the right of that displayed in Figure 7m, resembling $\varepsilon_\chi(B)$ with B replaced by B/2.

As another example, the errors in 1HBS (Figure 7r) don't seem to correlate significantly with B factor. However, the B factors in this 3 Å structure show little consistency from molecule to molecule in the asymmetric unit, questioning the validity of refining atomic B factors at this resolution. The recently introduced free R factor test (Brünger, 1992) is useful in determining when atomic B factors can be refined safely.

Conclusion

By analysis of 18 structures with multiple molecules in the asymmetric unit, a function has been derived which reproduces the positional differences observed between equivalent atoms in the chemically identical structures. We believe this function, $\mathcal{E}_\chi(B)$, truly represents the accuracy of a macromolecular structure because: 1) systemic differences (crystal contacts) were removed by using only normally distributed positional differences; 2) $\mathcal{E}_\chi(B)$ generates an overall level of error for a structure similar to that by obtained by the Luzzati method; 3) the predicted errors decrease with increasing resolution.

As shown above, $\mathcal{E}_\chi(B)$ can be used to predict the level of errors in other macromolecular crystal structures and thus can be used in evaluating the reliability of crystallographic coordinates. Because of the empirical nature of this study it is better to use $\mathcal{E}_\chi(B)$ when interpolating, rather than extrapolating. This means that $\mathcal{E}_\chi(B)$ should only be applied to structures between 1.5 Å and 3.0 Å resolution, and preferably only for those atoms with B factors less than 40 $Å^2$. It is clear that atoms with B factors above this will have larger errors, but the exponential form used here may not be appropriate in that range.

Table I. Structures used in analysis

PDB name	name of protein	res. (Å)	R factor %	# mol in asym unit	# of indep reflec-tions	# atoms in asymmetric unit	REFL/ ATOM
1THB	T state of hemoglobin	1.50	19.6	2	87000	4874	17.8
2CCY	Cytochrome C'	1.67	18.8	2	30533	2146	14.2
4CHA	α-Chymotrypsin	1.68	23.4	2	35274	3591	9.8
4DFR	Dihydrofolate reductase	1.70	15.5	2	32554	3041	10.7
2HHB	Deoxy-hemoglobin	1.74	16.0	2	56287	4779	11.8
3CYT	Tuna Cytochrome C (oxidized)	1.80	20.8	2	16831	1743	9.7
2AZA	Azurin	1.80	15.7	2	21980	2263	9.7
1GD1	Glyceraldehyde 3P dehydrogenase	1.80	17.7	4	93120	10984	8.5
1AZA	Azurin	2.00	19.0	2	15614	2133	7.3
1GP1	Glutathione peroxidase	2.00	18.6	2	26564	3102	8.6
1HMQ	Hemerythrin	2.00	17.3	4	40422	4296	9.4
2PKA	Kallikrein A	2.05	22.0	2	35500	3456	10.3
2PFK	Phosphofructo-kinase	2.40	16.8	4	59481	9371	6.3
1FCB	Flavocytochrome B2	2.40	18.8	2	61365	6948	8.8
4MDH	Malate dehydrogenase	2.50	16.7	2	22910	5675	4.0
4ATC	Aspartate trans-carbamylase	2.60	24.0	2	26912	7620	3.5
1FC1	Immunoglobin IGG	2.90	22.0	2	10342	3182	3.3
1HBS	Deoxyhemo-globin S	3.00	25.4	4	17662	9104	1.9

The four letter code refers to the PDB designation for each structure, for which the references are: 1THB, (Waller & Liddington, 1990); 2CCY, (Finzel, Weber, Hardman & Salemme, 1985); 4CHA, (Tsukada & Blow, 1985); 4DFR, (Bolin, Filman, Matthews, Hamlin & Kraut, 1982); 2HHB, (Fermi, Perutz & Shaanan, 1984); 3CYT, (Takano & Dickerson, 1980); 2AZA, (Baker, 1988); 1GD1, (Skarzynski, Moody & Wonacott, 1987); 1AZA, (Norris, Anderson & Baker, 1983); 1GP1, (Epp, Ladenstein & Wender, 1983); 1HMQ, (Strenkamp, Siker &

Jensen, 1982); 2PKA, (Bode, Chen & Bartels, 1983); 2PFK, (Rypniewski & Evans, 1989); 1FCB, (Xia & Mathews, 1990); 4MDH, (Birktoft, Rhodes & Banaszak, 1989); 4ATC, (Ke, Hozatko & Lipscomb, 1984); 1FC1, (Deisenhofer, 1981); 1HBS, (Padlan & Love, 1985)

Table II. Internal control and comparison of $\varepsilon_\chi(B)$ to Luzzati formula

Structure	s.d. of Z-score	$<\Delta R>$ Luzzati	$<\Delta R>$ from ε_x
1THB	1.08	0.25	0.12
2CCY	0.95	0.20	0.14
4CHA	0.83	--	0.12
4DFR	0.91	0.15	0.18
2HHB	1.08	0.18	0.14
3CYT	0.81	0.20	0.16
2AZA	1.04	0.15	0.15
1GD1	0.67	0.18	0.15
1AZA	0.98	--	0.20
1GP1	0.96	--	0.14
1HMQ	0.96	--	0.13
2PKA	1.37	0.20	0.14
2PFK	0.42	--	0.31
1FCB	0.74	--	0.23
4MDH	1.10	0.225	0.31
4ATC	1.17	--	0.36
1FC1	0.74	--	0.40
1HBS	0.90	0.40	0.55

Figure legends:

Figure 1. Scatter plot of the difference in position (ΔR) of an atom in the two molecules in the asymmetric unit of the 1GP1 structure, after superpositioning, as a function of the average atomic B factor assigned to the atom in the two molecules. The vertical bars indicate the atoms with a mean B factor of 12 $Å^2$ +/- 2 $Å^2$.

Figure 2. Histogram of the distribution of ΔR values for atoms in the 1GP1 structure with a mean atom B factor of 12$Å^2$ +/- 2 $Å^2$. The bin size is 1/17 the rms value of the ΔRs.

Figure 3. Histogram of $\Delta\chi$ values obtained from the ΔR values in Figure 2. The dashed line indicates the best fit Gaussian curve, where the standard deviation is $\sigma\chi(12.0)$.

Figure 4. Differences between the two molecules in the asymmetric unit of the 1GP1 structure as a function of B factor. The thin curve represents 341 values for $\sigma\chi(B)$ (from B=6.0 $Å^2$ to 40.0 $Å^2$ in steps of 0.1 $Å^2$). The thick curve is the best fit 3 parameter exponential function to these data points.

Figure 5. The values of $\sigma\chi(B)$ (from the smooth curve approximation) for the 18 structures in the study at 3 distinct B factors plotted as a function of exp(-2ATOM/REFL) for each structure. Circles, B factor = 10 $Å^2$; Squares, B factor = 20 $Å^2$; Triangles, B factor = 30 $Å^2$. Also indicated is the best fit line to the data points at each B factor.

Figure 6. Family of $\mathcal{E}_\chi(B)$ curves. Each value of ATOM/REFL yields a distinct member of this family. The curves shown span the range of values seen in this study. From top to bottom, the values of ATOM/REFL are 1/2, 1/4, 1/8 and 1/16 respectively.

Figure 7. Observed errors, $\sigma_\chi(B)$, and expected errors, $\mathcal{E}_\chi(B)$, for each of the 18 structures used in the study. Since the $\sigma_\chi(B)$ curves were used to derive the empirical formula, the degree to which the two curves in each figure match indicates how well all the information from all the curves has been reduced to six parameters (k1 through k6). The structures are displayed in order of resolution of the structure. In each figure, the choppy line is the standard deviation of $\Delta\chi$ in each B factor bin, the thick solid line is the 3 parameter $\sigma_\chi(B)$ curve, and the dashed line is the curve calculated for that structure from $\mathcal{E}_\chi$(B,ATOM/REFL). Note that ordinate of four figures (1MDH, 4ATC, 1FC1 and 1HBS) goes to 2.4 Å, while the ordinate of the other 14 goes to 0.6 Å.

Figure 8. The Normalized Error Score (N.E.S.) indicates how well $\mathcal{E}_\chi(B)$ accounts for different levels of errors in different subgroups of atoms. A N.E.S. value below 1.0 means $\mathcal{E}_\chi(B)$ overestimates the errors in that subgroup, while a value above 1.0 means $\mathcal{E}_\chi(B)$ underestimates the errors for that subgroup.

Figure 9. Application of the $\mathcal{E}_\chi(B)$ curve. The smooth curve represents the predicted one-dimensional standard deviation of the positional differences between two independently solved trypsin structures. The jagged curve shows the observed one-dimensional standard deviation of the positional differences.

Appendix I

Proof that $P(X) = \frac{1}{R}$ for $0<X<R$. The exact form used here assumes only positive values of X.

This result says that all values of the X component from a random 3 dimensional vector of length R are equally likely. Since this result is rather counter-intuitive, the derivation is presented below.

The probability of a given event X, P(X), is defined as the number of outcomes with a value between X and X+δx divided by the total number of outcomes.

Consider a sphere of radius R. P(X)dx is then that surface area of the sphere generated with an X component between X and X+dx, divided by the total surface area of the hemisphere (since we're only interested in positive values of X).

The desired surface area can be calculated from the following integral:

$$\int_{\theta=\cos^{-1}\frac{X}{R}}^{\theta=\cos^{-1}\frac{X+dX}{R}} \int_{\varphi=0}^{\varphi=2\pi} R^2 \sin\theta \, d\theta \, d\varphi, \qquad \text{where} \tag{I.1}$$

θ is the angle between the R vector and the x axis, ϕ is the angle of between the projection of R into the y-z plane and the z axis, and $R^2 \sin\theta d\theta d\phi$ is the surface area element in spherical coordinates.

Evaluating this double integral yields the value $2\pi R dX$. The surface area of a hemisphere is $2\pi R^2$, so that

$$P(X)dX = \frac{2\pi R dX}{2\pi R^2} = \frac{dX}{R}, \text{ or, equivalently} \qquad \text{(I.2)}$$

$$P(X) = \frac{1}{R} \qquad \text{(I.3)}$$

Appendix II

The ratio ATOM/REFL is related to the maximum resolution and the protein fraction in the unit cell. The following symbols are used:

ρ = density of protein = 1.3 Å^3/dalton

A = average molecular mass per non-hydrogen atom = 14 daltons

V = volume of unit cell

a = unit cell length, assuming cubic lattice

ATOM = non-hydrogen atoms per asymmetric unit

m = asymmetric units per unit cell

F_p = protein fraction in the unit cell

a* = reciprocal unit cell length = 1/a

d = maximum resolution

d* = maximum $|s|$ = 1/d

obs = number of observations

REFL = number of independent reflections

The number of possible observations is related to the volume of the sphere in reciprocal space.

$$obs = \frac{4}{3}\pi\left(\frac{d^*}{a^*}\right)^3 \qquad \text{(II.1)}$$

To get the number of unique reflections, divide by 2m. Also, we can replace d* and a* by 1/d and 1/a respectively to get:

$$REFL = \frac{4}{3}\pi\left(\frac{a}{d}\right)^3 / 2m \qquad \text{(II.2)}$$

The volume of the unit cell is given by:

$$V = (m\rho A)\ ATOM/F_p \qquad \text{(II.3)}$$

Assuming a cubic lattice, V = a^3. Thus, we can rewrite the ratio

$$ATOM/REFL = (ATOM)\ (2m)\ \frac{3}{4\pi}\left(\frac{d}{a}\right)^3 \qquad \text{(II.4)}$$

as

$$ATOM/REFL = \frac{(ATOM)(2m)3d^3F_p}{4\pi\rho(ATOM)m} = \frac{3}{2\pi\rho A}d^3F_p \qquad \text{(II.5)}$$

Taking ρ = 1.3 $Å^3$/dalton and A = 14 daltons, this simplifies to:

$$ATOM/REFL = \frac{F_p}{38Å^3}d^3 \qquad \text{(II.6)}$$

Taking an average value of 0.5 for F_p (50% solvent content) we can write:

$$ATOM/REFL = \frac{d^3}{76Å^3} \qquad \text{(II.7)}$$

Appendix III

Constructing an exponential curve expression for $\mathcal{E}_\chi$(B).

$\mathcal{E}_\chi$(B, ATOM/REFL) is defined by the following equation:

$$\mathcal{E}_\chi(B, ATOM/REFL) = Intercept(B) + Slope(B) * e^{(-2*ATOM/REFL)} \qquad (7)$$

where Intercept(B) and Slope(B) are defined by equations 4 and 5.

At one specific value of ATOM/REFL, $\mathcal{E}_\chi$(B) is an exponential function of the B factor, as are Intercept(B) and Slope(B). This dependence can be made explicit by recasting $\mathcal{E}_\chi$(B) as:

$$\mathcal{E}_\chi(B) = p1 + p2 * e^{(B/p3)} \qquad (8)$$

Any three points can be fit exactly by a three-parameter exponential. Thus, the values of $\mathcal{E}_\chi$(B) for B=10 Å^2, 20 Å^2 and 30 Å^2 calculated from Eq. 6 uniquely determine the three parameters, p1, p2 and p3, for any given value of ATOM/REFL. Namely:

$$p3 = 10\frac{\mathcal{E}_\chi(20, ATOM/REFL) - \mathcal{E}_\chi(10, ATOM/REFL)}{\mathcal{E}_\chi(30, ATOM/REFL) - \mathcal{E}_\chi(20, ATOM/REFL)} \qquad (III.1)$$

$$p2 = \frac{\mathcal{E}_\chi(20, ATOM/REFL) - \mathcal{E}_\chi(10, ATOM/REFL)}{e^{(20/p3)} - e^{(10/p3)}} \qquad (III.2)$$

$$p1 = \mathcal{E}_\chi(20, ATOM/REFL) - p2 * e^{(20/p3)} \qquad (III.3)$$

Appendix IV

The relationship between $\Delta\mathcal{R}_{mp}$ and ε_χ.

The following symbols are used:

$\Delta\mathcal{R}$ = the distance between the observed position for an atom and the "true" position for that atom.

$\Delta\mathcal{R}_{mp}$ = the most probable value for $\Delta\mathcal{R}$ given a Maxwellian distribution.

ε_χ = the expected one-dimensional standard deviation of the differences in atomic positions between two observations of the same "true" structure.

ε_r = the expected three-dimensional standard deviation of the differences in atomic positions between two observations of the same "true" structure.

$\sigma_{\mathcal{R}}$ = the three-dimensional standard deviation of the differences in atomic position between the "true" structure and an observed structure.

By propagation of error,

$$\varepsilon_r^2 = \varepsilon_\chi^2 + \varepsilon_\chi^2 + \varepsilon_\chi^2 \qquad \text{(IV.1)}$$

therefore

$$\varepsilon_r = \sqrt{3}\,\varepsilon_\chi \qquad \text{(IV.2)}$$

Likewise, by propagation of errors again,

$$\varepsilon_r^2 = \sigma_{\mathcal{R}}^2 + \sigma_{\mathcal{R}}^2, \qquad \text{(IV.3)}$$

so

$$\sigma_{\mathcal{R}} = \sqrt{\tfrac{3}{2}}\,\varepsilon\chi \tag{IV.4}$$

This three-dimensional standard deviation corresponds to the Maxwellian distribution:

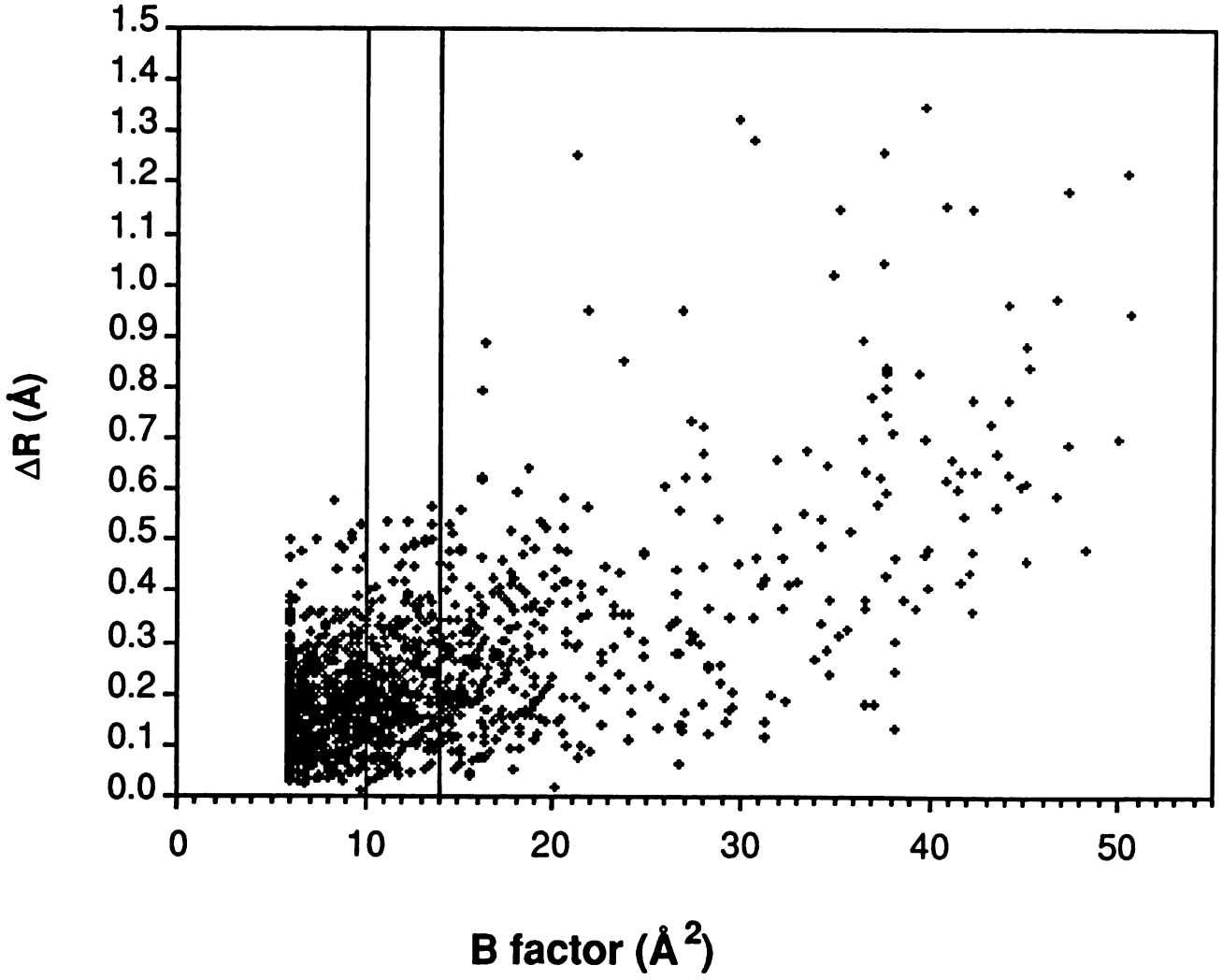
$$P(\Delta\mathcal{R}) = \frac{\Delta\mathcal{R}^2}{\sigma_{\mathcal{R}}^3}\sqrt{\frac{54}{\pi}}\exp\left(-\tfrac{3}{2}\frac{\Delta\mathcal{R}^2}{\sigma_{\mathcal{R}}^2}\right) \tag{IV.5}$$

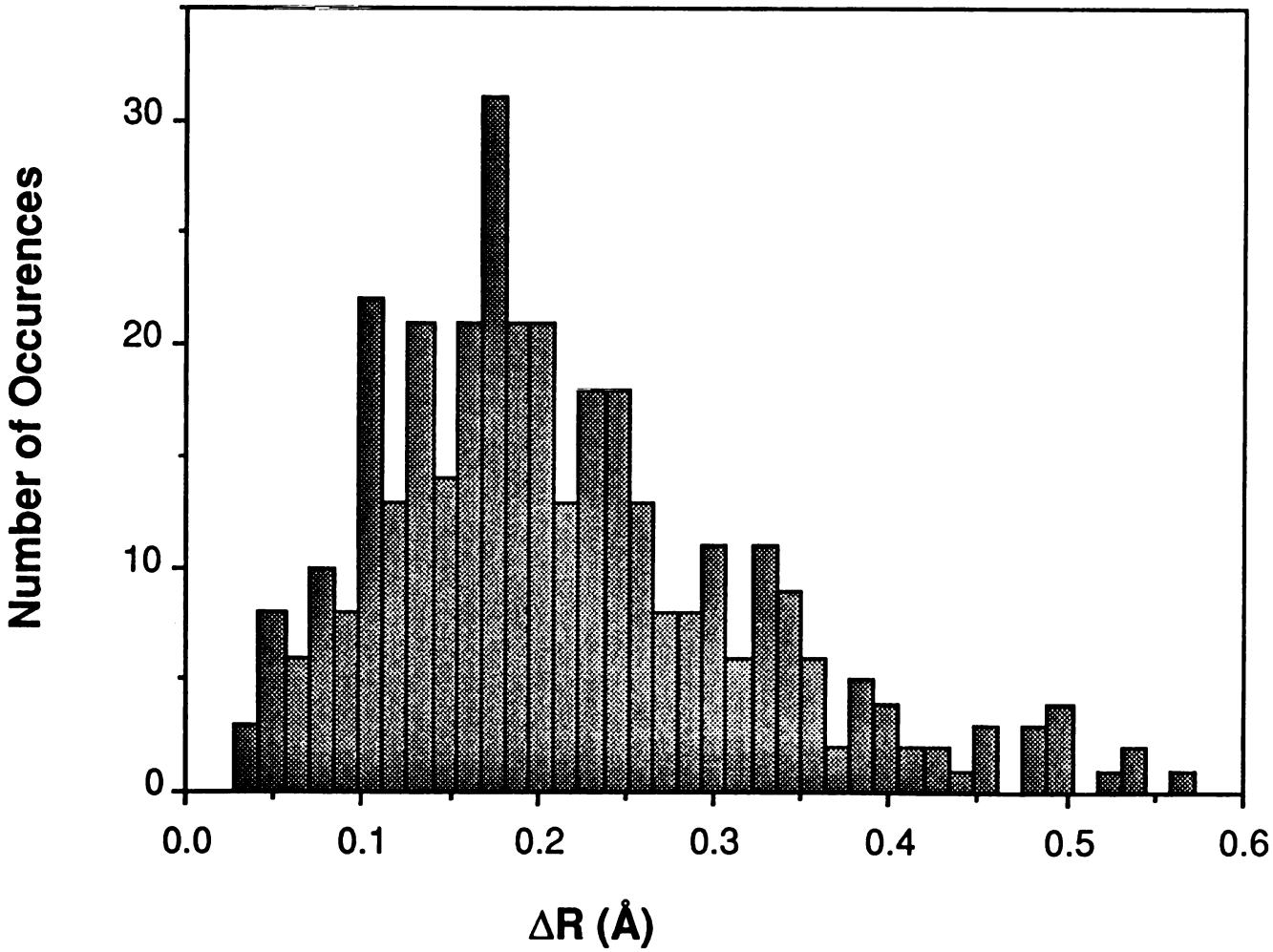
The maximum value of this function yields $\Delta\mathcal{R}_{mp}$, which is

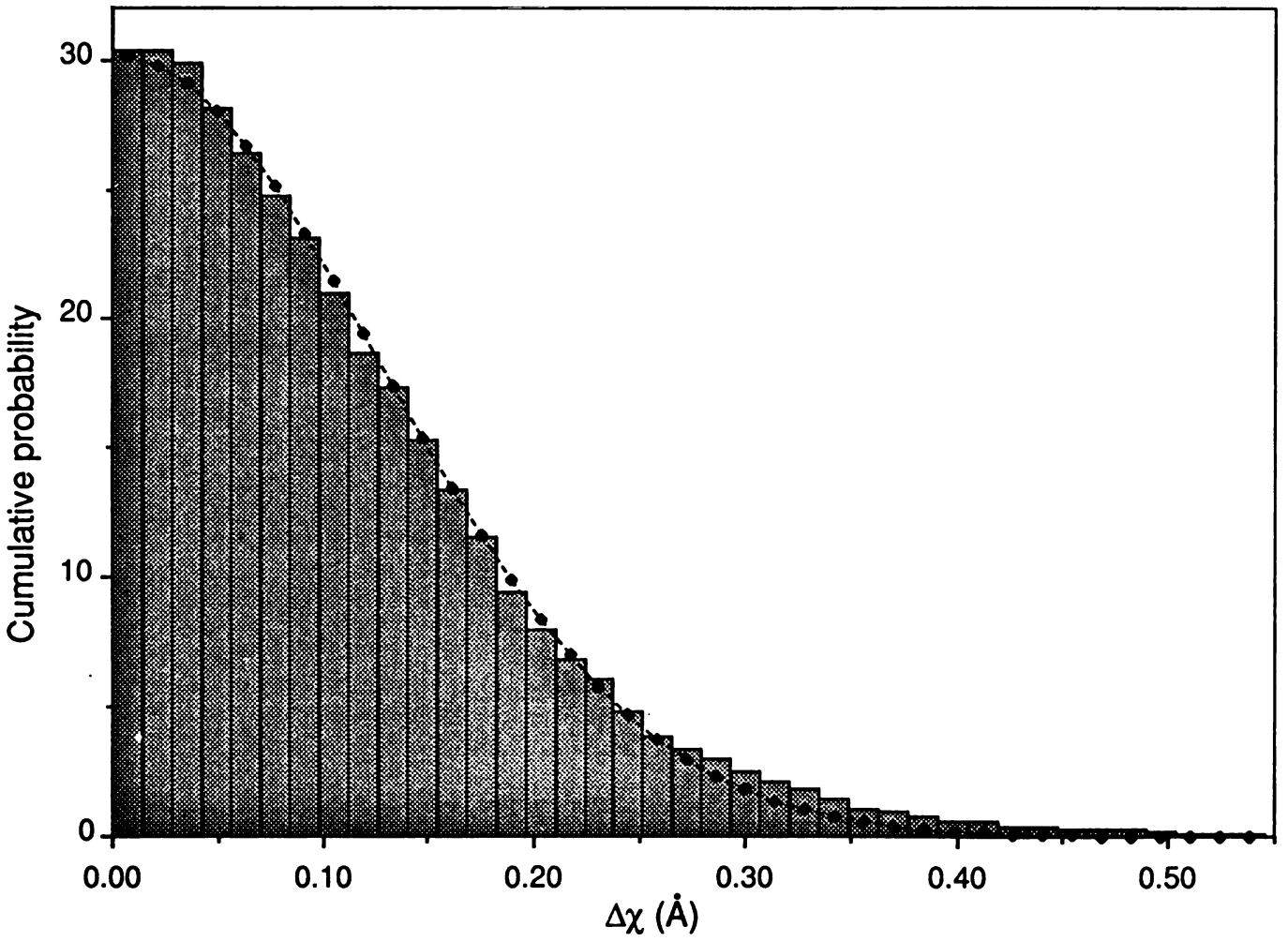
$$\Delta\mathcal{R}_{mp} = \sqrt{\tfrac{2}{3}}\,\sigma_{\mathcal{R}} \tag{IV.6}$$

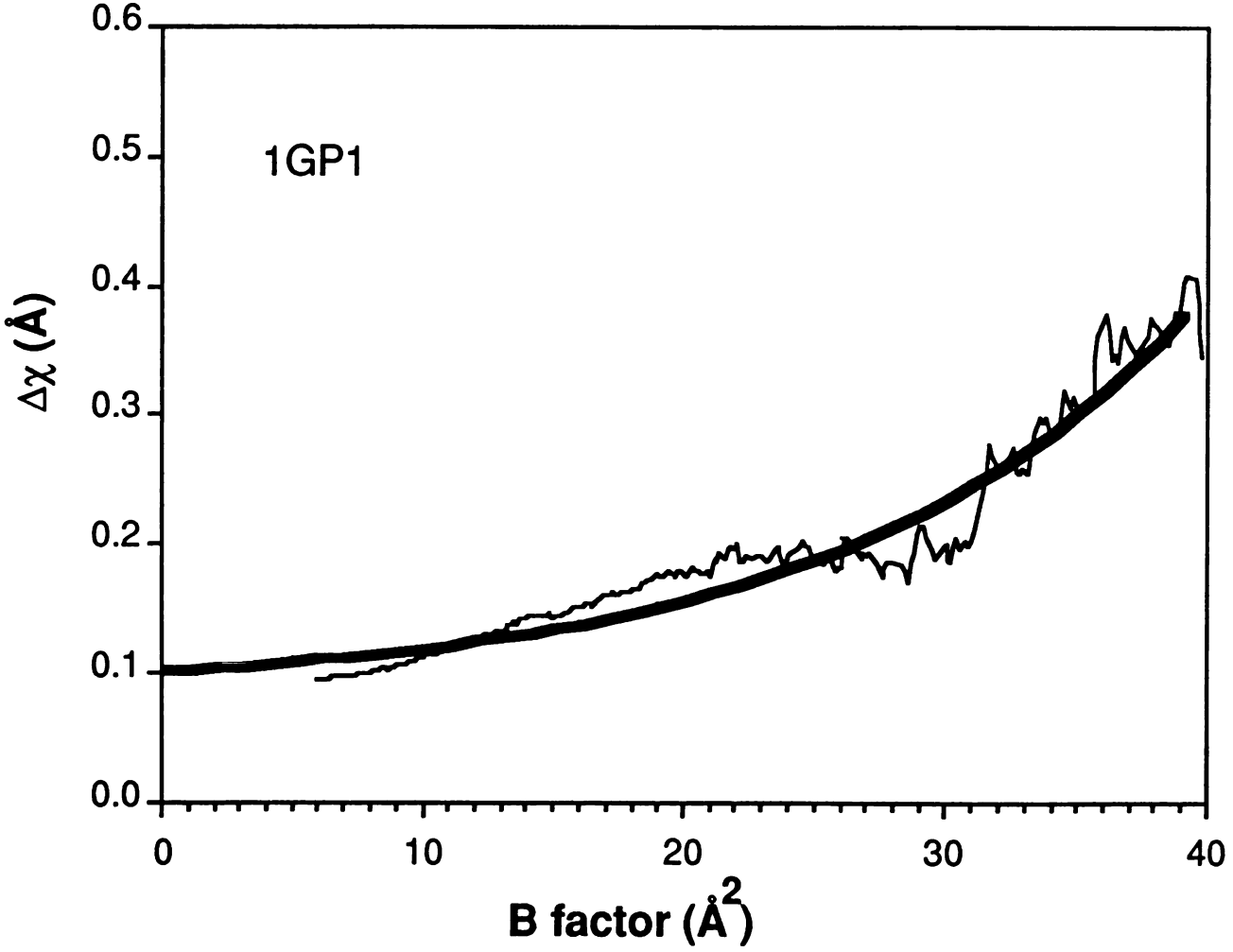
Therefore,

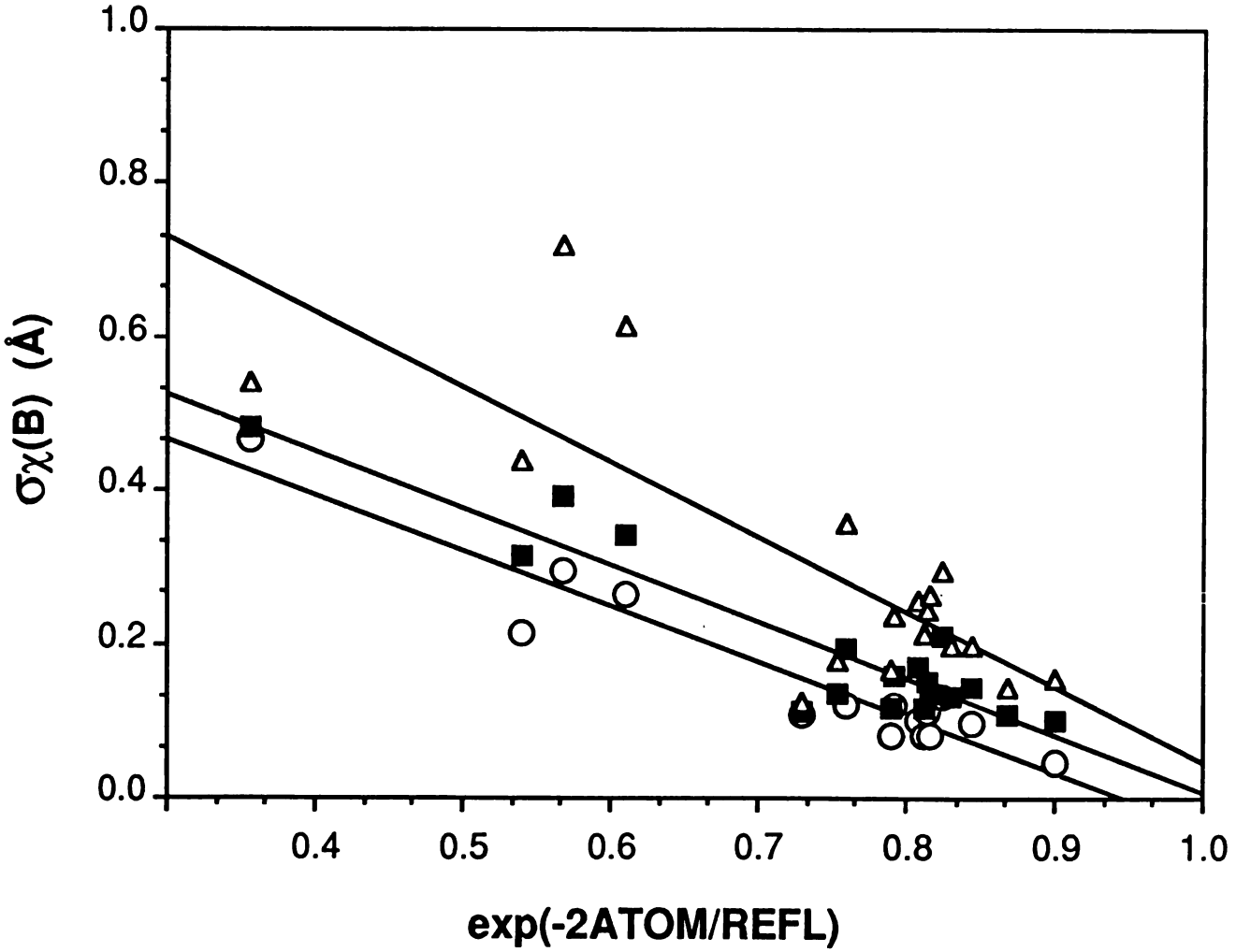
$$\Delta\mathcal{R}_{mp} = \sqrt{\tfrac{2}{3}}\,\sigma_{\mathcal{R}} = \sqrt{\tfrac{2}{3}}\sqrt{\tfrac{3}{2}}\,\varepsilon\chi = \varepsilon\chi \tag{IV.7}$$

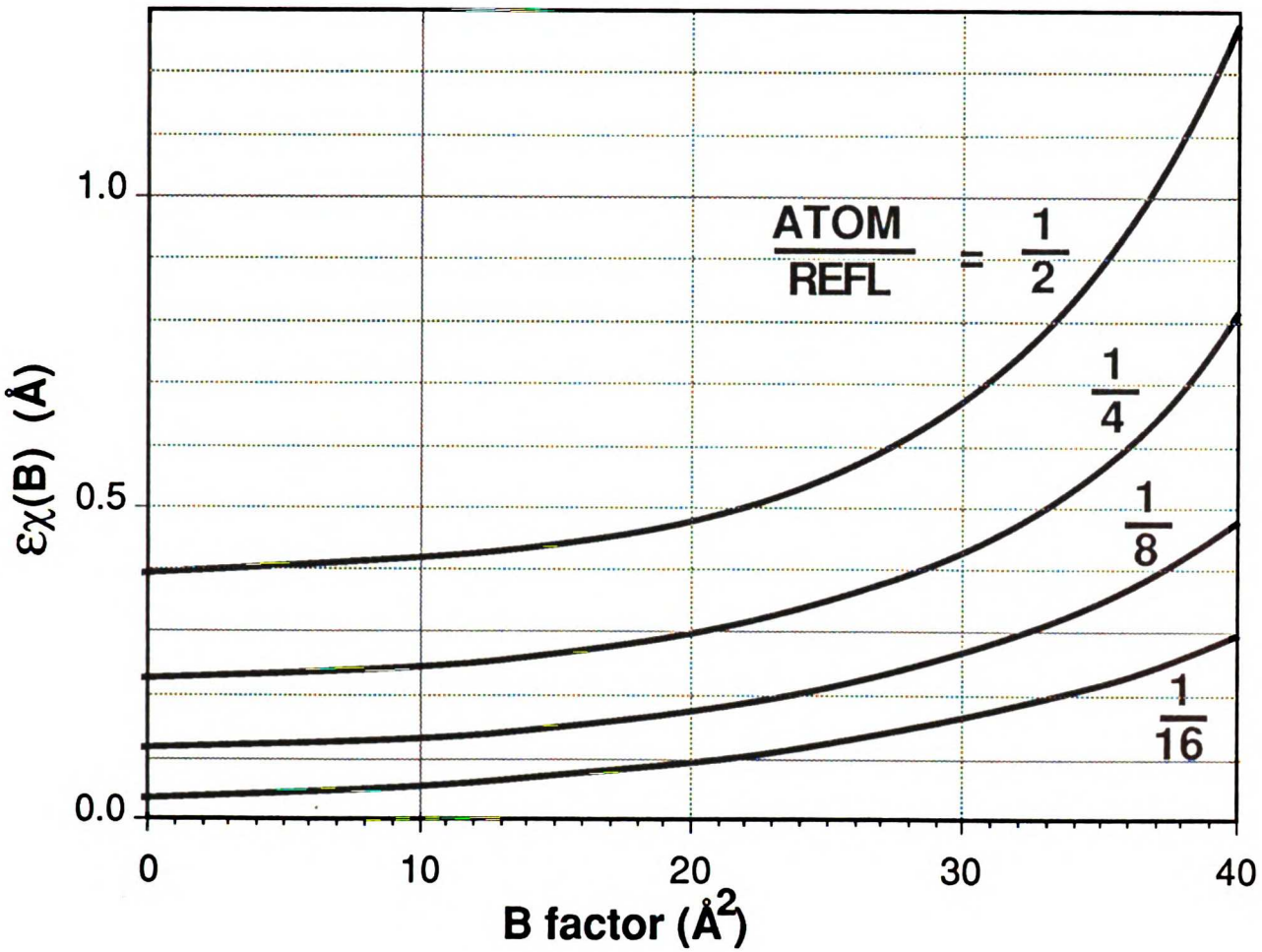

1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
ΔR (Å)
0
10
20
30
40
50
B factor (Å 2)

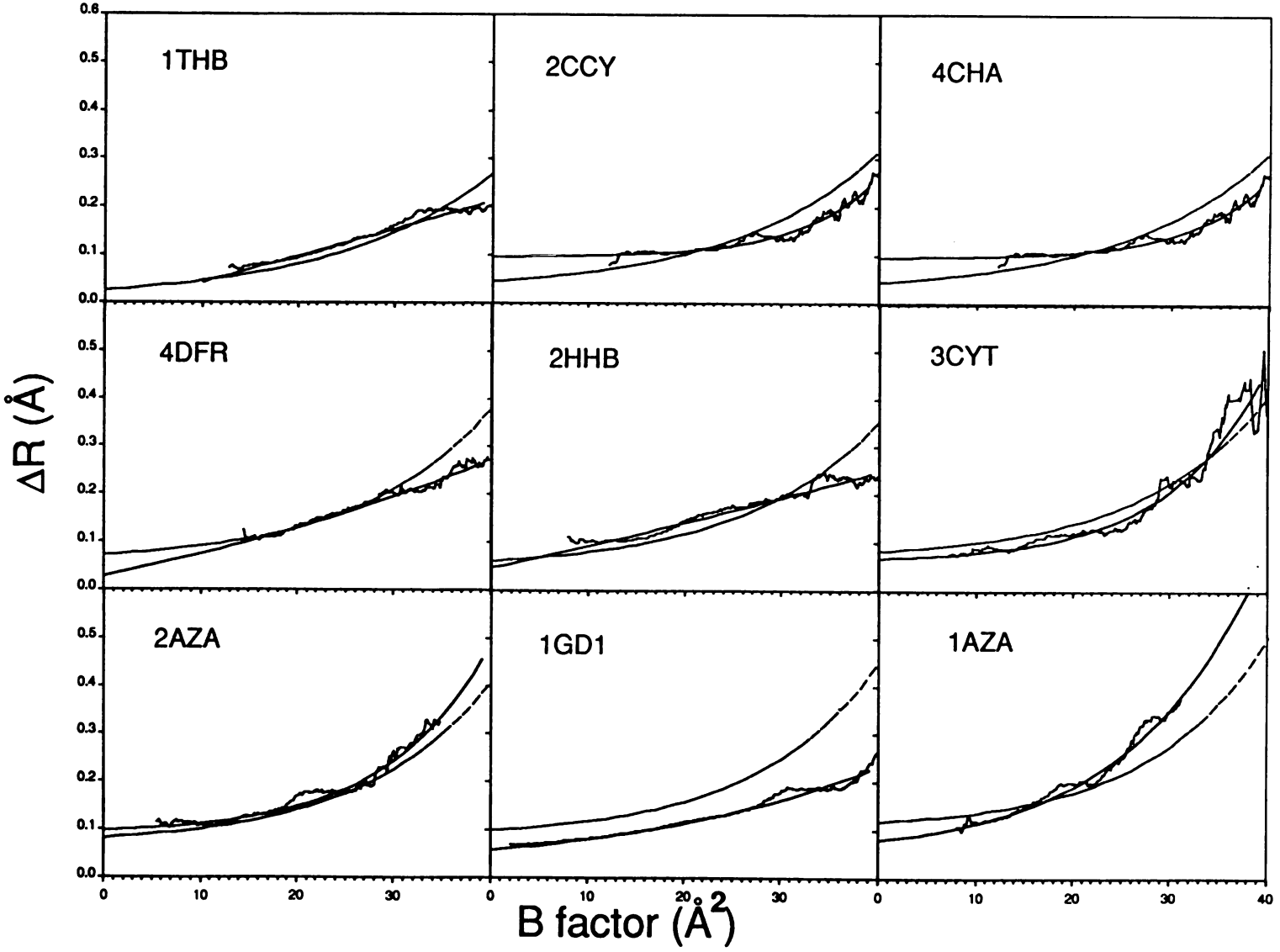

Number of Occurences
30
20
10
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
ΔR (Å)

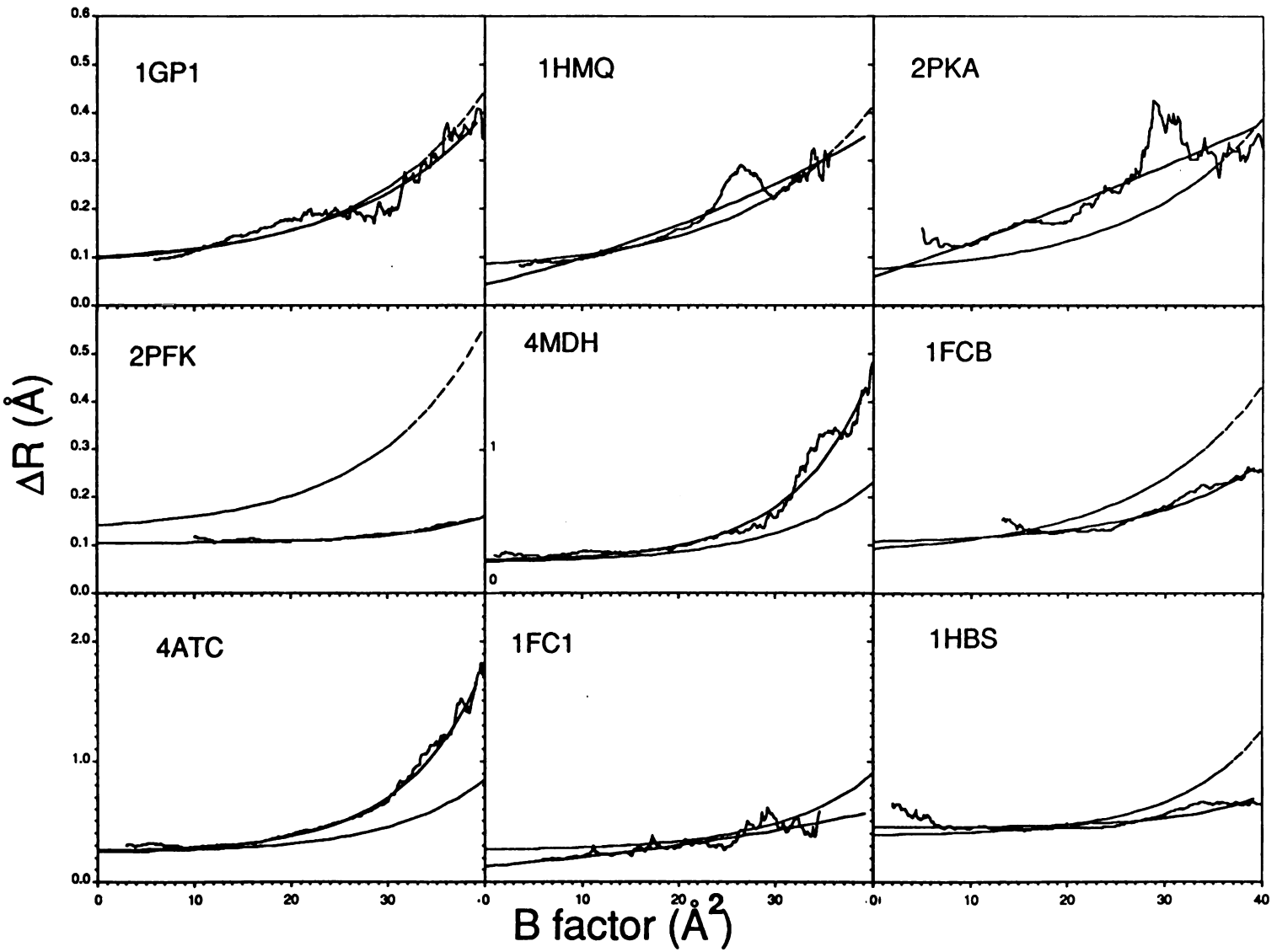

Cumulative probability
30
20
10
0
0.00
0.10
0.20
0.30
0.40
0.50
Δχ (Å)


1GP1
Δχ (Å)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
B factor (Å²)

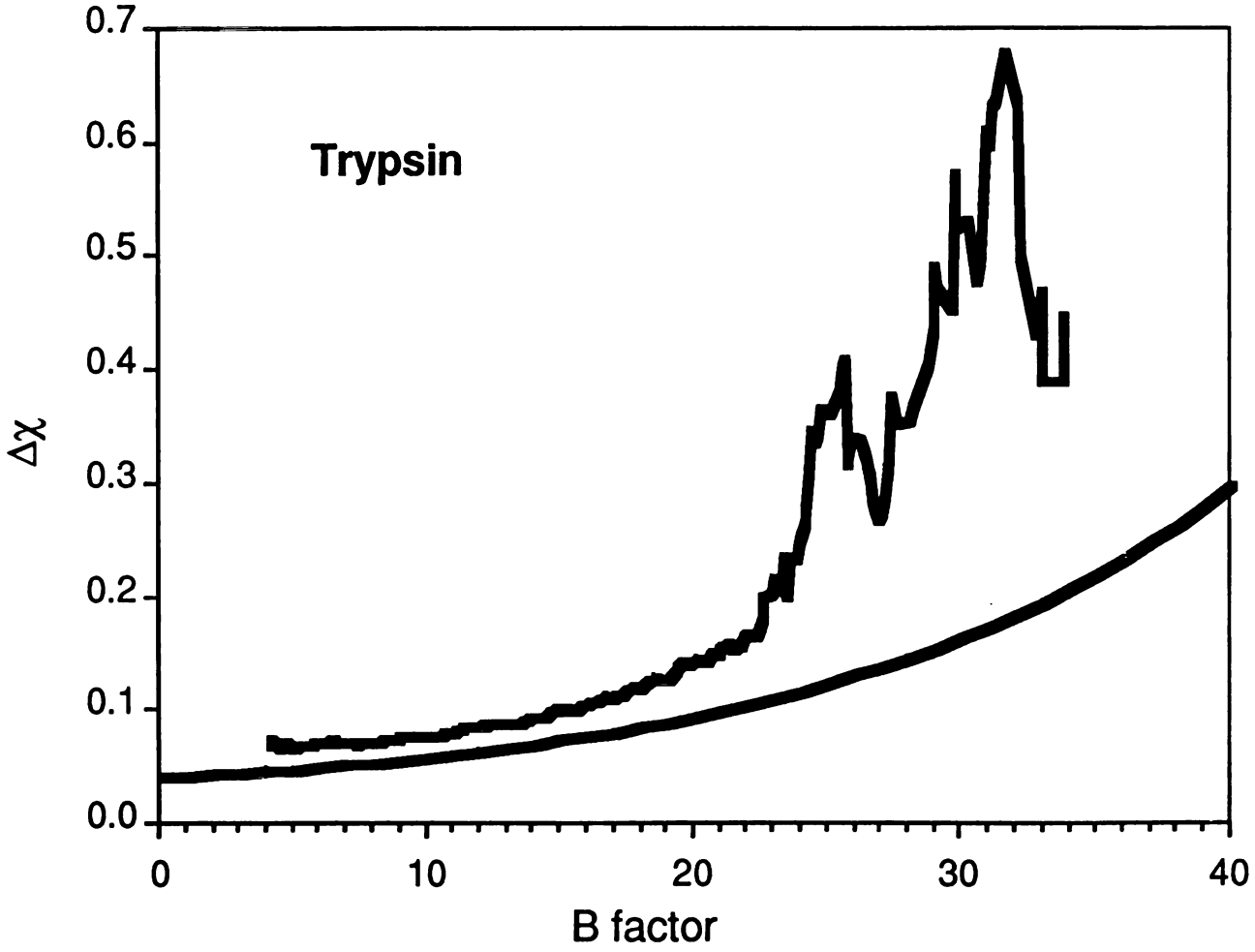

1.0
0.8
0.6
0.4
0.2
0.0
σχ(B) (Å)
0.4
0.5
0.6
0.7
0.8
0.9
1.0
exp(-2ATOM/REFL)


1.0
0.5
0.0
εχ(B) (Å)
0
10
20
30
40
B factor (Å²)
ATOM / REFL = 1/2
1/4
1/8
1/16


1THB
2CCY
4CHA
4DFR
2HHB
3CYT
2AZA
1GD1
1AZA
ΔR (Å)
B factor (Å²)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
10
20
30
40


1GP1
1HMQ
2PKA
2PFK
4MDH
1FCB
4ATC
1FC1
1HBS
ΔR (Å)
B factor (Å²)

Normalized Error Score
0
1
2
B factor
Element
Connectivity
B=5
B=10
B=15
B=20
B=25
B=30
B=35
S
O
N
C
main chain
sc 3 bonds
sc 2 bonds
sc 1 bond
subset


Trypsin
Δχ
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
10
20
30
40
B factor

Chapter 2.

Methods of Crystal Structure Comparison

A. Superpositioning

Before you can compare two structures, you must first decide how you are going to orient the two structures with respect to each other. Every crystal structure is in its own reference frame with essentially arbitrary x, y and z axes.

The standard way to superimpose two sets of coordinates is to pick some common set of atoms in the two structures, and rotate and translate one of the structures so as to minimize the rms deviation of the pairs of selected atoms. This is described in Perry, et al., 1990, which is discussed in the next section. Incidentally, minimizing the rms deviation to superimpose two structures implicitly assumes that all differences between the pairs of selected atoms are due to random, Gaussian distributed, errors. Then, the superpositioning which minimizes the rms deviation produces the "most probable" solution.

This begs the question of how one selects the set of atoms to use for superpositioning. The easiest choices are all the atoms, all the mainchain atoms, or all the alpha carbons. But what if there are some real differences between the structures, or some conformational change? My favorite illustration is to compare two left hands, one thumb up, hitchhiker style, and one with the thumb down, against the index finger. If you found these two hands in isolation, how would you orient them relative to each other so as to describe the conformational change?

This introduces the concept of a core, a subset of the structure which does not change. Changes in structure can then be described in reference to this core. My program, NEWDOME, was developed to select a core of alpha carbon atoms in a pair of structures. The programming details are left to the

appendix, but what NEWDOME does is to select the largest subset of alpha carbons in the structures whose positions relative to the other atoms in the core do not change. Furthermore, to fit our conceptual image of a core as a connected group of atoms, every atom in the core is within a certain distance from every other atom in the core. Indeed, the atoms are connected in the mathematical sense in that there is a path between any pair of atoms in the core constructed of other atoms in the core, with no two atoms in the path farther than the maximum cutoff distance. In practice, I have found maximum cutoff distance of 10 Å to be generally acceptable. I rationalize this value as the farthest apart two alpha carbons could be and still have their sidechains interacting.

The movements of atoms relative to each other can be described in a difference distance plot. Such a plot can be seen in Earnest, et al., 1990, described in the next section. In practice, NEWDOME sets an upper limit of 0.5 Å for change of interatomic distance between all pairs of atoms in the selected core (although this is a user definable parameter).

In the hand analogy then, NEWDOME would select the palm and the fingers as the basis for overlap, which would then assign all the motion to the thumb. This seems the natural answer, although *a priori* we can not be sure that in fact the thumbs have not moved and it is the rest of the hand which has changed. Also note that in the hand analogy that we could have used the entire hand as the basis for our overlap, but this would "smear" the observed motion over both the thumb and the fingers; the result would not have been as clean. This is essentially what NEWDOME does; it helps "focus" the motion by assigning all the motion to those pieces of the structure which move relative to the core.

All of the structure comparisons described in this thesis began by superimposing the two structures by NEWDOME.

B. Conformational Changes

Conformational changes in proteins can occur on a number of different levels or scales. These different sorts of molecular changes are presented in the following subsections, and methods appropriate to each are described.

1. large scale rigid body motion

The core, defined above, is conceived of as a rigid body; a subset of the structure which is unchanging. A domain may be defined in an analogous manner to the core as a subset of atoms which don't move relative to each other. If the entire protein can be described as a set of a few large subsets or domains, a rigid body description of conformational changes may be most useful.

Once the pair of structures to be compared has been superimposed, the motion of this domain can then be described as a rigid body transformation. Any rigid body transformation can be thought of as a screw transformation; that is, a translation and a rotation about a single axis. The SUPER function in GEM, described in the appendix, calculates this transformation.

Alternatively, an internal coordinate system for a domain can be defined from the principal axes of the set of atoms. Every object has three orthogonal principal axes which are axes about which the object can be rotated without any external torque. A rigid body transformation can then be described as a translation of the origin of the domain coordinate system and rotations about the principal axes.

A complete description of this method, and its application to the analysis of glycogen phosphorylase, is given in:

Michelle F. Browner, Eric B. Fauman, and Robert J. Fletterick (1992) Tracking Conformational States in Allosteric Transitions of Phosphorylase. Biochemistry *31*, 11297-11304.

Even if the entire protein cannot be split into separate domains, certain parts of the protein may still be adequately described by rigid body motion. For example, in my analysis of our HIV protease structure, described later, I used a principal axes analysis to define the conformational change in the flaps (two beta-ribbon structures covering the active site) and to define the rotation and translation observed for the ligand.

2. motion of secondary structural elements

If there are no obvious domains in the proteins being analyzed, or if the rigid subunits of the proteins and small and numerous, an alternative to the above approach may be in order.

In the case of thymidylate synthase, there were no large domains, but we noticed the different elements of secondary structure, helices, strands and loops, each seemed to move as rigid units, in going from the unbound to the bound form. We picked out the rigid units by first displaying the superimposed pair of crystal structures along with a small arrow on each alpha carbon indicating the direction and magnitude of the shift in position of that atom (program ALLARROW). Stretches of alpha carbons with similar shifts were grouped together, and as noted these generally corresponded to secondary structural elements.

Once the small rigid units were identified, the shifts of the units could be analyzed and compared. For instance, a total shift for a rigid unit was defined as the vector sum of the shifts of the individual alpha carbons. The statistical significance of the magnitude of the vector sum was analyzed by reference to Maxwellian distributions. A comparison of the directions of the shifts was effected by calculating dot products between all pairs of rigid units. By making a matrix of these dot products, one can quickly asses which units are moving in the same direction or which are moving in opposite directions. This analysis of direction and magnitude of the vector sums is performed by the program WHEREARROW.

The results of this type of analysis is given in the next chapter in the section of Segmental Accommodation.

3. generalized all-atom motion

When the shifts in atomic positions are very slight, it may be more useful to forego any type of rigid body analysis. For a comparison of room temperature to low temperature trypsin structures, the primary tool was simply a difference distance matrix, as described above for superpositioning techniques. This shows which alpha carbons have moved closer together or farther apart, on an atom by atom basis.

For a simple way to visualize the motion of all the atoms (not just the alpha carbons), I invented a method which plots the distribution of the angle between the vector shift of each atom to the vector to the center of mass of the protein (program CENTERDOT). In the case of trypsin, this technique clearly shows that the protein has shrunk as a result of the freezing. This same technique was used in my study of the product complex, described in the next chapter.

The most dramatic result in the trypsin study was not the change in atomic positions, but the change in B factors. All atomic B factors were shifted down in a very linear manner as a result of the freezing. The exact relationship was determined by the subroutine BVSB in the program GEM, described in the appendix.

The complete trypsin story can be found in:

Thomas Earnest, Eric Fauman, Charles S. Craik, and Robert Stroud (1991) 1.59 Å Structure of Trypsin at 120 K: Comparison of Low Temperature and Room Temperature Structures. Proteins: Structure, Function and Genetics, 10, 171-187.

C. Species to species comparison

In the preceding section the emphasis was on simple conformational changes; that is, different states for a protein with a fixed amino acid sequence. Things can get somewhat more complicated if there are widespread amino acid differences between the structure under analysis.

As of the writing, the structures of thymidylate synthase from four species are known (*L. casei*, *E. coli*, phage T4, human). The bacterial species were the first to be solved and are currently at the highest resolution (*L. casei* to 2.3 Å and *E. coli* to 1.8 Å).

When we compared the *E. coli* and *L. casei* TSes, the first problem was how to align the sequences. This is different than the problem of superimposing structures, described above, since the method given there assumes you already know which alpha carbons correspond in the two structures. Sequence alignment in TS is not terribly difficult because the

sequence is quite highly conserved, and we now have 20 sequences for comparison.

Once the sequences were aligned and the structures superimposed as described previously, the structures could be compared. Using the technique described for analysis of bound to unbound TS, I quantified the motion of the secondary structural elements in this species to species comparison.

However, there were also more subtle motions evident. In manner similar to that for described for the trypsin comparison above, I looked for individual atomic motions as a function of distance to specific sites in the protein. In this case, the sites were the centroids of amino acid substitutions, which allowed me to evaluate shifts as a function of distance from the nearest amino acid change.

In fact, there were two parameters which can serve as predictors for how much an atom will shift in evolution: the atomic B factor and the distance to the nearest amino acid change. The results of the *L. casei* to *E. coli* TS comparison are discussed in the next chapter in the section on plasticity.

Chapter 3.

Crystallographic Analysis of Thymidylate Synthase

A. Plasticity

Thymidylate synthase is one of the most highly conserved enzymes known. *Lactobacillus casei* is a gram positive bacterium and *Escherichia coli* is a gram negative bacterium, and so are separated by about two billion years of evolution. However, their TS proteins are 44% identical, and the tertiary folds of the proteins are identical.

Using the techniques from the last chapter, we discovered three major aspects in explaining the structural differences between crystal structures of these two TSes. First, in general, the backbones of secondary structural elements move as units. That is, all the alpha carbons in a helix or sheet tend to move in the same direction. Second, the amount an atom can be shifted can be predicted by its B factor. That is, the larger an atom's B factor, the more it is likely to be moved as the result of a sequence change. This effect is larger than would be accounted for by B factor related errors, as discussed in Chapter 1. Also, at least in the case of thymidylate synthase (TS), the greater mobility of atoms at the surface of the protein was completely explained by the atomic B factors.

The final aspect relevant to how much an atom will be shifted during evolution is how close the atom is to the amino acid substitution. In the analysis of the TS, I discovered that the farther an atom is from the nearest amino acid substitution, the less it will be shifted. The curve of shift in position versus the nearest change was fit to a decaying exponential, since this is each to manipulate. On theoretical grounds, one can justify a curve of the form $Y=R^{-2}$. If you imagine that the disruptive effects of a given change at a

distance of R are evenly spread out over the spherical shell, which has an area of $4\pi R^2$, so that any given atom at that distance feels only $\frac{1}{4\pi R^2}$ of the effect.

I put this model to the test in a computer simulation I called "The Dance of the 3 Å Happy Spheres." In this simulation, 200 hard spheres were allowed to move so as to reach an equilibrium state where each sphere was three Å away from its nearest neighbor. Next an extra sphere was added to the center of the group and equilibrium was reestablished. As expected, the amount an atom needed to move in reaching equilibrium the second time was proportional to R^{-2} where R was its distance to the added atom.

The difference between a decaying exponential ($Y=e^{-R}$) and the theoretical curve ($Y=R^{-2}$) can not be observed in the range of distances used in the crystallographic analysis (about 2 Å to 6 Å).

The manner in which the structural effects of mutations are absorbed by those atoms closest to the mutation we termed "plasticity." The word "plastic" comes from the Latin, where it means "that which may be molded." Thus, structural plasticity refers to the ease with which the atoms near a mutation may be molded so as to accommodate the mutation.

The full account this comparison of *E. coli* and *L. casei* TSes is given in:

Kathy M. Perry, Eric B. Fauman, Janet S. Finer-Moore, Gladys F. Maley, Larry Hardy, Frank Maley, and Robert M. Stroud (1990) Plastic Adaptation Toward Mutations in Proteins: Structural Comparison of Thymidylate Synthases. Proteins: Structure, Function and Genetics, 9:315-333.

B. Segmental Accommodation

In contrast to the widespread motion with only local directionality described in the previous section for multiple point mutations, the conformational change accompanying ligand binding in thymidylate synthase consisted of a systematic closing down on the active site. This makes sense, as the point mutations were spread all throughout the protein, whereas the ligands are well-localized to the active site.

The two structures used to analyze ligand binding in TS were that of *E. coli* TS bound to inorganic phosphate and *E. coli* TS bound to its substrate, dUMP, and an anti-folate, CB3717. As described in the previous chapter, in binding to its ligands, the secondary structural elements all moved as rigid units. They all moved inward to close off the active site in a consistent yet independent manner. We termed this motion "segmental accommodation." The protein accommodates the ligands by moving different segments in a concerted motion.

There are two main purposes we assigned to segmental accommodation. The first is to increase the interactions between the protein and the ligand, since if there's no conformation change upon ligand binding, the part of the ligand nearest the entrance to the active site will necessarily be devoid of protein interactions. The second purpose is to isolate the ligands from bulk solvent. We suspect that in some regards, the TS reaction is easy to carry out, and simply requires that the substrates be correctly positioned. The presence of unwanted water molecules can lead to extraneous side reactions.

A full description of the conformational change can be found in:

William R. Montfort, Eric B. Fauman, Kathy M. Perry, and Robert M. Stroud (1990) Segmental Accommodation: A Novel Conformational Change Induced Upon Ligand Binding by Thymidylate Synthase. In Current Research in Protein Chemistry: Techniques, Structure and Function, Villafrance, J. (ed.). Academic Press, Inc. San Diego, 367-382.

The relevance of this conformational change to the TS reaction is described in:

William R. Montfort, Kathy M. Perry, Eric B. Fauman, Janet S. Finer-Moore, Gladys F. Maley, Frank Maley, and Robert M. Stroud (1990) Structure, Multiple Site Binding and Segmental Accommodation in Thymidylate Synthase on Binding dUMP and an Anti-Folate. Biochemistry, 29(30), 6965-6977.

C. Water-Mediated Substrate/Product Discrimination

When I determined the structure of *E. coli* thymidylate synthase bound to the product of the reactions, deoxythymidylate (dTMP) and methylene tetrahydrofolate, I discovered several new aspects of the TS reaction. For one, the closed down conformation, observed for the ligand bound structure discussed in the previous section, is also observed for the product complex. This suggests that there is no conformation change during catalysis and thus that all the conformational changes occur to bind and release reactants.

Another discovery was the extent to which TS uses water molecules to help it perform its function. The improved resolution of the product

complex (1.83 Å vs. 1.97 Å for the previous best structure) allowed localization of many new water molecules in the structure. Many are involved in the binding of the folate, which previously had relatively few binding interactions. One water in particular, however, seems to be intricately involved in the catalysis. This water, which I called water^{C7}, because of its proximity to the C7 position of bound dTMP, seems to perform at least 2 functions: 1) by virtue of its position it disfavors binding of the product, tying part of the energy driving the reaction to energy favoring product release. and 2) waterC7 may be a proton acceptor, providing an answer to a long-standing mystery of the TS reaction.

Thus, our understanding of segmental accommodation is expanded to the realization that although bulk solvent is undesired for the correct completion of the TS reaction, highly ordered bound water molecules are essential to the TS reaction.

The complete story of the product complex structure is given in the following paper, which (as of January 1993) has been submitted for publication in Biochemistry.

Water-mediated Substrate/Product Discrimination: The Product Complex of Thymidylate Synthase at 1.83 Å†‡

† Research was supported by NIH Grants RO1-CA-41323 and GM24485 to R.M.S., by NSF Grant DMB90-03737 to G.F.M. and by Grant CA44355 to F.M. from the National Cancer Institute. E.B.F. is a Howard Hughes predoctoral fellow. E.E.R. is an American Cancer Society postdoctoral fellow.

‡ The coordinates for this structure have been submitted to the Brookhaven Protein Data Bank. The accession code is 1TYS.

Eric B. Fauman, Earl E. Rutenber,
Gladys F. Maley,§ Frank Maley,§
Robert M. Stroud*
Department of Biochemistry and Biophysics
University of California, San Francisco
San Francisco, California 94143-0448

§ Present address: Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, NY 12201-0509.

* To whom correspondence should be addressed.

RUNNING TITLE: Water-mediated substrate/product discrimination

KEY WORDS: crystallography, hydroxymethylase, carbamate, dihydrofolate

Textual Footnotes:

1 Abbreviations: TS, Thymidylate Synthase; ECTS, *E. coli* TS; dUMP, 2'-deoxyuridine-5'-monophosphate; dTMP, 2'-deoxythymidine-5'-monophosphate; H_2folate, dihydrofolate; CH_2-H_4folate, 5,10-methylenetetrahydrofolate; CB3717, 10-propargyl-5,8-dideazafolate; Å, Ångstrom = 10^{-10} m; rmsd, root mean square deviation.

2 Numbering of residues is by the *L. casei* convention, as in Hardy et al., 1987. *E. coli* equivalents are given in parentheses throughout the text.

3 $$Rsym = \sqrt{\frac{\sum_{hkl}\sum_{i=1}^{N}(I_{avg} - I_i)^2}{\sum_{hkl}\sum_{i=1}^{N}(I_i)^2}}$$

4 $$R = \frac{\sum_{hkl}\left|(|Fo| - |Fc|)\right|}{\sum_{hkl}|Fo|}$$

ABSTRACT: In an irreversible enzyme-catalyzed reaction, strong binding of the products would lead to substantial product inhibition. The X-ray crystal structure of thymidylate synthase (1.83 Å resolution, final R factor = 0.183 for all data between 7.0 Å and 1.83 Å) in complex with the reaction products displays how the enzyme uses a bound water molecule to disfavor binding of the product nucleotide. This water molecule is hydrogen bonded to absolutely conserved Tyr 146 (using the *Lactobacillus casei* numbering system), and is displaced by the C7 methyl group of thymidylate. The relation between this observation and kinetic and thermodynamic values is discussed. This high-resolution structure reveals a previously unobserved modified N-terminus, identified as carbamate, and new rotamer assignments for a large fraction of the sidechains.

The ternary product structure is compared to the previously determined structure of thymidylate synthase in complex with substrate and an inhibitory cofactor analog. The nearly identical arrangement of ligands in these two structures supports our model for the reaction progress and verifies the physiological relevance of the mode in which potent inhibitors bind to this target for rational drug design.

Thymidylate synthase (TS[2]) (EC 2.1.1.45) catalyzes the reductive methylation of deoxyuridylate (dUMP) to generate thymidylate (dTMP) in the sole *de novo* pathway for this DNA precursor (Figure 1). TS is a dimeric protein of identical monomers which uses methylenetetrahydrofolate (CH_2-H_4folate) both as the one-carbon source and as the reductant, converting it to dihydrofolate (H_2folate). Because of the need for thymidylate in rapidly dividing cells, TS has been a focus of attention for inhibitor design efforts against cancer.

[Fig. 1]

We have studied TS to develop a method for the rational design and improvement of potent inhibitors, and also to understand on an atomic level the kinetics and catalysis of this enzyme (Finer-Moore, et al., 1990). Many X-ray crystal structures of TS have been determined (Hardy, et al., 1987; Matthews, et al., 1990a,b; Montfort, et al., 1990b; Perry, et al., 1990; Schiffer, et al., 1991; Kamb, et al., 1992) with several representing key points along the reaction pathway. TS undergoes a large conformational change in going from the unbound form to the ternary complex with dUMP and CB3717 (Figure 1), a folate analog and potent TS inhibitor (Montfort, et al., 1990a). Various secondary structural elements move toward the active site to sequester the reactants away from bulk solvent in a motion we termed segmental accommodation. Other crystal structures reveal TS bound with dUMP alone is in the open form

---

[2] Abbreviations: TS, Thymidylate Synthase; ECTS, *E. coli* TS; dUMP, 2'-deoxy-5'-uridine monophosphate; dTMP, 2'-deoxy-5'-thymidine monophosphate; H_2folate, dihydrofolate; CH_2-H_4folate, 5,10-methylenetetrahydrofolate; CB3717, 10-propargyl-5,8-dideazafolate; Å, Ångstrom = 10^{-10} m; rmsd, root mean square deviation.

(unpublished results), while TS with a cofactor analog alone is in the closed form (unpublished results). Thus it is the folate which leads to closure of the enzyme.

When the ECTS•dUMP•CB3717 complex was determined it was noted that under non-reducing conditions the quinazoline portion of CB3717 occupied an alternate, well-defined binding pocket. The high degree of conservation of the residues in this alternate site suggested a functional role for this pocket, such as initial binding of the CH_2-H_4folate, or post-catalytic binding of H_2folate.

The nucleotide product of the reaction, dTMP, differs from the substrate, dUMP, by a single methyl group. However, dTMP binds 3-7 fold less tightly (Galivan, et al., 1976; Beaudette, et al., 1980; Santi & Danenberg, 1984), which is important for product release. Because of the hydride transfer step, the reaction is essentially irreversible. Thus, any degree of product binding contributes to product inhibition.

In order to understand the molecular basis of the nucleotide discrimination, to discern the binding mode of H_2folate and to learn the conformational state of the enzyme upon completion of the chemistry of bond rearrangements, we determined the structure of the enzyme bound to the reaction products, H_2folate and dTMP. This ECTS•dTMP•H_2folate complex is the first X-ray crystal structure to reveal the state of the enzyme after bond rearrangements.

Since some of the structures we need to complete our study of structures along the reaction pathway involve the substrate molecules, we have used a mutant of TS with the active site cysteine replaced by a serine to prevent turnover. To simplify

comparisons among these structures, this same mutant has been used for all crystal structures of complexes of TS with the naturally occurring ligands. The *E. coli* C198(146)S[3] mutant TS activity is about 1000 fold less than wild type by spectrophotometric analysis (unpublished results) and about 5000 fold less by tritium release (Dev, et al., 1988). The analogous mutant in *L. casei* TS fails to complement a TS deficient strain of *E. coli* (Climie, et al., 1990). Although k_{cat} is greatly affected, the dissociation constant for dUMP is relatively unchanged (Dev, et al., 1988).

MATERIALS AND METHODS

Protein and Crystals. The thyA gene in M13mp9 was mutated to C198(146)S using the method of Taylor, et al. (1985) and then transferred as a HindIII fragment to pUC19 which was transformed into *E. coli* X2913 (ΔthyA572). After induction to 5 to 10% of the cellular protein, the mutated enzyme was purified as described (Maley & Maley, 1988). Ternary complex crystals were grown by hanging drop cocrystallization of TS C198(146)S with the reaction products dTMP and H_2folate. A 5.0 μL drop of protein solution (20 mM KPO_4 pH 8.0, 1.5 mg/mL TS C198(146)S, 1.9 mM H_2folate, 5.8 mM dTMP, 3.8 mM $MgCl_2$, 3.8 mM dithiothriotol (DTT), 0.05 mM disodium ethylenediaminetetraacetic acid (EDTA) and 1.25 M $(NH_4)_2SO_4$ was equilibrated at 23° C against an excess of precipitant solution (2.5 M

---

[3] Numbering of residues is by the *L. casei* convention, as in Hardy et al., 1987. *E. coli* equivalents are given in parentheses throughout the text.

$(NH_4)_2SO_4$ and 20 mM KPO_4 pH 8.0). Golden, highly birefringent crystals with hexagonal bipyramidal morphology (450 μm x 250 μm x 250 μm) grew in 2-3 days. Diffraction intensities were collected at the Stanford Synchrotron Radiation Laboratory on Port 7-1 using a MAR imaging plate, 1.09 Å radiation, 2° frames and data collection times of 15 to 30 seconds. Two crystals cooled to 4°C each yielded complete datasets, for a total of 228,243 full observations to 1.83 Å with observable diffraction extending beyond 1.5 Å. Intensities were integrated using the program DENZO (Otwinowski, 1986), and the cell parameters were found to be a=b=71.97 Å, c=115.04 Å, α=β=90°, γ=120°. Observations were scaled and reduced in point group 321 using the method of Fox and Holmes (Fox & Holmes, 1966) to give 30,830 reflections with an overall Rsym[4] of 9.4%. Amplitudes were assigned to weak and negative intensities by fitting to an *a priori* distribution (Wilson, 1949; French & Wilson, 1978). Systematic absences along c* narrowed the choice of space groups to $P3_121$ or $P3_221$.

Structure solution. The structure was solved by molecular replacement using the first monomer of the reported ECTS•dUMP•CB3717 ternary complex (Montfort, et al. 1990b), with the ligands and water molecules omitted, as a model. A rotation search followed by rigid body Patterson correlation refinement

---

$$^4 Rsym = \sqrt{\frac{\sum_{hkl}\sum_{i=1}^{N}(I_{avg} - I_i)^2}{\sum_{hkl}\sum_{i=1}^{N}(I_i)^2}}$$

(Brünger, 1990) of the top 106 coalesced peaks generated a top solution with a Patterson correlation coefficient of 0.105 using amplitudes with $F/\sigma_F > 2.0$ between 15.0 Å and 4.0 Å resolution and Patterson space vectors between 30 Å and 5 Å. This solution was consistent with a monomer in the asymmetric unit, where the physiological dimer would be generated by the (x,x,0) crystallographic two-fold. This monomer was rotated by 120° about **c** so that a one-dimensional translation search could be conducted in the **a** direction. A translation search in $P3_121$ along (x,0,1/3) using amplitudes with $F/\sigma_F > 2.0$ between 8.0 Å and 3.5 Å gave a solution with an R factor[5] of 42%. A search in $P3_221$ along (x,0,1/6) with the same data gave a top solution with an R factor of 51%. The transformed coordinates were refined by rigid body least squares minimization using X-PLOR to give an R factor of 36.7% using all data between 15.0 Å and 4.0 Å.

A difference electron density map ($(F_o - F_c)\,\alpha_c$) calculated using the rigid-body refined model displayed clear, detailed density for the ligands H_2folate and dTMP, including distinct density for the C7 methyl group of dTMP, confirming that the rotation and translation solutions were correct. The ligands were modeled into the difference electron density using FRODO (Jones, 1985). The complete model was subjected to alternating cycles of positional and B-factor refinement, including one round of simulated annealing

---

[5] $$R = \frac{\sum_{hkl} \left| (|Fo| - |Fc|) \right|}{\sum_{hkl} |Fo|}$$

(Brünger, 1989) and hand rebuilding. As a final step, occupancies and B-factors for the water molecules only were refined in alternating cycles. The final R factor is 18.3% for all data between 7.0 Å and 1.83 Å for a model with a total of 2386 atoms, including 164 water molecules. Two residues, Met 10(8) and Leu 260(208) have been modeled with two alternate conformations.

The ECTS•dUMP•CB3717 structure, refined to 1.97 Å, was previously the highest resolution TS structure reported. That structure had been refined using PROLSQ (Hendrickson & Konnert, 1979). To facilitate comparison between the ECTS•dUMP•CB3717 and the ECTS•dTMP•H_2folate complexes and to distinguish differences due to the increased resolution, different space group, different refinement schemes and different ligands, the ECTS•dTMP•H_2folate complex coordinates were used to initiate a new refinement of the ECTS•dUMP•CB3717 structure. A dimer of the ECTS•dTMP•H_2folate coordinates was rotated into the $P6_3$ space group, dUMP and CB3717 were placed in the active sites, and the waters from ECTS•dTMP•H_2folate were retained. These coordinates were subjected to simulated annealing refinement to minimize the bias from the starting structure. This was followed by alternating cycles of hand-rebuilding and refinement through minimization. This new refined ECTS•dUMP•CB3717 structure was used for all further structural comparisons.

ECTS•dTMP•H_2folate was superimposed on ECTS•dUMP•CB3717 by selecting a core of Cα atoms in the dimer (Perry, et al., 1990) and minimizing the rms deviation in these Cα

positions in the two structures (Kabsch, 1978), using the programs NEWDOME and GEM (Appendices 1 and 2).

RESULTS

[Table I]
[Table II]

Quality of structure. As the structure reported here is but the latest of a large number of TS structures, including four with independent heavy atom solutions (Hardy, et al., 1987; Matthews, et al., 1990a; Matthews, et al., 1990b; Montfort, et al., 1990b), we can be sure there are no gross errors in the main chain. The quality of this structure is further reflected in the crystallographic statistics in Tables I and II, and in the electron density figures (Figures 2 and 3). A 2Fo-Fc electron density map contoured at a 1 σ level displays continuous density for all but some surface sidechains. The electron density for most aromatic groups displays a hole through the center of the ring, characteristic of high resolution structures. The structure has good stereochemistry (Table II), there are no Ramachandran violations and 95% of the sidechains can be matched to the rotamers identified by Ponder and Richards (Ponder & Richards, 1987) within 3 standard deviations, as assessed by the program RAMPLUS (Appendix 3).

[Fig 2]
[Fig 3]

New packing arrangement. E. coli TS is a dimer of identical 30 kd monomers. The trigonal space group $P3_121$, observed for ECTS•dTMP•H_2folate but previously unreported for an ECTS crystal, places a single monomer in the asymmetric unit. Crystal structures

of the phosphate- and nucleotide-bound binary complexes of the enzyme (Hardy, et al., 1987; Perry, et al., 1990; Schiffer, et al., 1991) also have a monomer in the asymmetric unit. Our previous ternary complex (Montfort, et al., 1990b; Kamb, et al., 1992) and folate-bound (unpublished results) crystal structures, however, have crystallized in space group $P6_3$, with a dimer in the asymmetric unit. Under our crystallization conditions, the presence of both dTMP and H_2folate are required to yield crystals in space group $P3_121$, since crystallization with either one of these ligands alone did not result in this space group.

In the ECTS•dUMP•CB3717 structure, from space group $P6_3$, the lattice contacts are different between the two monomers in the asymmetric unit (Figure 4). In ECTS•dTMP•H_2folate, there is a single monomer in the asymmetric unit, and the crystal contacts are the same to both monomers of the physiological dimer. These contacts are different from those observed for either of the monomers in the ECTS•dUMP•CB3717 structure, though they are more similar to the lattice contacts for the second monomer (Figure 4).

[Fig 4]

The number of crystal contacts (defined as symmetry atoms located less than 3.5 Å away) is the same in both the $P3_121$ and the $P6_3$ crystal forms. However, the number of interdimer hydrogen bonds (closer than 3.2 Å) is higher in the $P3_121$ crystals, with 24 hydrogen bonds/dimer, as compared to only 14 hydrogen bonds/dimer in the $P6_3$ crystal form. Since the packing density is roughly the same in both space groups (54% solvent in the $P3_121$ crystal form

vs. 51% in the $P6_3$ crystal form), these extra hydrogen bonds may contribute to the improved resolution of the data obtained from the ECTS•dTMP•H_2folate crystals. The most extensive crystal contacts in ECTS•dTMP•H_2folate occur at residues His 53(51) and Arg 55(53), concurrent with large shifts in position for these residues compared to ECTS•dUMP•CB3717 (Figure 4).

Structure comparison: In order to determine which differences in structure observed between the ECTS•dUMP•CB3717 and the ECTS•dTMP•H_2folate complexes are due simply to the improved resolution of the current stucture, ECTS•dUMP•CB3717 was rerefined starting with the ECTS•dTMP•H_2folate complex as a model. This newly refined structure has the same R factor as the previous structure, but better geometry (Table II). The Cα positions changed by only 0.19 Å rmsd. The largest change occurred between residues 155 and 156 where a flip of the peptide plane was discovered in both the ECTS•dUMP•CB3717 and the ECTS•dTMP•H_2folate complexes relative to the original reported ECTS•dUMP•CB3717 structure. Further, extra density was observed confluent with the Sγ of Cys 52(50) and Cys 244(192) in both monomers. This was modeled as a single additional sulfur, and may reflect interaction of the protein with β-mercaptoethanol which was added during crystal growth. The overall rmsd between the reported and the newly refined ECTS•dUMP•CB3717 structures is 0.73 Å for all atoms in the protein, mostly due to those sidechains assigned to new rotamers.

The protein in ECTS•dTMP•H_2folate is in the "closed down" conformation, first seen for the ECTS•dUMP•CB3717 complex. After

superimposition of the dimers, there is an rmsd of 0.35 Å in Cα position (0.66 Å for all atoms in the protein) between the two structures.

Although the monomers are similar in both space groups, there is a change in the association of the monomers. In ECTS•dUMP•CB3717, the monomers are asymmetrically disposed, related by a rotation of 179.5° about an axis with direction cosines of (-0.0012, 0.4302, 0.9027) with respect to $P6_3$ orthogonalized axes, with a slight translation of 0.08 Å along the rotation axis. Thus, not only is the axis not a perfect two-fold, but it is far from where the two-fold would have to be to put the ECTS•dUMP•CB3717 structure in a higher symmetry space group ($P6_322$). In contrast, the monomers ECTS•dTMP•H_2folate are related by a strict crystallographic two-fold rotation axis (parallel to the x axis).

Because of this change in association, the monomers of each structure superimpose better than the whole dimers. Monomer 1 of the ECTS•dUMP•CB3717 structure superimposes on the monomer in the ECTS•dTMP•H_2folate structure with a rmsd of 0.30 Å for Cα (0.60 Å all atoms) while monomer 2 superimposes with a rmsd of 0.29 Å for Cαs (0.66 Å all atoms). This is even smaller than the differences between the two monomers in ECTS•dUMP•CB3717, which have a rmsd of 0.33 Å for Cαs (0.75 Å all atoms).

Although the magnitude of the positional differences is very small, there is an inward bias to the shifts (Figure 5). Thus ECTS•dTMP•H_2folate is slightly more closed down than ECTS•dUMP•CB3717. This is reflected in the radius of gyration for

the dimer, which decreases from 23.12 Å for in ECTS•dUMP•CB3717 to 23.04 Å in ECTS•dTMP•H_2folate.

[Fig 5]

Ligand binding. The dTMP and H_2folate are well defined in the electron density maps (Figures 2 and 3). The positions of the ligands in the ECTS•dUMP•CB3717 and ECTS•dTMP•H_2folate complexes are nearly identical. There is an rmsd of 0.27 Å for all matching atoms in the nucleotide and 1.11 Å for all matching atoms in the folate. The ECTS•dUMP•CB3717 structure exhibits a covalent bond between the Sγ of Cys 198(146) and the C6 of the dUMP. The protein in ECTS•dTMP•H_2folate is a Cys to Ser variant and there is no covalent bond present. Instead the serine sidechain rotates from the gauche$^-$ rotamer (χ_1 = -52°) seen for Cys 198(146) in ECTS•dUMP•CB3717 to the trans rotamer (χ_1 = 151°) and establishes a hydrogen bond to the mainchain carbonyl of Ser 219(167).

[Table III]

All noncovalent interactions around the ligands reported for ECTS•dUMP•CB3717 (Finer-Moore, et al., 1990) are preserved in ECTS•dTMP•H_2folate, with the following exceptions: a) The C-terminal carboxylic acid and Nε of Trp 85(83), hydrogen bonded in monomer 1 of the ECTS•dUMP•CB3717 complex, are 3.8 Å apart in ECTS•dTMP•H_2folate. This hydrogen bond is also absent in monomer 2 of ECTS•dUMP•CB3717 . b) There are additional interactions between the phosphate moiety of the nucleotide and the quartet of arginines ligating it (Table III) so that now each arginine is forming two distinct hydrogen bonds to the phosphate. c) Phe 228(176),

which interacts with the PABA moiety of the folate, has shifted, with a change in χ_2 angle from -150° to -85°, moving away from the propargyl group of CB3717. The shift in the sidechain repositions two waters. An additional water takes the place of the propargyl group of CB3717 (Figure 6).

[Fig 6]

Thymidylate differs from deoxyuridylate by the presence of a methyl group, C7, at the C5 position in the pyrimidine ring. In ECTS•dTMP•H_2folate, the only residue which comes within 3.5 Å of this methyl group is the side chain of the absolutely conserved Trp 82(80), which also interacts with the pterin of the folate. This residue is virtually unperturbed, rotating only 3° between ECTS•dUMP•CB3717 and ECTS•dTMP•H_2folate.

A water molecule, which has been designated water^{C7}, has shifted 0.5 Å away from the ligands to a position 3.4 Å from the new methyl group (Figures 2 and 6). This water is coordinated by the carbonyl of Ala (196)144 and the sidechain of absolutely conserved Tyr 146(94), which is also hydrogen bonded to the mainchain nitrogen of Ser 198(146). More dramatic than the shift in position is the decrease in occupancy of water^{C7}. In the ECTS•dUMP•CB3717 structure, water^{C7} has full occupancy and a B-factor of 9 Å^2 (compared to an average of 9 Å^2 for the ligating atoms). However, in ECTS•dTMP•H_2folate, water^{C7} has a relative occupancy of only 0.50 and a B-factor of 16 Å^2 (compared to an average of 16 Å^2 for the ligating atoms). In addition, there is another water, not present in the ECTS•dUMP•CB3717 structure, located 3.3 Å further removed

from methyl C7 which has a relative occupancy of 0.31 and a B-factor of 20 $Å^2$.

Modified N -terminus. The improved resolution has revealed some novel features in the thymidylate synthase structure. One is the presence of a chemical modification at the N-terminal nitrogen consisting of three covalently attached atoms (Figure 7). It is clear that this N-terminal modification is present also in the ECTS•dUMP•CB3717 structure. These atoms are planar with the main chain nitrogen and the two terminal atoms are hydrogen bonded to the sidechains and mainchains of the highly conserved Thr 48(46) and Thr 49(47) (Thr 49(47) is a leucine in the TS cloned from *Lactobacillus lactis* (Ross, et al., 1990)).

[Fig 7]
[Table IV]

Since both terminal atoms are receiving hydrogen bonds from main chain nitrogens (Table IV) these atoms have been identified as oxygens, creating a carbamate involving the N-terminal nitrogen. The modification is not simply a formyl group with two positions for the oxygen as the two terminal oxygens have full occupancy and B-factors of 17 $Å^2$ and 18 $Å^2$, close to that for the backbone atoms of residue Met 3(1) (19 $Å^2$).

In solution, carbamic acids rapidly decompose to release free carbon dioxide. In the context of the protein it appears the carbamate is stabilized by the threonine pocket, where it is sheltered from bulk solvent.

DISCUSSION

Nucleotide binding. The pyrimidine nucleotides dUMP and dTMP differ by the replacement of hydrogen by a methyl group at the C5 position in dTMP. However, dTMP binds three to seven times more weakly than dUMP, either with or without cofactor (Table V) in studies on TS from *L. casei.* By thermal titration, Beaudette (1980) showed that dUMP binding is driven primarily by enthalpy (ΔG = -7.1 kcal/mol, ΔH = -5.4 kcal/mol, $-T\Delta S$ = -1.7 kcal/mol). Thymidylate binding, however, is purely entropy driven (ΔG = -6.7 kcal/mol, ΔH = 0.7 kcal/mol, $-T\Delta S$ = -7.4 kcal/mol). This is consistent with the observation that dTMP is more hydrophobic than dUMP (as measured by partitioning coefficients from water to octanol) (Hansch & Leo, 1979) since hydrophobically driven associations are usually accompanied by an increase in entropy and a small decrease in enthalpy (Dill, et al., 1989; Da, et al., 1992).

[Table V]

The added hydrophobicity of the C7 methyl group is illustrated dramatically at the atomic level by the perturbation of water^{C7} near the C7 of dTMP (Figures 2 and 6). Water^{C7} is coordinated by the hydroxyl oxygen of Tyr 146(94), which is absolutely conserved, and the mainchain carbonyl of Ala 196(144). The presence of the methyl group shifts the water molecule, and greatly decreases its occupancy at this site. Thus, it appears that the enzyme is able to discriminate between the substrate and product nucleotides through this water molecule. In the absence of the C7 methyl group, water^{C7} is present and makes two strong hydrogen bonds to the protein.

However, with the addition of the C7 methyl group, this water is displaced, costing in enthalpy through the loss of hydrogen bonds to the two protein atoms, but gaining in entropy through the partial release of a bound water molecule to bulk solvent.

The removal of a bound water molecule is thermodynamically similar to the melting of ice which is accompanied by a gain in entropy but a loss in enthalpy. This was illustrated by Vriend et al. (Vriend, et al., 1991) in an experiment in which an alanine was replaced by serine in the neutral protease of *Bacillus stearothermophilus*. The serine displaced a bound water molecule and replaced the hydrogen bonding interactions. This results in a more stable protein, since the water is released to bulk solvent and there is no net loss of hydrogen bonds within the protein. In TS, however, the hydrogen bonds to water^{C7} are not compensated for in the protein, and a small cavity is left behind, which is also energetically unfavorable (Rashin, et al., 1986). Thus, in the case of TS, the waterless state (ECTS•dTMP) is less stable than the water-bound state (ECTS•dUMP).

Water-mediated ligand selectivity is also exhibited by the L-arabinose-binding protein. Through crystallographic analysis, Quiocho and coworkers showed that bound waters were responsible for favorable interactions to hydroxyl groups in L-arabinose and D-galactose (which bind with K_d values of 98 and 230 nM, respectively), and unfavorable interactions with a methyl group in D-fucose (which has a K_d of 3.8 μM). In TS, the water^{C7} makes no favorable interactions to either ligand, and correspondingly the

degree of selectivity is less in TS than in the L-arabinose-binding protein.

Aside from its role in ligand binding, water^{C7} may be involved in the chemistry of TS. One step in the reaction involves the removal of the proton at C5 (Pogolotti & Santi, 1977; Finer-Moore, et al., 1990). No base has been identified, but at 4 Å away, water^{C7} is the closest non-ligand atom to C5 other than the Sγ of Cys 198(146) in the ECTS•dUMP•CB3717 structure, and could act as the base and accept the proton from C5. Note that at the time of hydrogen-abstraction, C5 of dUMP is tetrahedral, with the C7 methylene covalently attached to the folate and directed away from the water^{C7} position. Additionally, the C5 proton is directed away from the folate, which means the C5-C5 proton vector would point almost directly at water^{C7}. Removal of the hydrogen results in a planar C5, which brings the C7 methylene closer to the position seen in the ECTS•dTMP•H_2folate structure, displacing water^{C7}.

Water^{C7} is distinct from the water favored by Matthews et al. as the hydrogen acceptor (Matthews, et al., 1990b) . That water, Wat 1 (in (Finer-Moore, et al., 1990); Wat 401 in (Matthews, et al., 1990b)), is 4.5 Å from C5 and hydrogen bonds to the absolutely conserved Glu 60(58) and the highly conserved His 199(147). His 199(147) is not preserved in two of the most enzymatically active TS sequences known (from ϕ3T (Kenny, et al., 1985; Maley & Maley, 1989) and *L. lactis* (Ross, et al., 1990)), however, and is likewise absent from sequences of the related hydroxymethylase enzymes, discussed below. Furthermore, Wat 1 is on the same side of the pyrimidine ring as the folate, and is not in a good position to receive

the proton (Figures 2 and 6). Matthews et al. do report a water near the position of water^{C7} (Wat 415 in their nomenclature). However, both of their ternary complexes (ECTS•FdUMP•CB3717 (Matthews, et al., 1990a), ECTS•FdUMP•CH_2-H_4folate (Matthews, et al., 1990b)) include FdUMP, which has a fluorine on the C5 carbon, which apparently perturbs water^{C7} so that in those structures water^{C7} is farther from C5 of dUMP than is Wat 1.

Mutagenesis of Tyr 146(94) can test the involvement of water^{C7} described above. Since water^{C7} does not favor binding of dTMP, replacement of Tyr 146(94) with a smaller sidechain should reduce this effect, and the K_d for dTMP should be closer to the K_d for dUMP for that mutant than in the wild-type enzyme. Further, if water^{C7} is involved in removal of the hydrogen from C5, the k_{cat} for mutants at Tyr 146(94) should be related to the hydrogen-bonding potential of the sidechain, except for the basic residues, which should be even worse as these will destabilize a positive charge on water^{C7}. Effects on k_{cat} due to the interaction of water^{C7} and Tyr146(94) will be compounded, however, with the effects due to the hydrogen bond between Tyr146(94) and the backbone of the catalytic thiol, which presumably is important for positioning of that side chain.

Three mutants of Tyr146(94) have been made: alanine, serine and proline (Climie, et al., 1990). All three show complementarity in a thy$^-$ deficient strain of *E. coli*, although the TS T146(94)P containing colonies grow weakly. No kinetic measurements are available for these mutants, however.

There is another possible chemical role for water^{C7}. In a reaction similar to that of TS, the hydroxymethylases transfer a methylene group from CH_2-H_4folate to the C5 position of a pyrimidine nucleotide. However, instead of adding a hydride from the cofactor, the hydroxymethylases add a water to C7, resulting in a hydroxymethyl group. Sequences of two hydroxymethylases are known: a dCMP hydroxymethylase (Lamm, et al., 1987; Thylen, 1988) which is 24% identical to ECTS and a dUMP hydroxymethylase (Wilhelm & Ruger, 1992), which is 22% identical to ECTS (aligned using the program GAP (Devereux, et al., 1984)). Tyr146(94) is conserved in these enzymes, and we expect water^{C7} to be present in a similar position. In ECTS•dTMP•H_2folate, water^{C7} is 3.4 Å from C7, and is nearly aligned with the π orbital, with an angle of 114° from C5 to C7 to water^{C7}, in a good position to add to a C7 methylene. The pterin moiety is exquisitely positioned in the active site, however, with C6 of H_2folate, the donor in the hydride transfer, 3.6 Å away, making an angle of 85° with C5 and C7 of dUMP. Thus, C6 of H_2folate is better positioned in TS to donate to C7, resulting in the creation of a methyl group, rather than hydroxymethyl.

From the structure and from mutagenesis it is known that the C-terminus is greatly involved in the positioning of the folate cofactor. Mutants of Val 316(264), the last residue in *L. casei* TS, have a higher K_m for the cofactor, while K_m for the nucleotide is little changed (Climie, et al., 1992). Deletion of just this last residue results in a protein which can bind both ligands, but is catalytically inactive (Galivan, et al., 1977; Carreras, et al., 1992). Both hydroxymethylase sequences known, while preserving many key

residues, differ greatly at the C-terminus. Deoxycytidylate hydroxymethylase ends 40 amino acids before the TS C-terminus, while dUMP hydroxymethylase contains an extra 120 amino acids past the TS C-terminus. As both sequences were isolated from bacteriophage, it is possible that the hydroxymethylases are simply TSs which have diverged to the point where methylene transfer can take place, but the folate C6 hydride donor is no longer perfectly aligned, allowing water^{C7} to complete the reaction. A mechanism for dUMP hydroxymethylase consistent with these predictions has been proposed by Kunitani and Santi (Kunitani & Santi, 1980).

Folate binding and overall protein conformation. The ECTS•dUMP•CB3717 and ECTS•dTMP•H_2folate structures are very similar, with an rms deviation of only 0.35 Å in Cα position. The slight shrinkage illustrated in Figure 5 may be due to either the greater number of strong crystal contacts in ECTS•dTMP•H_2folate, or the lower data-collection temperature employed in collecting the ECTS•dTMP•H_2folate complex data (4° C compared to 22° C for ECTS•dUMP•CB3717).

The ECTS•dUMP•CB3717 complex contains a covalent bond between the Sγ of Cys 198(146) and the C6 of the dUMP and thus we believe it represents a point on the reaction pathway after ternary complex formation, but prior to methyl transfer. The ECTS•dTMP•H_2folate structure represents a point after methyl transfer and after bond cleavage, but prior to dissociation of the ternary complex. The high degree of overlap between the ligands in the two structures suggests that relatively little conformational change occurs during carbon transfer and reduction, as opposed to

the extensive segmental accommodation observed for ligand binding. Thus it is likely that CH_2-H_4folate and dUMP in the activated complex are bound in the same conformation observed for the ligands in these two crystal structures, as assumed for our model of the activated complex, previously reported (Finer-Moore, et al., 1990).

The location of the H_2folate pterin ring in the primary binding site identified for the quinazoline of CB3717 and absence from the alternate site, does not support the notion that the alternate site is used for binding of H_2folate as was speculated previously. Although the function of the alternate site is still unknown, it may be involved in the enzyme mechanism, for example through initial binding of methylenetetrahydrofolate.

The largest chemical difference between CB3717 and H_2folate is the presence of the triple-bond containing propargyl group at N10 of CB3717 (Figures 1 and 6), which is important for the tight binding of that inhibitor. In the ECTS•dUMP•CB3717 complex, the aromatic ring of Phe 228(176) stacks against the propargyl group. This specific interaction is lost in the ECTS•dTMP•H_2folate structure, where Phe 228(176) is rotated away from the site where the propargyl group would be, to make more extensive van der Waals contacts with the para-aminobenzoic acid moiety of H_2folate.

Carbamate. The improved resolution ECTS•dTMP•H_2folate reveals several new features, principal among these being the N-terminal modification. In thymidylate synthases, the C-terminus is highly conserved and always ends at residue 316(264), where the carboxylate contributes to interactions maintaining the closed state. In contrast, the N-terminus can have a number of residues

prior to residue 3(1), to the extent of having an entire additional protein in the case of the bifunctional DHFR-TS enzymes (Beverley, et al., 1986). However, all the TS sequences known either begin with a methionine at position 3(1) as in *E. coli* TS, or have an acidic sidechain at this position. In the *L. casei* TS crystal structure (Hardy, et al., 1987), which has a glutamate residue at this position, the sidechain makes hydrogen bonds in the same threonine pocket occupied by the carbamate. Thus this could be a conserved feature of TSs, which serves as a link between regions involved in segmental accommodation, similar to conserved Tyr 6(4) which links the A helix to the J helix, and conserved His 264(212) which links the K helix to the C-terminal strand.

Although carbamates are unstable in solution, they have been observed in protein crystal structures before (Arnone, 1974, Lundvist & Schneider, 1991). In ribulose-1,5-bisphosphate carboxylase, a carbon dioxide molecule modifies a lysine within the active site. In hemoglobin, carbon dioxide is carried by covalent addition to the N-terminus.

Conclusion. The x-ray crystal structure of the product complex of thymidylate synthase suggests that the enzyme uses a bound water molecule to disfavor binding of the product nucleotide. This water molecule, named water^{C7} for its proximity to the C7 atom of dUMP may also be involved in proton abstraction from C5 in the reaction mechanism. An analogous water in the hydroxymethylases may be the source of the hydroxyl group after the methylene is transferred from CH_2-H_4folate in that enzyme's catalytic mechanism.

The high resolution of this structure, besides identifying a previously unreported N-terminal modification, allows better positioning of all residues in the structure. Our ongoing inhibitor-design efforts should be enhanced by this improved structure. Although the positional differences between ECTS•dTMP•H_2folate and ECTS•dUMP•CB3717 are slight, the configurations of many sidechains are different. In the dimer, 179 of 528 residues have been assigned to new rotamers. This could have a large effect on molecular mechanics calculations based on the thymidylate synthase structure.

ACKNOWLEDGMENTS

E. coli strain X2913 was kindly provided by Dan Santi. We thank Partho Ghosh for assistance with synchrotron data and Janet Finer-Moore, Chris Carreras and Brian Shoichet for many useful discussions. Figures 2, 3, 6 and 7 were generated using MolScript (Kraulis, 1991).

Table I: Crystallographic Data Statistics

Resolution (Å)	∞ - 5.60	4.00	3.27	2.84	2.54	2.32	2.15	2.01	1.90	1.80
R_{sym} (I) (%)	5.6	5.5	6.7	8.6	11.3	14.3	17.3	21.6	27.5	38.3
Ave. I/sig(I)	8.8	8.9	8.5	6.7	5.3	4.3	3.6	3.0	2.4	1.7
unique reflections, possible	1 254	1 999	2 544	2 927	3 352	3 634	3 898	4 294	4 274	4 845
unique reflections, collected	1 117	1 942	2 506	2 889	3 313	3 597	3 860	4 254	4 235	3 117
observations, collected	8 195	15 485	20 114	23 009	25 514	27 296	28 793	30 135	30 773	18 929
Completeness of data (%)	89	97	99	99	99	99	99	99	99	64
Redundancy	7.3	8.0	8.0	8.0	7.7	7.6	7.5	7.1	7.0	6.1
Resolution bins for R factor (Å)		7.0 - 3.59	2.92	2.57	2.34	2.18	2.05	1.95	1.87	1.80
R_{cryst} of refined structure (%)		13.6	15.6	18.2	19.6	20.4	23.0	24.6	28.2	31.2

Table II: Structural statistics

RMS deviations from ideality in final model[a]	Bond Lengths Å	Bond Angles °	Dihedral Angles °	Improper Angles[b] °	Ponder-Richards[c] %	Rama-chandran outliers[d]
ECTS•dTMP•H_2folate	0.012	2.68	25.08	1.01	95	0
ECTS•dUMP•CB3717	0.010	2.78	25.68	1.23	94	3
Montfort et al.[e]	0.030	5.07	28.52	3.81	85	3

[a] Measured for all non-hydrogen atoms against X-PLOR target values. [b] Improper angles define chiral centers and planar groups of atoms. [c] Percent of residues which can be assigned to one of the rotamers identified by Ponder and Richards within 3 standard deviations. [d] Number of non-glycine residues which fall outside allowed regions for left-handed helix, beta-strand, alpha-helix, or the "saddle" region between beta-strand and alpha-helix. [e] Previously reported ECTS•dUMP•CB3717 structure, which was refined against PROLSQ target values.

Table III: Hydrogen bonds to the phosphate in ECTS•dTMP•H_2folate

Arginine atom	Phosphate atom	distance (Å)[a]	angle (°)[b]
Nε 21	OP1	2.82	161
Nη1 21	OR5	3.01	169
Nη1 166	OP3	2.73	152
Nη2 166	OP2	2.80	151
Nε 126	OP2	3.05	135
Nη1 126	OP1	3.01	163
Nε 127	OP3	2.84	142
Nη1 127	OP3	2.79	144

[a] Distance between the named atoms. [b] Optimal donor-hydrogen-acceptor angle, assuming a donor-hydrogen bond length of 1.0 Å and that the phosphate is the hydrogen bond acceptor in all cases.

Table IV: Hydrogen bonds to the N-terminal modification

atom1	atom2	distance (Å)[a]	angle (°)[b]
OT1	THR^{47} Oγ1	2.78	175
OT1	THR^{47} N	2.98	155
OT2	THR^{46} Oγ1	2.72	176
OT2	THR^{46} N	3.00	156
OT2	WAT^{574}	2.79	180

[a] Distance between the named atoms. [b] Optimal donor-hydrogen-acceptor angle, assuming a donor-hydrogen bond length of 1.0 Å and that the carbamate is the hydrogen bond acceptor in all cases.

Table V: Reported dissociation constants for dUMP and dTMP

Method	K_d dUMP (μM)	K_d dTMP (μM)	$\frac{K_d \text{ dTMP}}{K_d \text{ dUMP}}$
Equilibrium Dialysis[a]	1.80	5.75	3.2
Equil. Dial. w/H_2folate[b]	0.52	3.27	6.3
Kinetics[c]	0.32	2.37	7.4
Thermal Titration[d]	5.9	17.5	3.0
Competition with PLP[e]	0.38	1.60	4.2

[a,b](Galivan, et al., 1976). [c](Daron & Aull, 1978). [d](Beaudette, et al., 1980). [e](D.V. Santi, personal communication). PLP is pyridoxal phosphate, an inhibitor of TS.

Figure legends

FIGURE 1: Chemical structures of TS Ligands. Positions in the structures mentioned in the text are indicated, as are named subunits of the folates and folate-analog. The source and destination of the methylene and hydride are highlighted with dashed circles.

FIGURE 2: Cross-eyed stereo view of dTMP binding. The nucleotide ligand is clearly revealed in this Fo-Fc electron density omit map, contoured at 6 σ. Before calculating this map, both ligands were removed from the Fc calculations. The C7 methyl group, can be seen on the right side of the pyrimidine ring. Residues mentioned in the text are labeled. Selected hydrogen bonds (<3.1 Å) are indicated with dashed lines. Crystallographic waters are drawn as dark circles.

FIGURE 3: Cross-eyed stereo view of H_2folate binding. All atoms of the folate ligand are in continuous density in this Fo-Fc electron density omit map, contoured at 3 σ. Before calculating this map, both ligands were removed from the Fc calculations. Residues mentioned in the text are labeled. Selected hydrogen bonds (<3.1 Å) are indicated with dashed lines. Crystallographic waters are drawn as dark circles.

FIGURE 4: Changes in structure compared to crystal contacts. The top half of the figure plots the Δr in Cα position between

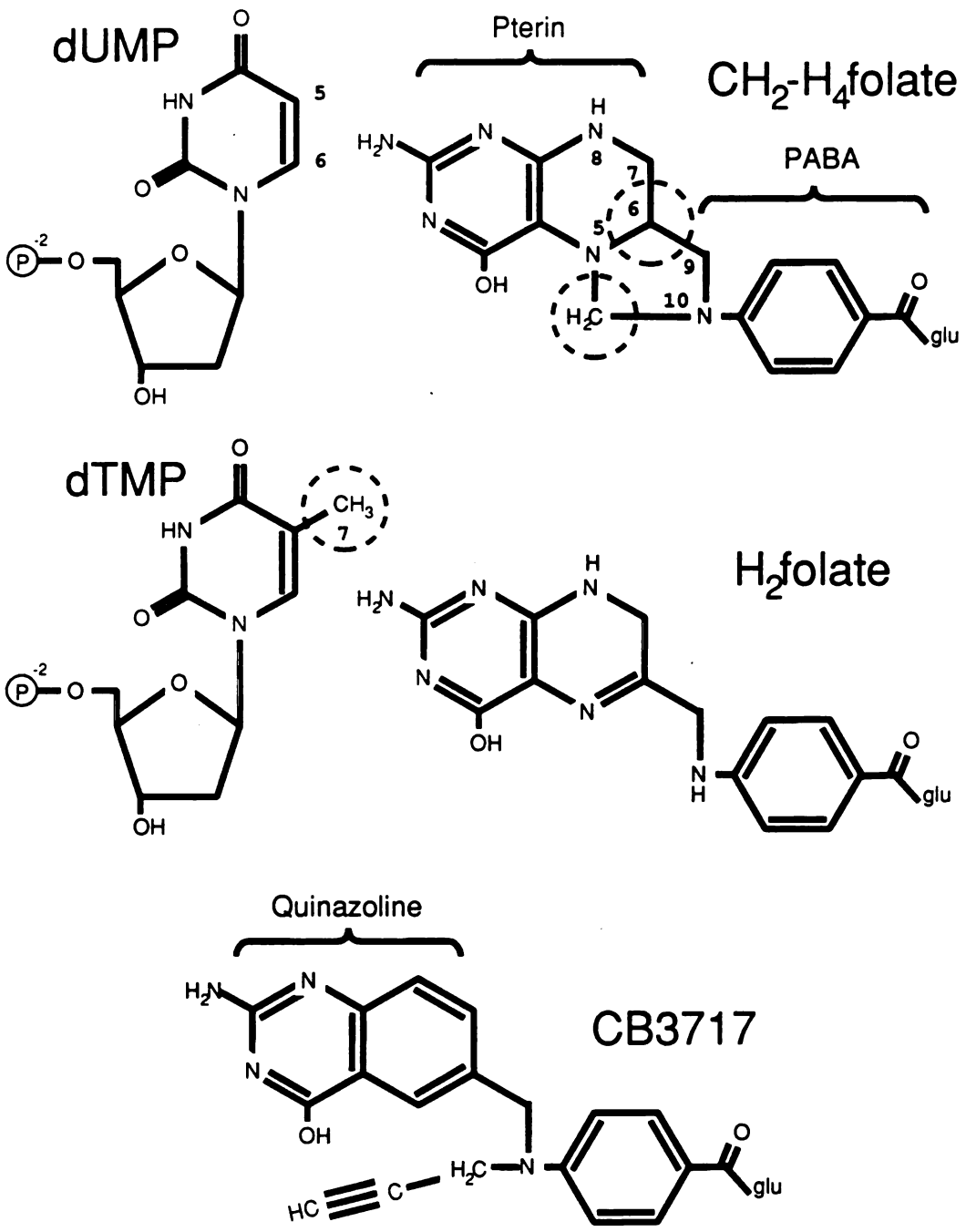
ECTS•dUMP•CB3717 and ECTS•dTMP•H_2folate after superimposing the dimers. The first monomer contains residues 3 to 316, the second 3' to 316'. The bottom half of the figure shows the number of different atom-atom contacts less than 3.5 Å there are involving each residue, as determined by the CONTACT routine in FRODO. Contacts in ECTS•dUMP•CB3717 are indicated below the line, while those in ECTS•dTMP•H_2folate are above the line.

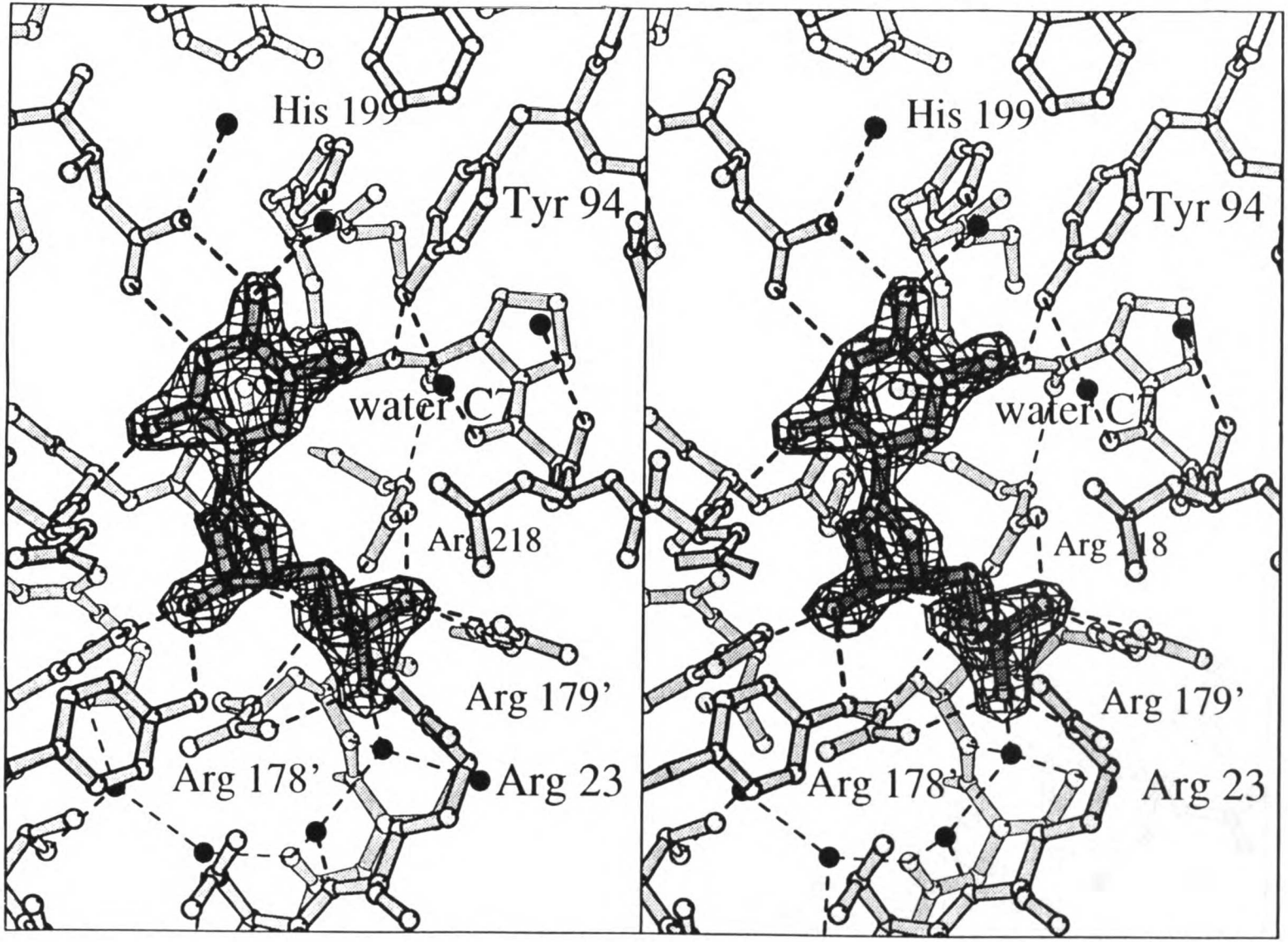
FIGURE 5: Shrinkage. Although the atomic shifts between ECTS•dUMP•CB3717 and ECTS•dTMP•H_2folate are very small, the inward bias of the direction of the shifts is revealed in this histogram. Theta is the angle between the vector from the atom to the center of mass for the dimer and the shift direction for that atom. Thus, a movement straight toward the center of mass would be an angle of 0°. The values are binned in 5° intervals and plotted as the number of occurrences in each bin, divided by the number expected based on random motion, which, for a bin from $\theta 1$ to $\theta 2$, is $\frac{\cos(\theta_1) - \cos(\theta_2)}{2(\theta_2-\theta_1)}$. The straight line at frequency=1.0 indicates the graph expected if there were no bias. The dimer, and also each monomer considered separately (not shown), all show a more shifts towards the center ($\theta < 90°$) than away ($\theta > 90°$), indicating that the ECTS•dTMP•H_2folate structure is closed down slightly more than the ECTS•dUMP•CB3717 structure.

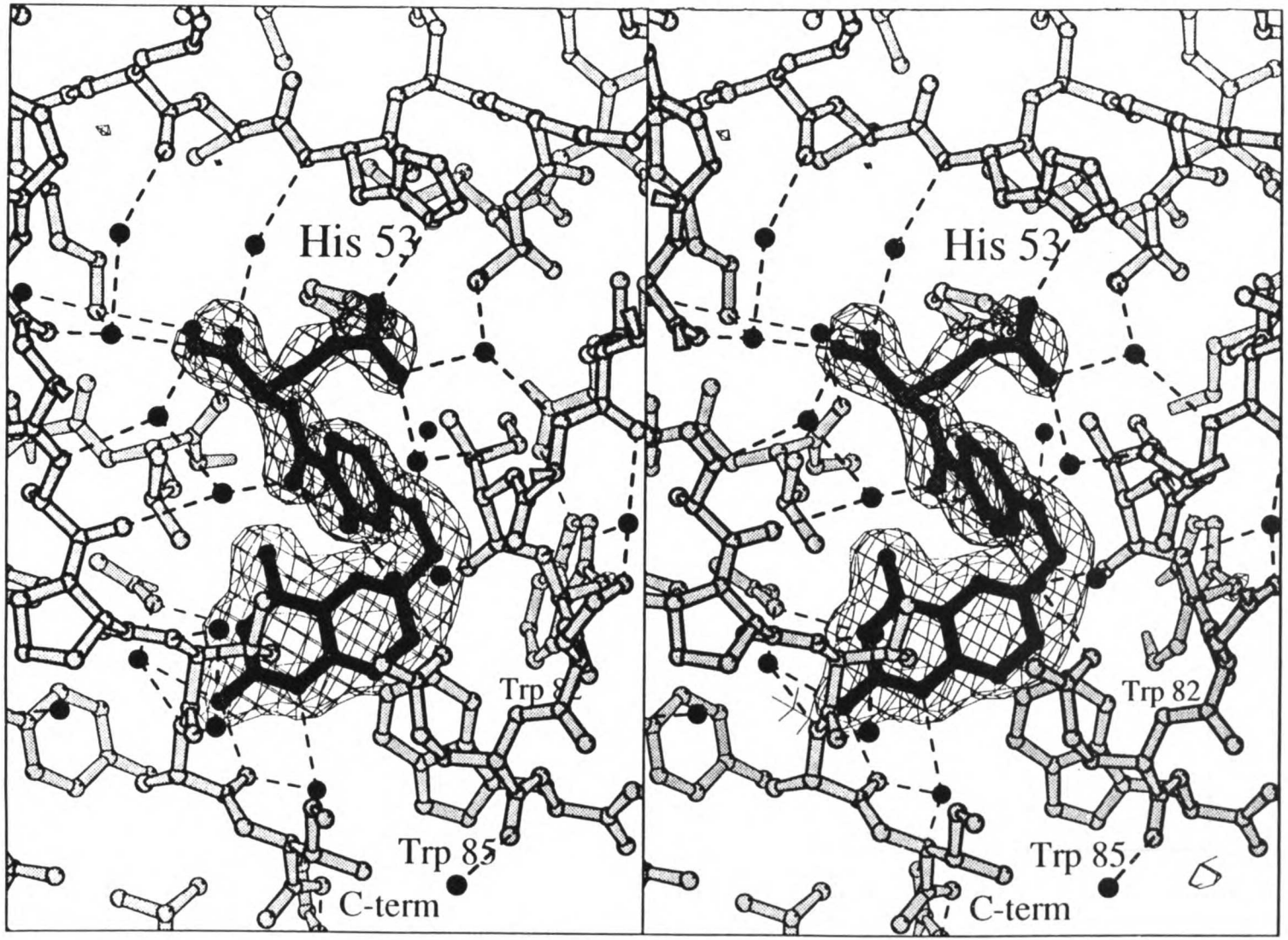
FIGURE 6: Cross-eyed stereo view of Phe 228(176). Phe 228(176) in ECTS•dTMP•H_2folate (black bonds) moves away from the site of the

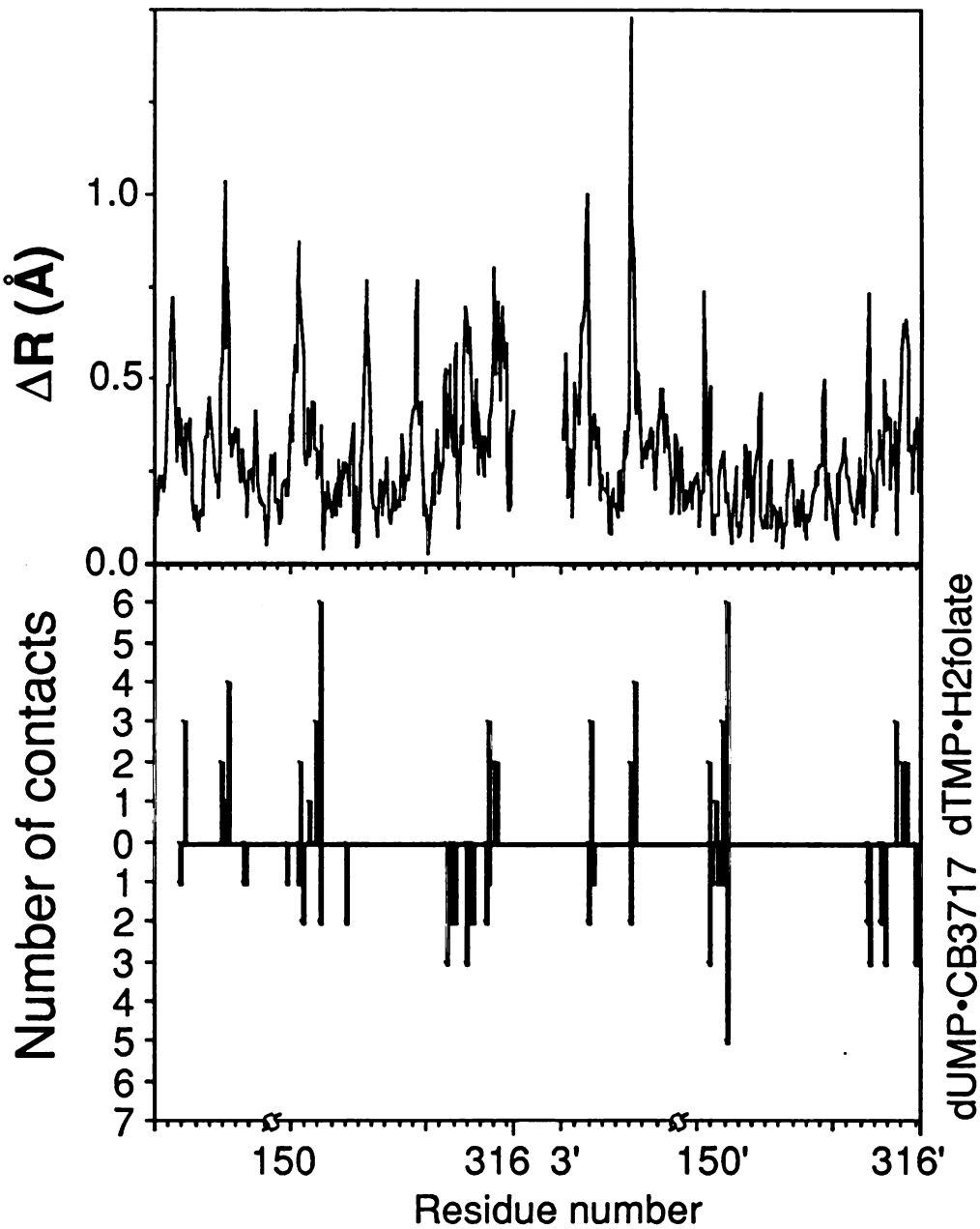
propargyl group in ECTS•dUMP•CH_2-H_4folate (white bonds). A water occupies the site of the propargyl, and two waters near the phenylaline move in concert with the sidechain. Most other waters have identical positions in both ternary complex structures, with the notable exception of water^{C7}, to the right of the nucleotide.

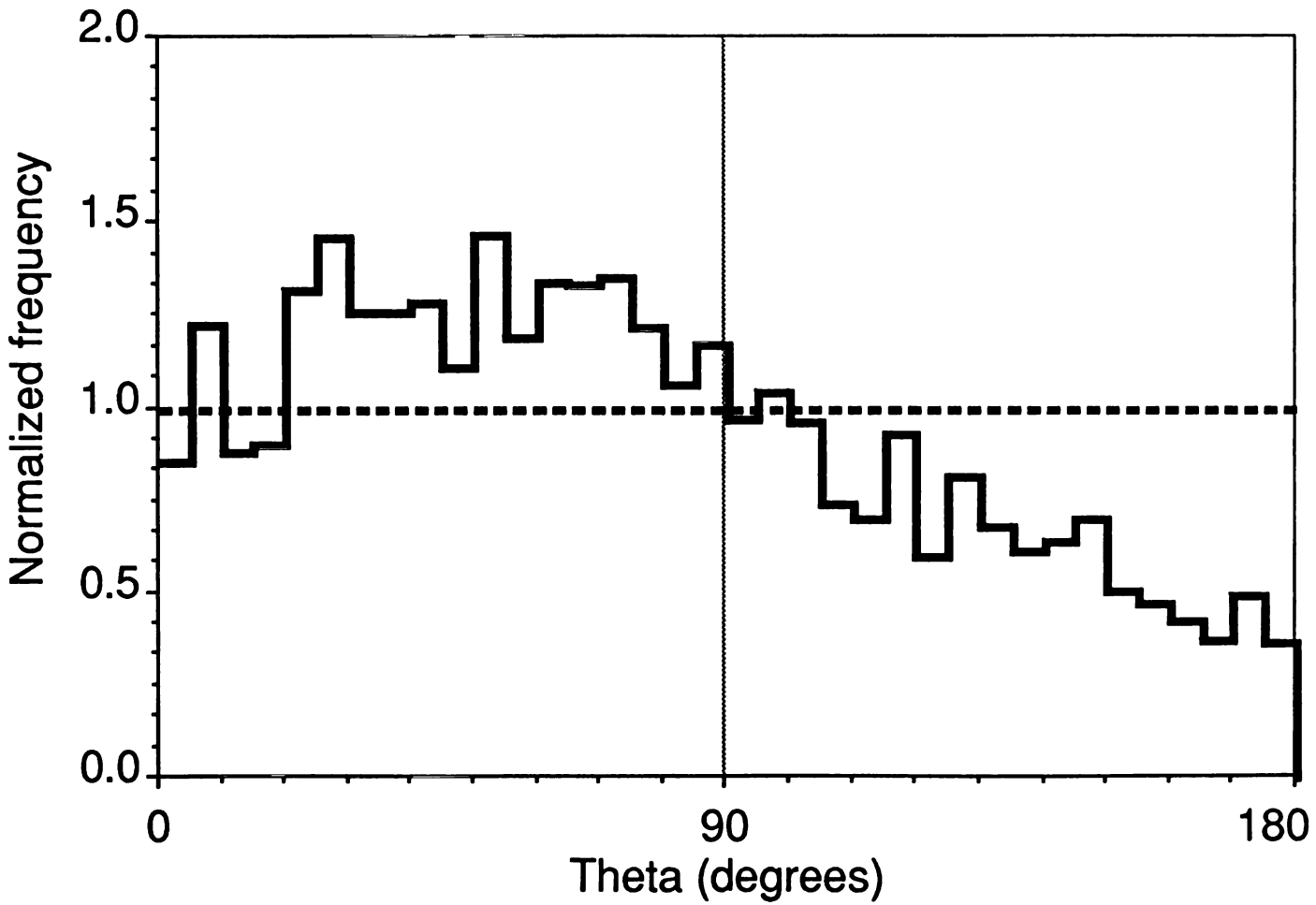
FIGURE 7: Cross-eyed stereo view of the N-terminal modification. The clear continuous density for the N-terminal modification is visible in this 2Fo-Fc electron density map, contoured at 1 σ. Carbons are displayed as light circles while non-carbons are drawn as dark circles. Hydrogen bonds involving the N-terminal modification are indicated by dashed lines.

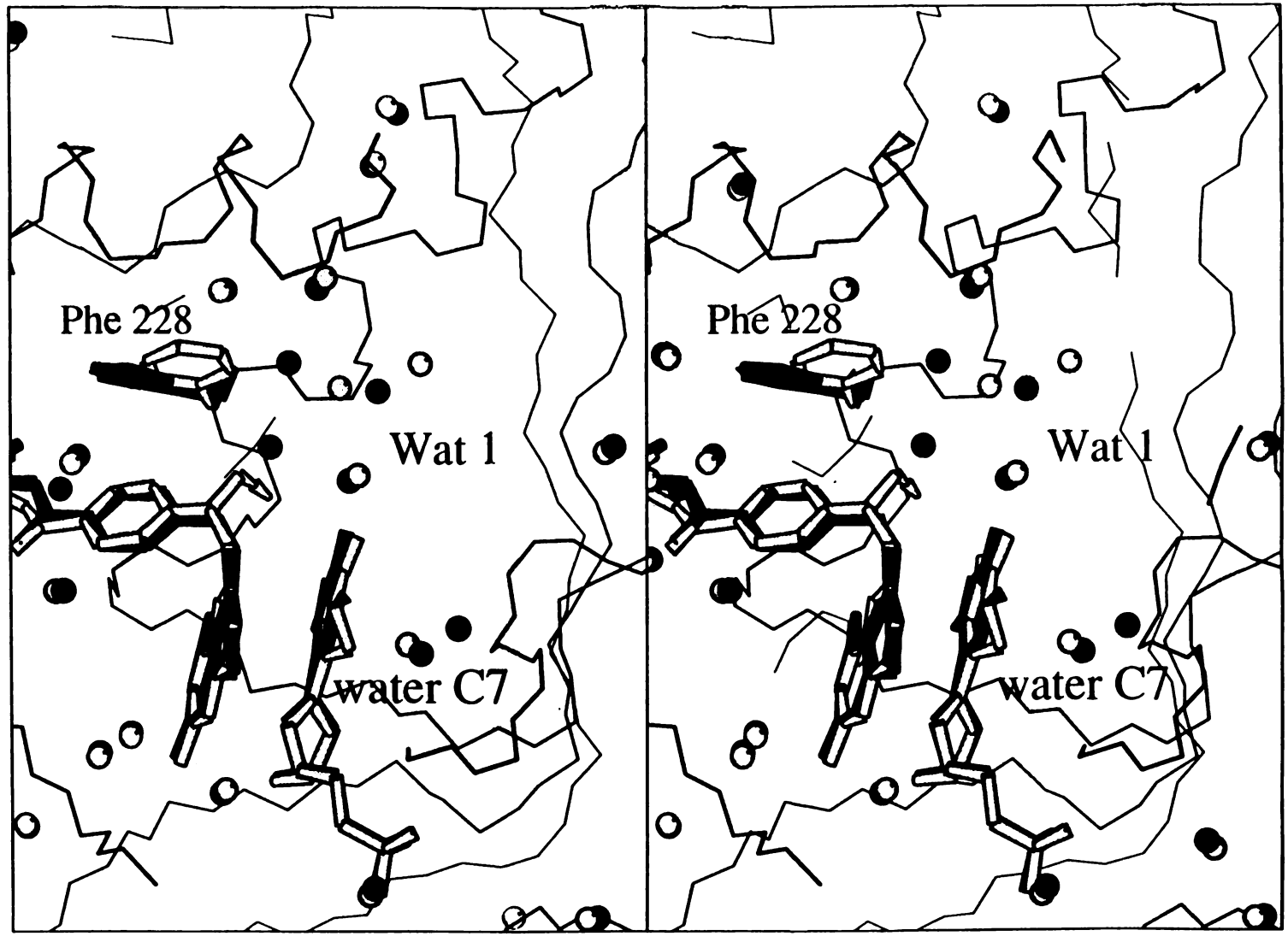

dUMP
Pterin
CH2-H4folate
PABA
dTMP
H2folate
Quinazoline
CB3717
glu


His 199
Tyr 94
water C7
Arg 218
Arg 179'
Arg 178'
Arg 23
His 199
Tyr 94
water C7
Arg 218
Arg 179'
Arg 178'
Arg 23


His 53
Trp 8
Trp 85
C-term
His 53
Trp 82
Trp 85
C-term


ΔR (Å)
1.0
0.5
0.0
Number of contacts
6
5
4
3
2
1
0
1
2
3
4
5
6
7
150
316
3'
150'
316'
Residue number
dUMP•CB3717 dTMP•H2folate


2.0
1.5
1.0
0.5
0.0
Normalized frequency
0
90
180
Theta (degrees)


Phe 228
Wat 1
water C7
Phe 228
Wat 1
water C7

Thr 48
Thr 49
Met 3
Thr 48
Thr 49
Met 3

D. Electrostatics and Ligand Binding

My thesis up until now has focused on short range type molecular interactions: hydrogen bonds, van der Waals interactions and covalent bonds. Electrostatics, on the other hand, can be very long interactions and several aspects of the TS mechanism require an analysis of this force. An examination of a electrostatic potential map of *E. coli* TS reveals two key areas of localized charge. One area is near the phosphate of the substrate nucleotide and the other is surrounding the para-aminobenzoic acid (PABA) moiety of the folate. The significance of both of these regions is discussed below.

1. the phosphate binding site

The following manuscript has been formatted for submission to Proteins: Structure, Function and Genetics.

Crystallographic Analysis of a Phosphate-Binding Arginine Quartet: Three mutants of R179 in *L. casei* Thymidylate Synthase

Eric B. Fauman and Robert M. Stroud

Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, California, 94143-0448

ABSTRACT **Arginine is the side-chain most often used in proteins to coordinate phosphate moieties. In the nucleotide-binding enzyme thymidylate synthase, the dianionic phosphate moiety is coordinate by four absolutely conserved arginines. Despite the high conservation of arginine 179, this amino acid can be replaced with alanine, glutamate, threonine and lysine with only minimal degradation of binding and catalysis. Interestingly, the substitution of lysine produced the greatest effect on Km. Crystallographic analysis of the R179A, R179K and R179E mutants complexed with inorganic phosphate explains the differential effects of the substitutions.**

Key words: electrostatics, plasticty, mutagenesis

INTRODUCTION

Thymidylate synthase (TS* , E.C. 2.1.1.45) catalyzes the reductive methylation of deoxyuridylate (dUMP) to deoxythymidylate (dTMP) using

* Abbreviations used: TS, thymidylate synthase; LCTS, TS from *Lactobacillus casei*; RMS, root mean square. Numbering of residues is by the L. casei convention, as in Hardy et al.(Hardy, et al., 1987). Residues from the second monomer are indicated by the apostrophe.

methylene-tetrahydrofolate (CH_2-H_4folate) both as a source for the carbon and the hydride. TS is a dimeric protein, composed of identical monomers, generating two identical active sites. Twenty sequences of TS have been reported from a wide variety of organisms, revealing that TS is one of the most highly conserved enzymes known(Perry, et al., 1990). Two absolutely conserved arginines (Arg-178′ and Arg-179′) from each monomer extend across the dimer interface to form contacts to the phosphate moiety of the nucleotide substrate in the other monomer(Montfort, et al., 1990). The phosphate binding site is completed with an additional pair of absolutely conserved arginines (Arg-23 and Arg-218) as well as a highly conserved serine (Ser-219).

The preferred sidechain for phosphate binding is arginine, although with the exception of TS no protein uses more than two arginines per phosphate moiety (Table I). In an effort to probe the roles of these four arginines in TS, all four amino acid positions have been subjected to mutagenesis in TS from *L. casei* (Climie, et al., 1990; Santi, et al., 1990). There are no catalytically active mutants at positions 23 or 218. However, both Arg-178′ and Arg-179′ can be replaced with little effect on kinetic parameters. Arg-179′ has been replaced with alanine, threonine, lysine and glutamate, and all four mutants are catalytically active.

Surprisingly, the substitution with the greatest effect on the kinetic parameters was R179K(Santi, et al., 1990). In order to understand the minimal effects of replacement of this absolutely conserved residue, and the especial effect of lysine at 179, we determined the crystal structures of three of these mutants in complex with inorganic phosphate. The structural results can explain the different kinetic parameters obtained for these mutants.

MATERIALS AND METHODS

The R179A, R179K and R179E mutants of LCTS were prepared, expressed and purified as described(Santi, et al., 1990). Hexagonal bipyramidal crystals were grown by vapor diffusion from 4 μl drops containing 2-6 mgs/ml protein, 100 mM potassium phosphate (pH 7.4), 10 mM DTT and 0.5% to 2.0% ammonium sulfate (percent saturated) suspended over a solution containing just the phosphate buffer, DTT and EDTA. Crystals appeared in 1-2 days, and grew to 400-600 μm in length.

Crystallographic data were collected on a Xentronics area detector using copper Kα radiation. Observations were integrated, scaled and reduced using the Xengen software package(Howard, et al.,). Total datasets required one or two crystals.

The previously determined structure of wild-type LCTS with bound phosphate(Finer-Moore, 1993) was used as a starting model in each case. The side-chain of Arg-179 was trimmed back to the Cβ, and the new side chain (if any) was built into an (Fo-Fc) electron density difference map, using Frodo(Jones, 1985). Structures were refined by simulated annealing refinement(Brünger, 1989) followed by cycles of positional refinement, B factor refinement (program X-PLOR(Brünger, 1989)) and hand rebuilding (program Frodo(Jones, 1985)).

RESULTS

All three mutants crystallized in space group $P6_122$ (a=b=78.3, c=243.2) as for the wild-type enzyme with inorganic phosphate or in binary complex with dUMP(Hardy, et al., 1987; Finer-Moore, 1993). Crystal statistics are

summarized in Table II. None of the mutants showed any large conformational change.

R179A

The replacement of an arginine with an alanine is simply the deletion of all atoms past the Cβ. The deletion of this absolutely conserved residue had relatively little effect on the binding of the inorganic phosphate ion seen in the crystal structure. As shown in Table III, the P_i is shifted slightly relative to its position in the wild-type LCTS structure solved in complex with a phosphate. The phosphorus atom of the P_i is shifted away from the neutral alanine at 179 and towards Arg-23.

R179K

Although the replacement of an arginine with a lysine is isoelectronic (at pH 7.4), this modification perturbed the placement of the P_i. Relative to the phosphorus position in the wild type structure, the phosphorus in R179K has shifted toward the Cα's at residues 23, 178 and 179. The phosphorus in R179K is nearly an Ångstrom from its position in the R179A mutant.

R179E

Due to the small size of the crystal and rapid crystal decay in the case of the R179E mutant, only a partial dataset was collected. The Fo-Fc difference electron density map calculated from this data and the R179A model is extremely noisy and is uninterpretable. It will be impossible to accurately

place the sidechain of Glu-179 until more data is collected on this mutant. However, it is evident at this point that the phosphate is present and there is little perturbation of the phosphate binding site.

DISCUSSION

Structural interpretation of Arg-179 mutants

The substitution of an arginine by a lysine is usually considered to be a conservative replacement. However, in the kinetic analysis of four mutations at position 179 in LCTS, a lysine substitution resulted in the greatest effect of kinetic parameters, greater than that seen for alanine, threonine or glutamate(Santi, et al., 1990) (see Table IV).

Since Arg-179 is known to interact with the dUMP substrate in the wild type structure(Finer-Moore, 1993), mutations at this site would be expected to have an effect on dUMP binding. As shown in Table III, Km for dUMP is increased in all the mutants relative to wild type. As derived by Santi(Santi, et al., 1990), Km/kcat for dUMP is equivalent to the on-rate for dUMP (k_1). Since the kcat values for all the mutants are the same, it is apparent that R179K has a reduced on-rate.

Pre-steady state kinetics of the wild type enzyme (Mittelstaedt & Schimerlik, 1986) have shown that binding of dUMP is a two step process; rapid formation of a weak complex followed by a slower isomerization. The K_d of the initial binding event is close to the K_i seen for inorganic phosphate, so it is possible that the first event is an electrostatic interaction between arginine quartet and the phosphate. Thus, in the case of the R179K mutant, the effect on k_1 would be due to interference with the isomerization event.

The structure of the R179K mutant with phosphate shows that while lysine is isoelectronic with arginine, it is not isosteric, and more importantly, the center of the positive charge is different in the two sidechains. Thus, the phosphorus is pulled closer to the Cα of 179 in R179K than in the wild-type. This relocation of the phosphate group could hinder the isomerization leading to binary complex formation.

The structure of R179K in complex with dUMP should show whether the full nucleotide is similarly slightly misplaced. If so, this could explain why only the R179K mutant has an effect on K_m of CH_2-H_4folate, as well as on K_m of dUMP. As derived by Santi et al. (Santi, et al., 1990), the effect on K_m of folate is probably attributable to interference with binding of the folate. Binding to TS is known to proceed in an ordered manner, with dUMP binding first(Danenberg & Danenberg, 1978). Further, dUMP forms part of the binding site for CH_2-H_4folate (Montfort, et al., 1990). Thus, if the dUMP is slightly misplaced, binding of CH2-H4folate would likewise be disturbed.

The R179A mutant had the least effect on kinetic parameters and also showed a smaller effect on the position of the phosphate relative to the binary complex than the R179K mutant. Understandably, the elimination of the electrostatic interaction between residue 179 and the phosphate causes the phosphate to move away from Ala-179. The remaining three arginines appear to be sufficient for initial binding and isomerization of dUMP binding. The structure of the binary complex of R179A with dUMP is expected to mimic the wild type binary complex structure.

The -2 change in charge from an arginine to a glutamate in the R179E had some effect on Km of dUMP, but not as large as the neutral substitution in R179K. Although the current structure does not allow placement of the Glu-179 sidechain, it is apparent that the phosphate group has not been

greatly displaced. With three other arginines in the phosphate binding pocket it is likely the Glu-179 sidechain is forming a salt bridge with one of them. For example, Glu-179 can be modelled to interact with the Nε of Arg-218. This keeps Glu-179 out of the binding pocket and shields the negative charge of Glu-179 from the negative charge of the phosphate behind the Nη1 and Nη2 of Arg-218. As with the R179A mutant, a binary complex of dUMP with R179E is expected to show no deviations from the binding seen in the wild-type binary complex.

Role of Arg-179

From mutagenesis it is apparent that Arg-179 is not essential for catalysis in thymidylate synthase. The high degree of conservation of this residue, however, argues that it serves an important function. This function is still unknown, but it is possible Arg-179 may be involved in dimer formation, in interaction with other proteins or substrate specificity or competition for substrates with other enzymes.

CONCLUSION

Structural analysis of 3 mutants of L. casei thymidylate synthase at position 179 provides a rational hypothesis for the kinetic parameters of the mutants. R179A, elimination of a positive charge, has little effect on binding of dUMP because there are still more than enough arginines to electrostatically interact with the phosphate moiety. R179K, on the other hand, hinders the formation of the wild-type dUMP binding mode because

the positive charge of lysine is positioned closer to the mainchain than is the positive charge of arginine. In R179E, the negatively charged Glu-179 may be positioned outside of the phosphate binding site, and thus behaves more like R179A, which has almost side chain, than like R179K, which actively disrupts phosphate binding.

ACKNOWLEDGEMENTS

This work was supported in part by an NIH grant to R.M.S. E.B.F. is a Howard Hughes Medical Institute Predoctoral Fellow. Some simulated annealing refinements were performed on a Cray YM-P at the Pittsburgh Super Computer Center. Purified mutants of LCTS were kindly provided by Dan Santi. The authors would like to thank Michael Shuster, Janet Finer-Moore and Kathy Perry for assistance with the crystallography.

Table I. Summary of Crystallographically Observed Phosphate Binding Sites

PDB	ligand	#P	mcN	D	E	H	K	N	Q	R	S	T	Y	H2O	Mg
2SNS	PTP	2		1						2			1		
3DFR	NDP	3	8						1	2		2		2	
1GD1	NAD	2	5		1	1							1	4	
4FXN	FMN	1	4					1						7	
1RNT	2GP	1									2	2			
2CFS	COA	3	1				1			2				3	
1PHH	FAD	2	1	1						2				1	
1PFK	FBP	2	1	1		1				4		1		7	1
	ADP324	2	5	1						1	1			6	
	ADP326	2	1		1					3	1			5	1
1WSY	PLP	1	5			1		1			1	1			
3GAP	CMP	1	1							1	1				
4MDH	NAD	2	2						1					7	
8CAT	NDP	3	1			1	1			1				1	
TS	DUM	1								4	1			1	
	totals		35	4	2	4	2	2	2	22	7	6	2	44	2

The Brookhaven Protein Databank (the PDB) was searched for crystal structures of proteins complexed with phosphate containing ligands. A group is listed above if one of its atoms is within 3 Å of any of the phosphate oxygens of the ligand (program PHOSLIG by E.B.F). Any symmetry matrices given in the REMARK records of the PDB file were included in determining interatomic distances. In this table, PDB indicates the identification given by the Databank, ligand is the residue name of the ligand in the file, #P is the number of phosphorus atoms in the ligand, mcN is the number of ligating main chain nitrogen atoms, H2O is the number of nearby crystallographically observed water molecules, and Mg is the number of ligating magnesium atoms. The remaining columns refer to the one-letter amino acid codes of any side-chains interacting with the phosphate. The proteins in the crystal structures used were staphylococcal nuclease(Cotton, et al., 1979), dihydrofolate reductase(Bolin, et al., 1982), glyceraldehyde-3-phosphate

dehydrogenase(Skarzynski, et al., 1987), flavodoxin(Smith, et al., 1977), ribonuclease(Arni, et al., 1987), citrate synthase(Remington, et al., 1982), hydroxybenzoate hydroxylase(Schreuder, et al., 1988), phosphofructokinase(Shirakihara & Evans, 1988), tryptophan synthase(Hyde & Miles, 1990), catabolite gene activator protein(Weber & Steitz, 1987), malate dehydrogenase(Birktoft, et al., 1989), catalase(Fita & Rossman, 1985) and thymidylate synthase(Montfort, et al., 1990).

Table II. Summary of Crystallographic Statistics

Mutant	Reso-lution	Reflect-ions	Redund-ancy	R-sym	Com-pleteness	R factor
R179A	2.55	12,977	7.5	14.0%	84%	19.5%
R179K	2.34	17,573	2.5	21.6%	89%	25.7%
R179E	2.67	5,772			48%	28.6%

Table III. Effect of mutations on phosphate position

	Shift of phosphorus in phosphate or in dUMP			Distance from phosphorus to Cα of:			
structure	wild type dUMP	wild type phos-phate	R179A phosphate	178′	179′	23	218
wild type dUMP				5.4	6.3	6.6	8.2
wild type phosphate	0.43			5.5	6.6	7.3	8.3
R179A phosphate	0.28	0.60		5.6	7.0	7.0	8.2
R179K phosphate	0.70	0.38	0.94	5.1	6.3	6.8	8.4

Shifts were measured after first superimposing the pair of structures by minimizing the RMS deviation of all Cα's in the dimer.

Table IV. Kinetic properties of Arg-179 Mutants

protein	Km dUMP (μM)	kM folate (μM)	kcat (sec^{-1})
wild type	2.7	20	7.8
R179A	5.2	20	2.4
R179T	6.9	20	1.9
R179E	17	16	1.9
R179K	38	66	2.0

Data from Santi, et al.(1990)

2. the PABA ring and the capacitor model

From the first time I learned of the closing of the active site of thymidylate synthase in response to ligand binding, I speculated that this conformational change may be electrostatically driven. Specifically, I constructed what I referred to as the "capacitor model." The idea this: the opposite sides of the active site have opposite charge, like a capacitor. When ligands are absent the active site is filled with water, which has a high dielectric, and thus the opposite sides of the active site are shielded from each other. When ligands are bound, the lower dielectric of these organic compounds allows the opposites sides to see each other, creating a driving force closing down the active site.

Data supporting this model followed the original conception. Specifically, it requires that the opposite sides of the active site be oppositely charged, and that the substrate fit between them. My first step was to construct an electrostatic potential map for the *E. coli* unbound and bound structures. This immediately revealed that the bulk of the protein is positively charged, while the small domain (the B and C helices) are negatively charged. Furthermore, from comparison with other TS sequences, many of these charges are conserved.

From examination of the electrostatic potential map and the sequence alignments, I was able to identify two ellipsoidally shaped regions (see function AXES in the program GEM in the appendix). The amino acids in these two regions are given in the tables below. The residue numbers given are those in the *E. coli* sequence (to convert to *L. casei* numbering add 2 to any number less 89 and add 52 to the rest). The letters to the left of each residue number are the single letter amino acid codes for the residue at that position in each of 18 TS sequences, viz., transposon TN4003 from

staphylococcus aureas, *Lactobacillus casei, Escherichia coli, Bacillus subtilus, Leshmania amazonensis, Leshmania major, Crithidia fasiculata, Plasmodium falciparum, Candida albicans, Saccaromyces cerivicia, Pneumocystis carinii,* human, mouse, herpes virus saimiri, herpes virus atales, Vericella zoster, phage T4, phage ϕ3T. Complete references for each of these sequences can be found in the bibliography at the end of this thesis. Positively charged amino acids are indicated by **BOLD** letters, negatively charged amino acids by *ITALICIZED* letters and uncharged amino acids by lower case letters.

Positive Domain

residue number	amino acids																	
48	**K**	**K**	**K**	**K**	**K**	**K**	**K**	**K**	**K**	**K**	**K**	**K**	**K**	**K**	**K**	**K**	**K**	**K**
211	n	n	n	n	s	n	n	n	*D*	*D*	*D*	n	n	*D*	n	n	n	**R**
212	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**	**H**
213	m	l	m	i	i	v	v	i	i	i	i	i	i	i	v	i	v	i
215	a	q	q	q	a	a	a	s	a	a	a	p	p	a	a	a	q	n
216	i	i	t	v	l	l	l	l	l	l	l	l	l	l	l	l	c	l
217	**H**	**K**	**H**	n	**K**	**K**	**K**	**K**	**K**	**K**	q	**K**	**K**	**K**	t	**K**	**K**	**K**
218	t	*E*	l	l	a	a	*E*	i	*E*	*E*	q	i	i	m	*E*	v	*E*	i
219	q	q	q	q	q	q	q	q	q	q	q	q	q	q	q	q	i	q
220	l	l	l	l	l	l	l	l	f	i	l	l	l	l	l	l	l	m
221	s	s	s	*E*	*E*	*E*	q	n	*E*	t	t	q	q	t	t	a	**R**	*E*
222	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**
255	**H**	y	**H**	**H**	**H**	**H**	y	**H**	y	**H**	y	**H**	**H**	**H**	y	**H**	**H**	g
257	a	a	g	**H**	p	a	p	**K**	p	**R**	t	t	t	i	s	p	p	**R**
258	i	i	i	i	i	i	i	i	i	i	i	i	i	i	i	l	i	l
259	**K**	**K**	**K**	**K**	**K**	**K**	**K**	s	**K**	*E*	**K**	**K**	**K**	**K**	**K**	**K**	**K**	l
260	a	a	a	g	m	m	m	m	m	m	m	m	m	m	m	m	g	f
261	p	p	p	a	*E*	*E*	*E*	*D*	**K**	**K**	**K**	*E*	*E*	**H**	p	*E*	**K**	*E*

Negative Domain

residue number	amino acids																	
57	t	s	**H**	**H**	*E*	*E*	*E*	*E*	**H**	l	*E*	*E*	*E*	*E*	*E*	*E*	a	**K**
58	*E*	*E*	*E*	*E*	*E*	*E*	*E*	*E*	*E*	*E*	*E*	*E*	*E*	*E*	*E*	*E*	*E*	*E*
59	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l
61	w	w	w	w	w	w	w	w	w	w	w	w	w	w	w	w	w	w
66	*D*	*D*	*D*	*D*	*E*	*E*	*E*	*E*	s	*D*	*E*	s	s	s	s	s	s	**K**
67	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	s
68	n	n	n	n	n	s	n	n	*D*	*D*	*D*	n	n	*D*	*D*	*D*	n	n
69	i	i	i	v	a	a	a	g	a	a	s	a	a	s	s	s	v	v
70	q	**R**	a	**R**	q	q	**H**	n	**K**	n	l	**K**	**K**	**K**	**K**	**K**	n	t
71	y	f	y	y	l	l	v	t	i	l	**K**	*E*	*E*	*E*	*E*	*E*	*D*	*E*
72	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l
73	l	l	**H**	q	a	a	a	l	s	s	**R**	s	s	s	a	a	**R**	n
74	**K**	q	*E*	*E*	*D*	*D*	*D*	n	*E*	*E*	*E*	s	s	a	a	a	l	**K**
75	y	**H**	n	n	**K**	**K**	**K**	**K**	**K**	q	**K**	**K**	**K**	a	s	**K**	i	m
76	n	**R**	n	g	*D*	*D*	*D*	n	g	g	n	g	g	g	g	*D*	q	g
77	n	n	v	v	i	i	i	v	v	v	u	v	v	v	v	i	**K**	v
78	n	**H**	t	**R**	**H**	**H**	**H**	**R**	**K**	**K**	**H**	**K**	**R**	**H**	**H**	**H**	t	**H**
79	i	i	i	i	i	i	i	i	i	i	i	i	i	i	i	i	v	i
80	w	w	w	w	w	w	w	w	w	w	w	w	w	w	w	w	w	w
81	n	*D*	*D*	n	*D*	*D*	*D*	*E*	*E*	*D*	*D*	*D*	*D*	*D*	*D*	*D*	*D*	*D*
82	*E*	*E*	*E*	*E*	g	g	g	a	g	g	a	a	a	a	a	i	*E*	q
83	w	w	w	w	n	n	n	n	n	n	n	n	n	n	n	y	n	w
84	a	a	a	a	g	g	g	g	g	g	g	g	g	g	g	g	y	**K**
85	f	f	*D*	*D*	s	s	s	r	s	s	s	s	s	s	s	s	*E*	q
86	*E*	*E*	*E*	*E*	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	**R**	s	n	*E*
87	n	**K**	n	n	*E*	*E*	*E*	*E*	*E*	*E*	*E*	*D*	*D*	s	s	**K**	q	*D*
88	g	g	g	g	m	m	m	n	g	g	g	g	g	g	g	g	g	g
89	n	*D*	*D*	*E*	*D*	*D*	*D*	*D*	*D*	*D*	*D*	*D*	*D*	*D*	*D*	*D*	*E*	t
90	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	i

The above tables show that the positive domain summed over the 18 sequences contains a total of 107 positive amino acids and only 24 negative amino acids. From the random distributions in globular proteins, there should be only 45 positive amino acids and 38 negative ones. Likewise, the negative domain contains 118 negative amino acids, and only 61 positive ones, as opposed to the 59 negative and 70 positive amino acids expected from random distribution. Thus the positive and negative regions required by my capacitor model exist and appear to be evolutionarily conserved.

Throughout evolution, then, the positive domain maintains a net charge of +5 e^-, while the negative domain maintains a net charge of -3 e^-. The two domains communicate through the PABA moiety of the folate through two nearly absolutely conserved hydrophobic residues (Ile-79 on the negative domain side, which is a valine in one TS sequence, and Leu-172 on the positive domain side, which is an isoleucine in one sequence and a tyrosine in another). Thus, the main hypothesis of the capacitor model is that the protein should not appear in the closed down state (ternary complex like) unless there is something between Ile-79 and Leu-172.

I have attempted to perform some electrostatics calculations to estimate the magnitude of the electrostatic interaction across the active site. The problem with doing such calculations is that numbers calculated for two different states of the enzyme can't be directly compared. Perhaps an extremely careful analysis of this system can minimize this problem. What analysis I've done suggests that the key residues for this interaction are His-212 and Lys-48 on the positive side and Glu-58 on the negative side. Unfortunately, these residues are highly conserved and may be directly involved in catalysis or ligand binding, instead of solely indirectly through

Chapter 4.

The Importance of Conformational Change to Rational Drug Design

The preceding two chapters have discussed how one can analyze and interpret conformational changes observed in protein structures. The knowledge of how a protein's structure can be altered by various stimuli may be useful in predicting how a protein will respond to a new stimulus.

Being able to accurately predict the conformational change a protein will undergo in binding to a small molecule is necessary to the success of rational drug design. Some parts of the protein may move closer to the ligand, as in segmental accommodation, described in the last chapter. This contributes additional binding energy which must be included if one wants to predict the total binding energy of the ligand. Other parts of the protein may move away from the ligand, allowing the protein to bind ligands which might appear to be too large for the active site.

Because of the complexity of protein structures, rational drug design efforts usually assume that the protein is rigid. This assumption led to a surprise in our study of the binding of UCSF8 (the thioketal derivative of haloperidol) to HIV-1 protease. One attempt at predicting the binding site of UCSF8 used the protein from a complexed structure which has a "closed down" active site. However, in the actual UCSF8•HIV-1 protease structure, the parts which close down on the active site (called "the flaps") are moved away from the active site, creating a new binding site which is occupied by the ligand, UCSF8. [Another attempt used a structure with an "open" active site, but in this case, not knowing the structure of the ligand is what led to an incorrect prediction. The whole issue is made more complicated by the solution of a structure with UCSF8 and a Q7K mutant of HIV-1 protease, in which UCSF8 does bind in the predicted site.]

The full HIV-1 protease drug design story is presented in the following manuscript which (as of January 1993) has been submitted to Nature:

Earl Rutenber, Eric Fauman, Robert Keenan, Susan Fong, Paul Furth, Paul Ortiz de Montellano, Elaine Meng, Irwin D. Kuntz, Dianne DeCamp, Rafael Salto, Jason Rosé, Charles Craik, and Robert M. Stroud (submitted) Structure of HIV-1 Protease Complexed with a Non-peptide Inhibitor: Initiating a Cycle of Structure-Based Drug Design.

Although our understanding of why the predicted site was correct in one case and not in another is incomplete, it is clear that a protein can undergo large conformational changes in response to ligand binding, and such changes must be anticipated.

The easiest method to predict what changes will occur is to start with a crystal structure of the protein bound to something analogous to desired ligand. For example, the structure of thymidylate synthase (TS) bound to deoxyuridylate (dUMP) is very similar to that of TS bound to inorganic phosphate alone, and the ternary complex structure of TS with dUMP and the antifolate is very similar to the product ternary complex, discussed in the last chapter.

Another method would take advantage of our knowledge of crystallographic B factors as indicators of protein plasticity, as discussed in the last chapter. Ideally, a curve of plasticity versus B factor would be constructed from a number of crystal structures of the target protein. However, the curve presented for the *E. coli* to *L. casei* TS comparison (Perry, et al., 1990) should provide a rough estimate of the allowed magnitude of positional shifts in a protein structure.

In contrast to the global plasticity observed for a large number of mutational changes, however, the conformational change observed for ligand

binding is localized, and tends to treat secondary structural elements as rigid units. A more sophisticated approach to prediction of protein response to ligand binding would probably incorporate this rigid motion of secondary structural elements with respect to mainchain atoms, but would still allow sidechains to move in a more plastic manner to maximize interactions.

In these ways, our understanding of proteins as flexible molecules can enhance our use of static crystal structures in the understanding of dynamic processes.

Conclusion

Macromolecular crystallography is not really feasible without computers to keep track of the thousands of intensity observations and the thousands of atomic positions. Likewise, it requires the use of a computer to help a crystallographer understand the conformational changes taking place, or other significant structural changes.

I have tried to show in each sort of structure comparison I did how I attempted to tailor the analysis to the system under study. That is, my analysis has always proceeded in two steps: first, I step back and try to evaluate what type of measurements are needed, a sort of "meta-analysis." Only then do I run (or most likely first write) the necessary programs to make the measurements.

Thus, the most important lesson from this thesis may not be the new methods which I've developed for exploring protein structures, but perhaps the methods by which these methods were developed. As Thomas Earnest is fond of saying: "If your only tool is a hammer, you tend to see every problem as a nail." I have strived to investigate how new tools are invented.

Two original first author manuscripts are presented in this thesis, and while they each have their own conclusions there are a couple points worth emphasizing here.

In the section on errors and B factors, the observation is made that the level of positional uncertainty in a protein is related to the ratio of the observations to parameters. Using this information, it should possible to optimize our refinement schemes to minimize the number of parameters we have to refine to have a more favorable degree of overdeterminancy. I point for example to recent attempts to use normal modes analysis in the refinement of B factors. Or perhaps using rigid body models for side chains

would actually yield more accurate structures. Our current models to explain the electron density we see are only models, and we are free to construct new models if needed. Now that we know of the free R-factor, we can use this technique to evaluate whatever models we may come up with.

We have known the structure of thymidylate synthase for several years, but we are only just starting to get a good picture of the structure of the water around the protein. The 1.83 Å structure reported in this thesis greatly expanded our knowledge of the extent of the involvement of bound water molecules in binding the ligands and in the reaction mechanism itself. The crystal form which yielded that structure holds the promise of a 1.5 Å or even higher resolution structure. I predict that we have still underestimated the role of water in the thymidylate synthase structure, and this will become apparent when the 1.5 Å structure is solved.

Bibliography

Arni, R., Heinemann, U., Maslowska, M., Tokuoka, R. & Saenger, W. (1987) Restrained least-squares refinement of the crystal structure of the ribonuclease T=1=*2'-guanylic acid complex at 1.9 angstroms resolution. Acta Crystallogr. B43, 549.

Baker, E. (1988) Structure of azurin from Alcigenes denitrificans. Refinement at 1.8 Å resolution and comparison of the two crystallographically independent molecules. J. Mol. Biol. 203, 1071-1095.

Beaudette, N. V., Langerman, N. & Kisliuk, R. L. (1980) A Calorimetric Study of the Binding of 2'-Deoxyuridine-5'-Phosphate and Its Analogs to Thymidylate Synthase. Arch. Biochem. and Biophys. 200(2), 410-417.

Belfort, M., Maley, G., Pederson-Lane, J. & Maley, F. (1983) Primary structure of the Eschericia coli thyA gene and its thymidylate synthase product. Proc. Natl. Acad. Sci. U.S.A. 80, 4914-4918.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. J., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) The Protein Data Bank. A computer-based archival file for macromolecular structures. J. Mol. Biol. 112, 535-542.

Beverley, S. M., Ellenberger, T. E. & Cordingley, J. S. (1986) Primary structure of the gene encoding the bifunctional dihydrofolate reductase-thymidylate synthase of Lesihmania major. Proc. Natl. Acad. Sci. U.S.A. 83, 2584-2588.

Birktoft, J., Rhodes, G. & Banaszak, L. (1989) Refined crystal structure of cytoplasmic malate dehydrogenase at 2.5 Å resolution. Biochemistry 28, 6065-6081.

Bode, W., Chen, Z. & Bartels, K. (1983) Refined 2 Å X-ray crystal structure of porcine pancreatic kallikrein A, a specific trypsin-like serine proteinase.

Crystallization, structure determination, crystallographic refinement, structure and its comparison with bovine trypsin. J. Mol. Biol. 164, 237-282.

Bolin, J., Filman, D., Matthews, D., Hamlin, R. & Kraut, J. (1982) Crystal structures of Escherchia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 Å resolution. J. Biol. Chem. 257(22), 13650-13662.

Bott, R. & Frane, J. (1990) Incorporation of crystallographic temperature factors in the statistical analysis of protein tertiary structures. Protein Engineering 3(8), 649-657.

Browner, M., Fauman, E. & Fletterick, R. (1992) Tracking conformational states in allosteric transitions of phosphorylase. Biochemistry 31, 11297-11304.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987) Crystallgraphic R Factor Refinement by Molecular Dynamics. Science 235, 458-460.

Brünger, A. T. (1989) Crystallographic Refinement by Simulated Annealing: Application to Crambin. Acta Cryst. A45, 50-61.

Brünger, A. T. (1990) Extention of Molecular Replacement - A New Search Strategy Based on Patterson Correlation Refinement. Acta Cryst. A. A46, 46-57.

Brünger, A. T. (1992) Free R-value - a novel statistical quantity for assessing the accuracy of crystal structures. Nature 335(6359), 472-475.

Bzik, D. J., Li, W.-B., Horii, T. & Inselburg, J. (1987) Molecular cloning and sequence analysis of the Plasmodium falciparum dihydrofolate reductase-thymidylate synthase gene. Proc. Natl. Acad. Sci. U.S.A. 84, 8360-8364.

Carreras, C. W., Climie, S. C. & Santi, D. V. (1992) Thymidylate Synthase with a C-terminal Deletion Catalyzes Partial Reactions but is Unable to Catalyze Thymidylate Formation. Biochemistry 31(26), 6038-6044.

Chambers, J. L. & Stroud, R. M. (1979) The accuracy of refined protein structures: Comparison of two independently refined models of bovine trypsin. Acta Cryst. B35, 1861-1874.

Chothia, C. & Lesk, A. M. (1986) The relationship between the divergence of sequence and structure in proteins. Embo J. 5(4), 823-826.

Chu, F. K., Maley, G. F., Maley, F. & Belfort, M. (1984) Intervening sequence in the thymidylate synthase gene of bacteriophage T4. Proc. Natl. Acad. Sci., U.S.A. 81, 3049-3053.

Cisneros, R. J. & Dunlap, R. B. (1990) Characterization of the parameters affecting covalent binding stoichiometry in binary and ternary complexes of thymidylate synthase. Biochim. Biophys. Acta 1039, 149-156.

Climie, S., Ruiz-Perez, L., Gonzalez-Pacanowska, D., Prapunwattana, P., Cho, S.-W., Stroud, R. & Santi, D. V. (1990) Saturation Site-directed Mutagenesis of Thymidylate Synthase. J. Biol. Chem. 265(31), 18776-18779.

Climie, S. C., Carreras, C. W. & Santi, D. V. (1992) Complete Replacement Set of Amino Acids at the C-terminus of Thymidylate Synthase: Quantitative Structure-Activity Relationships of Mutants of an Enzyme. Biochemistry 31(26), 6032-6038.

Clore, G. M. & Gronenborn, A. M. (1991) Comparison of the solution nuclear magnetic resonance and X-ray crystal structures of human recombinant interleukin-1 beta. J. Mol. Biol. 221(1), 47-53.

Cotton, F. A., Hazen, E. E. J. & Legg, M. J. (1979) Staphylococcal Nuclease. proposed mechanism of action based on structure of enzyme-thymidine 3',5'-biphosophate-calcium ion complex at 1.5-Ångstrom resolution. Proc. Nat. Acad. Sci. U.S.A. 76(6), 2551-2555.

Cruickshank, D. W. J. (1949) The accuracy of electron-density maps in X-ray analysis with special reference to dibenzyl. Acta Cryst. 2, 65-82.

Da, Y.-Z., Ito, K. & Fujiwara, H. (1992) Energy Aspects of Oil/Water Partition Leading to the Novel Hydrophobic Parameters for the Analysis of Quantitative Structure-Activity Relationships. J. Med. Chem. 35, 3382-3387.

Danenberg, P. V. & Danenberg, K. D. (1978) Effect of 5,10-methylenetetrahydrofolate on the dissociation of 5-fluoro-2'-deoxyuridylate from thymidylate synthase: evidence for an ordered mechanism. Biochemistry 17, 4018-4024.

Daron, H. H. & Aull, J. L. (1978) A Kinetic Study of Thymidylate Synthase from Lactobacillus casei. J. Biol. Chem. 253(3), 940-945.

Deisenhofer, J. (1981) Crystallographic refinement and atomic models of a human Fc fragment and it complex with fragment B of protein A from Staphyloccus aureus at 2.9- and 2.8-Å resolution. Biochemistry 20(9), 2361-2370.

Dev, I. K., Yates, B. B., Leong, J. & Dallas, W. S. (1988) Functional role of cysteine-146 in Escherichia coli thymidylate synthase. Proc. Natl. Acad. Sci. U.S.A. 85, 1472-1476.

Dev, I. K., Yates, B. B., Atashi, J. & Dallas, W. S. (1989) Catalytic Role of Histidine 147 in Escherichia coli Thymidylate Synthase. J. Biol. Chem. 264(32), 19132-19137.

Devereux, J., Haeberli, P. & Smithies, O. (1984) A Comprehensive Set of Sequence Analysis Programs for the VAX. Nucleic Acids Res. 12(1), 387-395.

Dill, K. A., Alonso, D. O. V. & Hutchison, K. (1989) Thermal Stabilities of Globular Proteins. Biochemistry 28, 5439-5449.

Earnest, T., Fauman, E., Craik, C. S. & Stroud, R. (1991) 1.59 Å structure of trypsin at 120 K: Comparison of low termperature and room temperature structures. Proteins: Structure, Function and Genetics 10, 171-187.

Edman, U., Edman, J. C., Lundgren, B. & Santi, D. V. (1989) Isolation and expression of the Pneumocystis carinii thymidylate synthase gene. Proc. Natl. Acad. Sci., U.S.A. 86(17), 6503-6507.

Epp, O., Ladenstein, R. & Wender, A. (1983) The refined structure of the selenoenzyme glutathione peroxidase at 0.2-nm resolution. Eur. J. Biochem. 133, 51-69.

Fermi, G., Perutz, M. & Shaanan, B. (1984) The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. J. Mol. Biol. 175, 159-174.

Finer-Moore, J. S., Montfort, W. R. & Stroud, R. M. (1990) Pairwise Specificity and Sequential Binding in Enzyme Catalysis: Thymidylate Synthase. Biochemistry 29(30), 6977-6986.

Finzel, B., Weber, P., Hardman, K. & Salemme, F. (1985) Structure of ferricytochrom c' from Rhodospirillum at 1.67 Å resolution. J. Mol. Bio. 186, 627-643.

Fita, I. & Rossman, M. G. (1985) The NADPH binding site on beef liver catalase. Proc. Nat. Acad. Sci., U.S.A. 82(6), 1604-1608.

Fox, G. C. & Holmes, K. C. (1966) An Alternative Method of Solving the Layer Scaling Equations of Hamilton, Rollett and Sparks. Acta Crystallogr. A20, 886-891.

French, S. & Wilson, K. (1978) On the Treatment of Negative Intensity Observations. Acta Crystallogr. A34, 517-525.

Galivan, J. H., Maley, G. F. & F., M. (1976) Factors Affecting Substrate Binding in Lactobacillus casei Thymidylate Synthetase as Studied by Equilibrium Dialysis. Biochemistry 15(2), 356-362.

Grumont, R., Washtein, W. L. & Santi, D. V. (1986) Bifunctional thymidylate synthase-dihydrofolate reductase from Leishmania tropica: Sequence homology with the corresponding monofunctional proteins. Proc. Natl. Acad. Sci., U.S.A. 83, 5387-5391.

Hansch, C. H. & Leo, A. (1979) Substituent constants for correlation analysis in chemistry and biology, Wiley, New York.

Hardy, L. W., Finer-Moore, J. S., Montfort, W. R., Jones, M. O., Santi, D. V. & Stroud, R. M. (1987) Atomic Structure of thymidylate synthase: Target for rational drug design. Science 235, 448-455.

Hendrickson, W. A. & Konnert, J. H. (1981) in Biomolecular Structure, Conformation, Function and Evolution, (Srinivasan, R., Subramanian, E. & Yathindra, N., Ed.) 43-57, Pergamon, Oxford.

Honess, R. W., Bodemer, W., Cameron, K. R., Niller, H.-H., Fleckenstein, B. & Randallm, R. E. (1986) The A+T-rich genome of herpesvirus saimiri contains a highly conserved gene for thymidylate synthase. Proc. Natl. Acad. Sci., U.S.A. 83, 3604-3608.

Howard, A. J., Neilsen, C. & Xuong, N. H. (1985) Software for a diffractometer with multiwire area detector. Methods Enzymol. 114, 452-472.

Hughes, D. E., Shonekan, O. A. & Simpson, L. (1989) Structure, genomic organization and transcription of the bifunctional dihydrofolate reductase-thymidylate synthase gene from Crithidia fasiculata. Mol. Biochem. Parasit. 34, 155-166.

Hyde, C. C. & Miles, E. W. (1990) The tryptophan synthase multienzyme complex. Exploring structure-function relationships with X-ray crystallography and mutagenesis. Bio-technology 8(1), 27-32.

Iwakura, M., Dawata, M., Tsuda, K. & Tanaka, T. (1988) Nucleotide sequence of the thymidylate synthase B and dihydrofolate reductase genes contained in one Bacillus subtilis operon. Gene 64, 9-20.

Jones, T. A. (1985) Interactive computer graphics: FRODO. Methods Enzymol. 115, 157-171.

Kabsch, W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. Acta Cryst. A34, 827-828.

Kamb, A., Finer-Moore, J., Calvert, A. H. & Stroud, R. M. (1992) Structural Basis for Recognition of Polyglutamyl Folates by Thymidylate Synthase. Biochemistry 31, 9883-9890.

Ke, H.-M., Hozatko, R. & Lipscomb, W. (1984) Structure of unligated aspartate carbamoyltransferase of Escherichia coli at 2.6 Å resolution. Proc. Natl. Acad. Sci., U.S.A. 81, 4037-4040.

Kenny, E., Atkinson, T. & Hartley, B. S. (1985) Nucleotide sequence of the thymidylate synthase gene (ThyP3) from the Bacillus subtilis phage f3T. Gene 34, 335-342.

Kossiakoff, A. A., Randal, M., Guenot, J. & Eigenbrot, C. (1992) Variability of conformations at crystal contacts in BPTI represent true low-energy structures - correspondence among lattice packing and molecular

dynamics structures. Proteins:Structure, Function and Genetics 14(1), 65-74.

Kraulis, P. J. (1991) MolScript. J. Appl. Cryst. 24, 946.

Kunitani, M. G. & Santi, D. V. (1980) On the Mechanism of 2'-Deoxyuridylate Hydroxymethylase. Biochemistry 19, 1271-1275.

Lamm, N., Tomaschewski, J. & Ruger, W. (1987) Nucleotide sequence of the deoxycytidylate hydroxymethlase gene of bacteriophage T4 (g42) and the homology of its gene product with thymidylate synthase of E. coli. Nucleic Acids Res. 15(9), 3920.

Luzzati, V. (1952) The statistical treatment of errors in crystal structures. Acta Cryst. 5, 802-810.

Maley, G. & Maley, F. (1988) Properties of a defined mutant of Eschericia coli Thymidylate Synthase. J. Biol. Chem. 263(16), 7620-7627.

Matthews, D. A., Appelt, K., Oatley, S. J. & Xuong, N. H. (1990) Crystal Structure of Escherichia coli Thymidylate Synthase Containing Bound 5-Fluoro-2'-deoxyuridylate and 10-propargyl-5,8-dideazafolate. J. Mol. Biol. 214, 923-936.

Matthews, D. A., Villafranca, J. E., Janson, C. A., Smith, W. W., Welsh, K. & Freer, S. (1990) Stereochemical Mechanism of Action for Thymidylate Synthase Based on the X-ray Structure of the Covalent Inhibitory Ternary Complex with 5-Fluoro-2'-deoxyuridylate and 5,10-Methylenetetrahydrofolate. J. Mol. Biol. 214, 937-948.

Michaels, M. L., Kim, C. W., Matthews, D. A. & Miller, J. H. (1990) Escherichia coli thymidylate synthase: Amino acid substitutions by suppression of amber nonsense mutations. Proc. Natl. Acad. Sci. U.S.A. 87, 3957-3961.

Mittelstaedt, D. M. & Schimerlik, M. I. (1986) Stopped-flow studies of 2'-deoxynucleotide binding to thymidylate synthase. Arch. Biochem. Biophys. 245, 417-425.

Montfort, W. R., Perry, K. M., Fauman, E. B., Finer-Moore, J. S., Maley, G. F., Hardy, L., Maley, F. & Stroud, R. M. (1990) Structure, Multiple Site Binding, and Segmental Accommodation in Thymidylate Synthase on Binding dUMP and an Anti-Folate. Biochemistry 29(30), 6964-6977.

Montfort, W. R., Fauman, E. B., Perry, K. M. & Stroud, R. M. (1990) Segmental Accommodation: A Novel Conformational Change Induced Upon Ligand Binding by Thymidylate Synthase, in Current Research in Protein Chemistry: Techniques, Structure and Function, (Villafranca, J. J., Ed.) 367-382, Harcourt Brace Jovanovich, San Diego.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992) Stereochemical quality of protein structure coordinates. Proteins 12, 345-364.

Nelson, K., Alonso, G., Langer, P. J. & Beverly, S. M. (1990) Sequence of the dihydrofolate reductase-thymidylate synthase (DHFR-TS) gene of Leishmania amazonensis. Nucleic Acids Res. 18(9), 2819.

Norris, G., Anderson, B. & Baker, E. (1983) Structure of azurin from Alcigenes denitrificans at 2.5 Å resolution. J. Mol. Biol. 165, 501-521.

Otwinowski, Z. (1986) DENZO manual, University of Chicago, Chicago.

Padlan, E. & Love, W. (1985) Refined crystal structure of deoxy-hemoglobin S I. Restrained least squares refinement at 3.0-Å resolution. J. Biol. Chem. 260(14), 8272-8279.

Perry, K. M., Fauman, E. B., Finer-Moore, J. S., Montfort, W. R., Maley, G. F., Maley, F. & Stroud, R. M. (1990) Plastic Adaptation Toward Mutation in

Proteins: Structural Comparison of Thymidylate Synthases. Proteins-Struct. Funct. Genet. 8, 315-333.

Perryman, S. M., Rossana, C., Deng, T., Vanin, E. F. & Johnson, L. F. (1986) Sequence of a cDNA for mouse thymidylate synthase reveals striking similarity with the prokaryotic enzyme. Mol. Biol. Evol. 3, 313-321.

Pogolotti, A. L. J. & Santi, D. V. (1977) The Catalytic Mechanism of Thymidylate Synthase, in Bioorganic Chemistry, Ed.) 277-311, Academic Press, New York.

Ponder, J. W. & Richards, F. M. (1987) Tertiary templates for proteins. J. Mol. Biol. 193, 775-791.

Quiocho, F. A., Wilson, D. K. & Vyas, N. K. (1989) Substrate specificity and affinity of a protein modulated by bound water molecules. Nature 340, 404-407.

Rashin, A. A., Iofin, M. & Honig, B. (1986) Internal Cavities and Buried Waters in Globular Proteins. Biochemistry 25, 3619-3625.

Remington, S., Wiegand, G. & Huber, R. (1982) Crystallographic refinement and atomic models of two different forms of citrate synthase at 2.7 and 1.7 angstroms resolution. J. Mol. Biol. 158(1), 111-152.

Richter, J., Puchtler, I. & Fleckenstein, B. (1988) Thymidylate synthase gene of herpesvirus ateles. J. Virol. 62, 3530-3535.

Ross, P., O'Gara, F. & Condon, S. (1990) Cloning and characterization of the thymidylate synthase gene from Lactococcus lactis subsp. lactis. Appl. Env. Micro. 56(7), 2156-2163.

Rouch, D. A., Messorotti, L. J., Loo, L. S. L., Jackson, C. A. & Skurry, R. A. (1989) Trimethoprim resistance transposon Tn4003 from Staphylococcus aureus encodes genes for a dihydrofolate reductase and

thymidylate synthetase flanked by three copies of IS257. Mol. Microbiol. 3(2), 161-175.

Rypniewski, W. & Evans, P. (1989) Crystal structure of unliganded phosphofructokinase from Escherichia coli. J. Mol. Biol. 207, 805-821.

Santi, D. V. & Danenbergm, P. V. (1984) Folates in Pyrimidine Nucleotide Biosynthesis, in Chemistry and Biochemistry and Folates, (Blakely, R. L. & Benkovic, S. J., Ed.) 345-399, John WIley & Sons, New York.

Santi, D. V., McHenry, C. S., Raines, R. T. & Ivanetish, K. M. (1987) Kinetics and Thermodynamics of the Interaction of 5-Fluoro-2'-deoxyuridylate with Thymidylate Synthase. Biochemistry 26, 8606-8613.

Santi, D. V., Pinter, K., Kealy, J. & Davisson, V. J. (1990) Site-directed mutagenesis of arginine 179 of thymidylate synthase. A nonessential substrate-binding residue. J. Biol. Chem. 265(12), 6770-6775.

Schiffer, C. A., Davisson, V. J., Santi, D. V. & Stroud, R. M. (1991) Crystallization of Human Thymidylate Synthase. J. Mol. Bio. 219(2), 161-163.

Schreuder, H. A., van der Laan, J. M., Hol, W. G. J. & Drenth, J. (1988) Crystal structure of p-hydroxybenzoate hydroxylase complexed with its reaction product 3,4-dihydroxybenzoate. J. Mol. Biol. 199(4), 637-648.

Shirakihara, Y. & Evans, P. R. (1988) Crystal structure of the complex of phosphofructokinase from Escherichia coli with its reaction products. J. Mol. Biol. 204(4), 974-994.

Singer, S. C., Richards, C. A., Ferone, R., Benedict, D. & Ray, P. (1989) Cloning, purification, and properties of Candida albicans thymidylate synthase. J. Bacteriol. 171, 1372-1378.

Skarzynski, T., Moody, P. & Wonacott, A. (1987) Structure of holo-glyceraldehyde-3-phosphate dehydrogenase from Bacillus stearothermophilus at 1.8 Å resolution. J. Mol. Biol. 193, 171-187.

Smith, W. W., Burnett, R. M., Darling, G. D. & Ludwig, M. L. (1977) Structure of the semiquinone form of flavodoxin from clostridium mp. Extension of 1.8 Angstroms resolution and some comparisons with the oxidized state. J. Mol. Biol. 117(1), 195-225.

Strenkamp, R., Siker, L. & Jensen, L. (1982) Restrained least-squares refinement of Themiste dyscritum methydroxohemerythrin at 2.0 Å resolution. Acta Cryst. B38, 784-792.

Takano, T. & Dickerson, R. (1980) Redox conformation changes in refined tuna cytochrome c. Proc. Natl. Acad. Sci., U.S.A. 77(11), 6371-6375.

Takeishi, K., Kaneda, S., Ayusawa, D., Shimizu, K., Gotoh, O. & Seno, T. (1985) Nucleotide sequence of a functional cDNA for thymidylate synthase. Nucleic Acids Res. 13, 2035-2043.

Taylor, G. R., Lagosky, P. A., Storms, R. K. & Haynes, R. H. (1987) Molecular characterization of the cell cycle-regulated thymidylate synthase gene of Saccharomyces cerevisiae. J. Biol. Chem. 262, 5298-5307.

Thompson, R., Honess, R. W., Taylor, L., Morran, J. & Davison, A. J. (1987) Varicella-zoster virus specifies a thymidylate synthase. J. Gen. Virol. 68, 1449-1455.

Thylen, C. (1988) Expression and DNA Sequence of the Cloned Bacteriophage T4 dCMP Hydroxymethylase Gene. J. Bacter. 170(4), 1994-1998.

Tilton, R. F. J., Dewan, J. C. & Petsko, G. A. (1992) Effects of temperature on protein structure and dynamics: X-ray crystallographic studies of the protein ribonuclease-A at nine different temperatures from 98 to 320 K. Biochemistry 31(9), 2469-2481.

Tsukada, H. & Blow, D. (1985) Structure of a-chymotrypsin refined at 1.68 Å resolution. J. Mol. Biol. 184, 703-711.

Vriend, G., Berendsen, H. J. C., van der Zee, J. R., van den Burg, B., Venema, G. & Eijsink, V. G. H. (1991) Stabilization of the neutral protease of Bacillus stearothermophilus by removal of a buried water molecule. Protein Eng. 4(8), 941-945.

Waller, D. A. & Liddington, R. C. (1990) Refinement of a partially oxugenated T state haemoglobin at 1.5 Angstroms resolution. Acta Crystallogr B46, 409-418.

Weber, I. T. & Steitz, T. A. (1987) Structure of a complex of cataboilite gene activator protein and cycle AMP refined at 2.5 angstroms resolution. J. Mol. Biol. 198(2), 311-326.

Wilhelm, K. & Ruger, W. (1992) Deoxyuridylate-Hydroxymethylase of Bacteriophage SPO1. Virology 189, 640-646.

Wilson, A. J. C. (1949) The Probability Distribution of X-ray Intensities. Acta Crystallogr. 2, 318-321.

Xia, Z. & Mathews, F. S. (1990) Molecular structure of flavocytochrome B2 at 2.4 Å resolution. J. Mol. Biol. 212(4), 837-863.

Appendix 1. NewDome

Introduction to NewDome

The purpose of NewDome is to select a set of carbon alphas to be used for overlapping a pair of structures. The input is two PDB files and the output is a list a residue numbers. NewDome assumes you already know the sequence alignment of one structure to the other. It simply prunes the list of alpha carbon pairs to pick a "core." NewDome also does not perform the actual superpositioning. This can be done by the program GEM (Appendix 2, this thesis).

The program internally constructs a difference distance matrix and then selects the largest subset of alpha carbons which fulfills the following two criteria: 1) each carbon alpha in the subset moves less than a fixed amount (by default 0.5 angstroms) from every other carbon alpha in the subset 2) the subset is "connected" in a mathematical sense. That is, a path can be drawn from every carbon alpha in the subset to every other carbon alpha in the subset by hopping through other carbon alphas in the subset, no single hop exceeding (by default) 10 angstroms.

NewDome is written in C for a VMS environment. The entire program is written in standard C except for the second fopen() which opens a text file for writing. This line will have to be modified for non-VMS systems.

Input for NewDome

When you run NewDome you will be asked for the following input:

1) file #1

2) file #2

These are both "minimal" PDB files, meaning they contain only ATOM records. In addition, residue numbers should be numbers only, for example "65A" does not work. Finally, neither file should have an insertion relative to the other. Thus NewDome assumes the first residue read from file #1 corresponds to the first residue read from file #2 and so on. If the PDB do need to be edited, the program GEM (Appendix 2 of this thesis) can be used to select and write out the desired CA records.

3) output

This is the output file. By default, it takes the suffix '.domains'

4) residue number range, file 1

5) residue number range, file 2

You select a range of residues to be used for file 1 and file 2. You can select a different range for each file, but the results will be reported in terms of residue numbers in file 1 only.

6) number of domains to find

NewDome will report the top N domains it finds. The default is 10. For most purposes, only the first domain is used however.

7) minimum number of residues per domain

NewDome will only report domains with greater than X residues. The default is 10. NewDome may report fewer than N domains (see above) if the domains it finds have fewer than X residues.

8) radius of contact

Each carbon alpha in a domain is within R angstroms of another carbon alpha in the domain. The default is 10 angstroms. This features ensures that the domains reported are "connected." See the section describing the algorithm below.

9) maximum movement in the core

Every carbon alpha in the core moves less than M angstroms relative to every other carbon alpha in the core. The default for M is 0.5 angstroms. This value can be raised or lowered to obtain larger or smaller domains.

10) output in GEM format (Y/n)

Selecting Y produces a list of residue numbers suitable for input into GEM. In fact the output will be a complete command file for GEM which will read in the named PDB files, select the core alpha carbons, calculate the transformation matrix, apply the matrix (rotating the first coordinate set onto the second) and write out the transformed coordinates into a file called ROTATED.PDB.

Selecting N produces a list of residue numbers in a slightly more readable format. The output will look like this:

1-4,8,13-17

which means use residues 1 through 4, residue 8, and residues 13 to 17.

Example header information from output of NewDome

File 1:ts:kdime.pdb First: 1. Last: 564. Count: 528

File 2:ts:bdime.pdb First: 1. Last: 564. Count: 528

Residue numbers derived from file 1 <- stats on diff dist matrix

Residues 323 and 562 move 3.232 A nearer <- largest negative value

Residues 264 and 564 move 12.334 A farther <- largest positive value

Average movement = 0.113 A <- average diff dist matrix val

RMS movement = 0.720 A <- rms diff dist matrix value

Radius of contact = 10.00 A <- user-entered values

Max allowable movement = 0.50 A

Core specific stats: <- statistics on subset #1

Residues 151 and 411 move 0.498 A nearer

Residues 99 and 513 move 0.498 A farther

Average movement = -0.006

RMS movement = 0.182

Algorithm employed in NewDome

The goal of NewDome is to find the largest subset of alpha carbons in the two structures which fulfills the two criteria: 1) the difference distance for every pair of atoms in the subset is less than some threshold, which by default of 0.5 Å, and the subset is connected with a path between every pair of atoms each step of which is smaller than the "radius of contact," which by default is 10 Å.

Both these cutoffs were determined empirically. A difference distance threshold of 0.5 Å seems to ensure that the core atoms when overlapped will have an RMS deviation of around 0.5 Å or less, about what would be expected from errors alone for medium resolution structures (around 2.5 Å). The 10 Å "radius of contact" seemed to me to be the farthest apart two alpha carbons could be and still be interacting directly through their side-chains.

The algorithm I developed for identifying such a subset seems to work reliably for the conditions found in proteins; however, I have not proven that it will always find the largest subset. In fact, one can construct "pathologic" cases where my algorithm in fact fails.

My algorithm assumes that some atom will be the center of the core. We can begin by assuming that alpha carbon #1 is the center of the core. At first, this alpha carbon is the only atom in the core. Then we go through the following steps.

1. Identify all atoms which are both within 10 Å of the current core and which move less than 0.5 Å with respect to all atoms in the current core. These atoms are potential recruits to be added to the core. If no potential recruits can be found, the core is done.
2. If any pair of recruited atoms have a difference distance value of greater than 0.5 Å, the atom with the greater average difference distance to the current core is labeled unacceptable, but remains in the set of recruited atoms since it may still be the basis of excluding yet another recruited atom.
3. All recruited atoms which remain acceptable are added to the core.
4. Go back to step 1 with the new larger core.

Once the core stops growing, the size of the core (the number of residues in the core) is noted. Then the process is begun again with the next alpha carbon in the structure. Once each alpha carbon has been used as the start of a core, the largest core is reported.

A secondary core is defined as that core which has the most alpha carbons which are not found in the first core. Note this need not be the second largest core, since that core may overlap substantially with the first core identified.

An important point in this algorithm is that the distance between atoms is that in the first structure. Thus, if this distance is greater than 10 Å in one structure, but less than 10 Å in the other, the cores identified may be different depending on which structure is listed first in the input to NewDome.

Appendix 2. GEM

OUTLINE

I. Introduction to GEM

II. Running GEM

1. Starting
2. Entering commands
3. Getting information
4. Exiting

III. Reading and writing files

1. IN
2. OUT
3. WRITE
4. DUMP

IV. Atom Selection commands

1. ATOM
2. RESIDUE
3. CHAIN
4. RANGE
5. OCCUPANCY
6. BFACTOR
7. AVGB
8. SEGID

V. Record modification commands

1. NEWCHAIN
2. NEWRANGE
3. NEWOCC

4. NEWBFACTOR
5. NEWSEGID
6. APPLY

VI. Functions using 1 PDB file

1. LIST SEQ
2. LIST ATOMS
3. LIST RES
4. RGYR
5. AXES
6. BCOUNT
7. BSEQ
8. SORTPDB

VII. Functions using 2 PDB files

1. PRIMARY/SECONDARY
2. LIST DIFF
3. RSEQ
4. PDBRMS
5. SUPER
6. AXES
7. BVSB
8. BSEQ
9. BSIGMA

VIII. Miscellaneous commands

1. DO
2. GO
3. SET
4. WINDOW

Appendix 2-A. PDB file fields

Appendix 2-B. Glossary of GEM commands and functions

I. Introduction to GEM

GEM is a collection of routines for analyzing, comparing and manipulating PDB files. Most of the functions in GEM are relatively simple calculations, but by being included in GEM are made much more flexible due to the extensive atom selection commands. Some of the functions are more sophisticated, but by being included in GEM are made much more accessible and easy to use. A glance through the Table of Contents should give you an idea of the scope of GEM. GEM was written by Eric Fauman and grew and evolved over the years 1988-1993.

II. Running Gem

1. Starting

GEM lives in deq:[fauman.gem]. To run GEM, type "run deq:[fauman.gem]gem".

If you put a statement like:

```
$gem :== "$deq:[fauman.gem]gem"
```

in your login.com, then you can run GEM at any time simply by typing the command GEM. Moreover, if you've done this you can use command line arguments. Specifically, you can type something like:

```
$GEM file1 file2
```

where file1.pdb and file2.pdb are two PDB files and GEM will begin by reading in those files. Note that the default extension .pdb is not required.

2. Entering commands

GEM is designed to be run interactively. When running GEM, you enter COMMANDS, which alter various parameters and set up the calculations, and FUNCTIONS, which actually perform the calculations. GEM sends responses to COMMANDS to the screen, while the results of calculations from FUNCTIONS go both to the screen and to the current output file (see the OUT command). Typically more information is sent to the output file than to the screen. Input to GEM can be in upper case or lower case. All arguments to commands and functions will be converted to upper case unless they are enclosed in quotes; for example: segid "a" GEM can accept commands from a file. Typing "@gem-in", for example, will cause GEM to look to the file GEM-IN.COM for input. Control will return to the user when the end of the command file is reached. Note that a default extension of .com is added if no extension is specified. This redirection is only one level deep; that is, a command file cannot use the "@" command to redirect input from another command file.

3. Getting information

GEM has two kinds of user information. Typing "NEWS" will display a brief summary of the latest enhancements to GEM. Typing "?" or "HELP" will list all the commands and functions known to GEM. Typing "?" or "HELP" followed by one of these keywords will display a few lines of information about how to use that command.

4. Exiting

The commands QUIT, EXIT, END, STOP and BYE get you out of GEM. Also, 20 blank lines in a row exit from GEM. This is necessary if GEM is used in a batch job and no EXIT command has been supplied.

III. Commands for Reading and writing files

1. IN

The IN command reads in PDB files. A detailed description of a PDB file is given in Appendix A. The command "IN file1 file2" reads the coordinates from the files file1.pdb and file2.pdb and puts them in the first two coordinate slots. GEM has room for up to 4 coordinate sets of up to 10,000 atoms apiece. Typing "IN" alone will show which files have been read in so far. The command "IN RESET" will erase the coordinates in memory to allow new files to be read in.

2. OUT

All the functions in GEM produce some information which is sent to an output file. If no file is specified, GEM will open a default file called "GEM.LOG". To redirect output to a different file, use the command "OUT outfile". All the successive output will be sent to the file OUTFILE until a new "OUT" or "WRITE <filename>" command is given.

3. WRITE

The WRITE function writes out a PDB file of the currently selected atoms (See section IV) from the PRIMARY coordinate set (See section VII). The WRITE function by itself sends the output records to the current output file. "WRITE" followed by a filename will create a new file with the specified name. To see what coordinates you will be saving, use the DUMP function. WRITE does not invoke the Record Modification commands (section VI). To modify the coordinate records you must use the APPLY function (section VI).

4. DUMP

The DUMP function is like the WRITE function, but the output goes to the screen. It is very useful for examining which atoms have been selected by the current selection criteria. The listing will pause every 20 lines and you

can scroll forwards or backwards through the listing, 20 lines at a time. DUMP with no argument lists all selected atoms. DUMP followed by a number lists only that many atoms. For example, "DUMP 10" lists the first ten selected atoms.

IV. Atom Selection Commands

Often, a calculation or operation is desired for only a subset of the atoms in a PDB file, for example if you want to overlap two structures using only the alpha carbons, or if you want to calculate the average B factor for arginine side chains. GEM has an extensize and flexible system for specifying a subset selection of atoms. Nearly every field of the PDB file can be used for selection purposes (see Appendix 2-A for a full description of the fields in a PDB file). The commands for selection based on the various fields are ATOM, RESIDUE, CHAIN, RANGE, OCCUPANCY, BFACTOR, AVGB and SEGID. To specify the selection, use the appropriate command followed by the selection critera as detailed below.

Some of these commands (ATOM, RESIDUE, RANGE) allow multiple criteria. When this is the case, the logic within a given field is a logical OR. However the logic between different fields is a logical AND. Thus only atoms that match criteria for all selected fields will be selected. To check the selection criteria for a specific field, just use the command by itself. For example, typing "SEGID" will show the current Segment ID selection criteria. The default is to include all atoms. For example, if no BFACTOR command has been issued, the atomic B factor will be ignored in deciding which atoms to include. To check which fields are being scrutinized for selection purposes use the REVIEW command. The REVIEW command will also show how many atoms have been selected by the selection criteria. To turn off selection

based on a specific field, use the RESET command with the selection command. For example, "SEGID RESET" will return SEGID to the default, non-selecting, condition. The RESET command can be used to reset several selection criteria at once. For example, "RESET ATOM BFACTOR" will turn off selection based on atom type or atomic B factor.

Following is a detailed description of each of these fields.

1. ATOM - Atom name

The GEM command ATOM allows you to select or deselect specific atoms based on the atom name. The following are some examples of the ATOM command.

```
ATOM CA          ! selects only alpha carbons
ATOM CA N C O    ! selects all main chain atoms
ATOM -CB         ! deselects beta carbons
ATOM H*          ! selects all hydrogens
ATOM N* -N       ! selects all nitrogens except the main chain nitrogen
ATOM *E*         ! selects all epsilon atoms
ATOM N:O*        ! selects all asperigine oxygens
ATOM CA -GLY:CA      ! selects all alpha carbons, except those of glycines
```

As illustrated, an atom type can be included by typing "ATOM" followed by the desired atom type. Alternatively, an atom can be excluded by preceeding the atom name with a minus sign. Multiple atom selection criteria can be entered on a single line. Additional ATOM commands will add to the current criteria. To enter a new set of selection criteria, you must first use the

"ATOM RESET" command. The wildcard character '*' can be used. A '*' in the middle of a word matches only one character. A '*' at the end of a word matches any number of characters. Specific ATOMS of specific residues may be selected by preceeding the atom name with the desired residue name and a colon (':'). As discussed in the next section, either 1 or 3 letter amino acid codes are acceptible.

2. RESIDUE - Residue Type

The RESIDUE command allows selection or deselection of specific residue types.

The following are some examples of the RESIDUE command:

```
RES LEU          ! selects leucines
RES K ARG H      ! selects all basic residues
RES -D -E        ! selects all non-acidic residues
RES T*           ! selects residues with names beginning with T
                 !(tyr and thr)
RES AS*          ! selects asp and asn
RES G* -G        ! selects glu and gln
```

As shown above, a specific residue is selected by typing "RESIDUE" (or "RES" for short) followed by the name of the residue. For amino acids, the standard 1-letter amino acid codes can be used. As with the ATOM command, a specific type can be excluded by preceeding it with a minus sign. Also, the wild-card character, '*' can be used. When used at the end of a word, the '*' matches any number of characters.

3. CHAIN - Chain Identifier

The Chain command allows selection of a specific chain from the PDB file. For example:

```
CHAIN A
CHAIN "b"
```

The first example selects all the atoms with a Chain ID of A. Gem automatically converts all input commands into uppercase. If you wanted to specify a Chain ID which was a lower case letter, surround the Chain ID specification in quotes, as illustrated above.

4. RANGE - Residue Number

The Range command allows selection of specific ranges of residue numbers. Some examples of the RANGE command:

```
RANGE 1 10                ! selects residues 1 through 10, inclusive
RANGE 20 24 26 30     ! selects resides 20 to 24 and 26 to 30
RANGE @myrange.dat    ! reads range selections from the file myrange.dat
```

Up to 100 different ranges can be selected as illustrated above. Subsequent RANGE commands add to the ranges already selected. To clear the current RANGE selections, type "RESET RANGE". The Range command can accept input from an external file. This file must contain the ranges as two columns of numbers, one pair per line. For example:

```
----------sample range file
```

```
20 24
26 30
----------sample range file
```

5. OCC - Occupancy

One range of allowed occupancy values can be specified with the OCC command. For example:

```
OCC 0.9 1.0
```

selects all atoms with an occupancy value between 0.9 and 1.0. The order of the numbers doesn't matter. If only one number is entered, the occupancy must match that value exactly. Subsequent OCC commands override previous commands.

6. BFACTOR - B Factor

As with the OCC command, the B factor command allows selection of atoms based on their atomic B factors. For example:

```
BFACTOR 0 40
```

selects all atoms with a reasonable B factor. Subsequent BFACTOR commands override previous commands.

7. SEGID - Segment Identifier

This command allows selection of atoms based on their segment ID. The following are some examples:

SEGID WAT

SEGID A*

SEGID "tiny"

SEGID WAT A* "tiny" -APP

SEGID RESET PROT

As with the ATOM and RES commands, the wild card character '*' may be used. As with the CHAIN command, if you want to specify a Segid with lowercase letters, surround the Segid with quotes. Be advised, however, that X-PLOR expects only uppercase letters in the Segid.

Multiple Segids can be selected at once, as in the 4th example above.

V. Record modification commands

Gem can change the values of fields in the PDB file. This is used, for example, in changing the residue numbering of a PDB file, adding or removing SEGIDs, assigning occupancies, creating a dimer from monomer coordinates, and so on. Typically, you will read in your PDB file, select which atoms to modify (see section IV), specify the changes to make, and then APPLY these changes. The APPLY function makes the changes in an internal coordinate set. To make a PDB file with the desired changes you must first APPLY the changes, and then WRITE a PDB file (see section III).

The REVIEW command will show which Record Modification commands are in effect. A single modification command can be reviewed by typing that command alone, with no arguments. To remove a Record Modification instruction, use the RESET command, as in "MATRIX RESET"

or "RESET NEWCHAIN". The RESET command can operate on several commands at once, as in

"RESET NEWSEGID NEWRANGE."

1. NEWCHAIN

The NEWCHAIN command is used to add, change or remove the Chain Identifier field. For example:

```
NEWCHAIN A
NEWCHAIN "a"
NEWCHAIN " "
```

The first example will set the Chain Identifier field of the selected atoms to "A" once the APPLY command is used. Since GEM converts all input to uppercase, if you want a lowercase letter for a Chain ID, you must surround it in quotes, as in the second example.

To clear the Chain ID, the third example is required: a space enclosed in quotes.

2. NEWRANGE

The NEWRANGE command is used to renumber the residues of the selected atoms. The residues can be numbered sequentially starting at some arbitrary number, or can be shifted by a fixed amount from their current values. For example:

```
NEWRANGE   1 SEQ
NEWRANGE 101 NOSEQ
```

The first example will renumber the residues starting at 1 following the APPLY function. All Insertion Codes (see Appendix 2-A) will be lost. In the second example, residue number 1 will be changed to residue number 101 and so on; that is, a shift of +100 to all residue numbers. Insertion Codes present in the selected residues will remain intact; residue 53B will become residue 153B.

3. NEWOCC

The NEWOCC command allows you to assign values to the occupancy field, for example, to set all the occupancies to 1 or to the number of electrons for each atom type. Examples:

NEWOCC 7.0
NEWOCC DELTA -0.5

With the APPLY function, the first example will set the occupancy of all selected atoms to 7.0. The second example will cause 0.5 to be subtracted from the occupancies of all selected atoms.

4. NEWBFACTOR

NEWBFACTOR is used to assign values to the B-factor field. For example:

NEWBFACTOR 15.0
NEWBFACTOR DELTA 10

The first NEWBFACTOR command will cause APPLY to set the B factor to 15 for all selected atoms. The second example directs APPLY to add 10.0 to the B factor for all selected atoms.

5. NEWSEGID

NEWSEGID allows you to add, delete or modify SEGIDs. For example:

```
NEWSEGID MAIN
NEWSEGID "solv"
NEWSEGID " "
```

The first two examples direct APPLY to change the SEGID field to MAIN and solv respectively for the selected atoms. Since GEM changes all input to upper case, if you want lower case letters in the segid you must surround the argument with quotes. The last example will cause the SEGID field to be cleared following the APPLY function.

6. MATRIX

The MATRIX function allows you to enter a rotation/translation matrix to be applied to the coordinates. The transformation is A' = rA + t, where A is the current coordinates, A' is the transformed coordinates, r is a 3x3 rotation matrix and t is a translation vector. If you type "MATRIX" the current rotation/translation matrix will be displayed. To enter a new matrix, tyoe "MATRIX ENTER." You will then be prompted for each line of the matrix. Often used matrices can be stored in a file such as:

```
----------sample matrix file-------
! A comment in the matrix file
! r matrix       translation
 0  1  0       12.5
 1  0  0      -12.5
 0  0 -1        0.0
! end of matrix
----------sample matrix file-------
```

To read in a matrix file, use the command "MATRIX @filename." Once a matrix has been entered, GEM will display the total rotation and total translation generated from the matrix. A rotation matrix must have a determinant of 1.0, and GEM checks to make sure this is true. If it displays the message: "Warning: Determinant of matrix is x!" where x is some number other than 1, then there is a mistake in matrix you entered. Double check it and re-enter it. As with the other record modification commands, the MATRIX command won't have an effect on the coordinates until the APPLY function is used.

7. APPLY

Apply is a FUNCTION, meaning it will affect the coordinates (as opposed to the COMMANDS which have no effect on the coordinates directly). APPLY generates new coordinate records from those in the PRIMARY coordinate set based on the current atom selection criteria (section V) and record modification commands. To check which atoms will be affected and what changes have been specified, use the REVIEW command. Only those commands which are not set to their default values will be displayed. The APPLY command has several options:

APPLY

APPLY REPLACE {n}

APPLY APPEND {n}

APPLY MODIFY

The first example stores the new coordinate records in an unused coordinate slot. That is, if you have two files read in, the modified records will be put in coordinate slot #3 (see the IN command). The second example overwrites whatever coordinate records are in slot #n with the modified

records. For example, if you wanted to delete all the hydrogens in the primary set, you could deselect hydrogens (see the ATOM command) and then type "APPLY REPLACE."

When used without a number, APPLY REPLACE overwrites the coordinates in the PRIMARY coordinate set. The function "APPLY APPEND n" appends the new modified coordinates to the end of the records in coordinate slot #n. Again, if no number n is supplied, the new records go to the PRIMARY coordinate set. "APPLY MODIFY" always affects only the PRIMARY coordinate set. It is used, for example, if you wanted to set the B factors of all the hydrogen atoms to 20. You would select all the hydrogens (see the ATOM command), set NEWBFACTOR to 20 (see above) and then use "APPLY MODIFY". If you used "APPLY REPLACE" instead, the new PRIMARY coordinate set would contain only the modified hydrogens. All the other atoms would be deleted.

VI. Functions using 1 PDB file

The following are functions which need only one PDB file. Most apply only to the PRIMARY coordinate set. All send output both to the screen and to the current output file (see OUT), though typically moreinformation is sent to the output file.

1. LIST SEQ

Often it is useful to examine the sequence of your structure, perhaps to double check it against a published sequence. The "LIST SEQ" function sends a list of the sequence of the selected atoms to the current output file. The default is to write the sequence in the 3-letter code present in the PDB file. If the structure is a protein, you can direct GEM to write the sequence with the

1-letter amino acid code by typing "SET ONE" command prior to using "LIST SEQ". Typing "SET THREE" restores the use of the 3-letter code.

2. LIST ATOMS

"LIST ATOMS" generates an alphabetic list of the names of all the selected atoms in the PRIMARY coordinate set. The number of each type of atom is listed next to the atom name. This function is useful, for example, to see what atom types have been selected by the atom selection criteria or to check if there are any unusual atom types in the structure.

3. LIST RES

"LIST RES" produces an alphabetic list of the names of the residues of all selected atoms, with a count of the number of each type of residue. This is useful to see how many residues of a given type are present or to see if there are any unusual residue types in the structure.

4. RGYR

The RGYR function calculates a radius of gyration for the selected atoms of the PRIMARY coordinate set. The radius of gyration is defined as the mass-weighted rms (root mean square) distance of all atoms from the center of mass. The mass used is the standard atomic mass for each atom type. The center of mass is defined as the mass-weighted average position of the selected atoms. RGYR also reports the total mass of all selected atoms as well as the moment of inertia, which is the sum of the product of the atomic mass times the distance from the center of mass squared.

5. AXES

Every physical object has 3 mutually orthogonal principal inertial axes. The axes are defined by the property that an object can be rotated about a principal axis with no external torque. AXES calculates the principal axes for the selected atoms, taking the center of mass, defined as for RGYR, to be the

center of rotation. The length of each axis is defined to be mass-weighted rms distance from the center of mass in the direction of the axis (projection onto the unit vector in the direction of the axis). These lengths are related to the radius of gyration and in fact the sum of the squares of the lengths of the axes is equal to the square of the radius of gyration. These shape-weighted axes are written to the current output file in a form readable by INSIGHT as a USER file. This file can also be read by the programs Ellipse and Bigllips which generate USER files for Insight which represent the ellipse generated by the axes. The ellipse generated by Bigllips has axes longer by a factor of square root of five, thus displaying the solid ellipsoid which would give rise to the axes calculated by AXES. This object is called an "equivalent ellipsoid" and in some sense in the best ellipsoidal approximation to the given set of points. If you calculate AXES first for one domain and then for a second domain, AXES will report the distance between the domains and the angles between the axes defining the two domains.

6. BCOUNT

The BCOUNT function generates a histogram of the B factors of all selected atoms. The output is in the form of a CURVY file (see OUT). The average B factor of the selected atoms is reported to the screen. BCOUNT is useful for examining the distribution of B factors in the file. It has been observed that B factors in a well refined structure, fall in a maxwellian distribution (see Chapter one of this thesis on B factors). This is perhaps not surprising if the B factor represents some sort of energy which has been partitioned among the atoms.

Using the command "SET NOHIST" will cause BCOUNT to plot only the peaks in each bin, and not the entire rectangle as in a histogram. This is

useful if you want to fit a Maxwellian function to the output curve. Using "SET HIST" will return BCOUNT to plotting histograms.

7. BSEQ

When only one coordinate set is in use, BSEQ generates a plottable graph of the average B factor for a residue versus the residue number (or sequence position). The average is calculated over all the selected atoms of each residue. The average can be calculated over several residues in sequence with the WINDOW command.

8. SORTPDB

SORTPDB takes all records in the primary coordinate set and sorts them according to predefined criteria and creates a new coordinate set containing the sorted records. GEM puts the new coordinate set in the first available slot (see the IN command, above).

Records are sorted first on SEGID. The order of the SEGIDs is set, not alphabetically, but by which segid has the lowest residue number in it. After SEGID, records are sorted by residue number. Within each residue number, the order of the atoms is defined as N, H, sidechain, C, O. The sidechain atoms are sorted from closest to mainchain to farthest based on the Romanized Greek positional nomenclature assigned to each atom, i.e., A, B, G, D, E, Z, H (which represent α, β, γ, δ, ε, ζ, η respectively).

VII. Functions using 2 PDB files

1. PRIMARY/SECONDARY

GEM can manipulate up to four coordinate sets at once. Files are read in with the IN command and stored sequentially in coordinate slots 1, 2, 3 and 4. By default, functions of one file apply to coordinate slot #1 and

functions of two files apply to coordinate slots #1 and #2. You can override these defaults with the PRIMARY and SECONDARY commands.

PRIMARY n
SECONDARY n

"PRIMARY n" will direct the functions of one file to use coordinate set n (n=1,2,3,4). "SECONDARY n" will direct GEM to use coordinate set n as the comparison set for functions requiring two files. For example, if you wanted to SUPERimpose coordinate set #4 on coordinate set #3, you would use "PRIMARY 3" and "SECONDARY 4" before executing the SUPER function. "PRIMARY" or "SECONDARY" alone, with no arguments, shows which coordinate sets are currently being used. This can also be checked with the "IN" command with no arguments. This generates a list of the files currently read in. A "[1]" following a file name indicates the file is the PRIMARY coordinate set, while a "[2]" follows the file name of the SECONDARY coordinate set.

2. LIST DIFF

For all functions requiring a comparison of two coordinate sets, atoms from the PRIMARY set must be matched to atoms in the SECONDARY set. Two atoms are matched if they have the same Chain ID, Residue number, Insertion code, Atom name and Segment ID, although the Chain ID and Segment ID matching can be disabled with the "SET NOCHAIN" and "SET NOSEGID" commands (see the SET command). Note specifically that Residue Type is not considered. This means that atoms with the same name from a PHE and a TYR, for example, can be matched with eachother. If two files have different Segids or are numbered differently, you will have to

modify one or both files (with the Record Modification Commands) to match the atoms you want matched. The "LIST DIFF" function lists all the selected atoms in the PRIMARY coordinate set which have no matching selected atoms in the SECONDARY coordinate set. This listing is sent to the screen as well as to the current output file.

3. RSEQ

The RSEQ function generates a graphable file of Delta R versus residue number (sequence position) for all the selected atoms. Delta R is the distance between an atom in the PRIMARY coordinate set and its matching atom in the SECONDARY coordinate set. If more than one atom is selected in a given residue, the rms (root mean square) average Delta R value over all the atoms in the residue is reported. With the WINDOW command, the rms average can be extended to be over several residues.

4. PDBRMS

PDBRMS calculates the rms (root mean square) average of Delta R over all selected atoms. With the atom selection commands, you can quickly determine the rms deviation of any selected subset of atoms: all carbon alphas, all mainchain, and so on.

5. SUPER

The SUPER function calculates the optimal matrix to superimpose the selected atoms of the PRIMARY set onto those of the SECONDARY set. The optimal matrix is that which minimizes the resulting rms average delta R of the selected atoms. The actually transformation is not performed, only the matrix is calculated. The matrix can be reviewed with the MATRIX command. The matrix is written to the current output file, and can be read back in at a later time with the "MATRIX @filename" command (see MATRIX). This function uses the Kabsch analytic algorithm to calculate the

matrix. SUPER reports the rms deviation you would obtain for the selected atoms if the transformation were performed. Note that if you've calculated the matrix from a subset of the atoms, but you want to transform the entire molecule, you'll have to undo the Atom Selection Criteria.

6. AXES

A full description of the AXES calculation is given in the section on functions of one coordinate set. AXES is also a useful way of looking at conformational change. If there two files currently read in, AXES will calculate the principal axes for the selected atoms in both coordinate sets and then calculate how much the center of mass moved and how much the axes have rotated.

7. BVSB

BVSB generates a graphable plot of the B factors in the PRIMARY coordinate set versus the B factors in the SECONDARY set for all selected atoms. BVSB also performs a linear least sqaures fit to this plot and reports the correlation coefficient, slope and intercept of the fit. This is a measure of how closely related the B factors are in two different files, or how much the B factors have changed during B factor refinement.

8. BSEQ

When used with two coordinate sets, BSEQ generates a plottable file of the difference in B factor between the primary coordinate set and the secondary coordinate set versus the residue number. If more than one atom is selected in a given residue, the value plotted is the average delta B over all the atoms in the residue. As with RSEQ (above), the WINDOW command allows you to extend the averaging to be over several residues.

9. BSIGMA

BSIGMA calculates the rms (root mean square) average of all the delta B's for all the selected atoms, where Delta B is the B factor for an atom in the PRIMARY set minus the B factor for its matching atom in the SECONDARY coordinate set. BSIGMA is analagous to the PDBRMS function and is a measure of the similarity in B factor between two coordinate sets.

VIII. Miscellaneous commands

1. DO

The DO command selects which function is to be performed on the coordinates. DO with no arguments lists all of GEM's functions and prompts the user to select one. DO with a function name, for example "DO PDBRMS" selects that function as the next function to be executed. The actual function is not performed, but the GEM's prompt is changed to reflect the selection. To actually execute a function you need to type in the function name, or use the "GO" function.

2. GO

The GO function executes the current default selected function. The function to be performed is reflected in the current prompt which is set by the DO command, or by performing some function by typing in its name. For example, if the last function you did was "RGYR" and you want to do that again (as for a different subset of atoms) you can either type "RGYR" again, or just type "GO".

3. SET

The SET command is for setting certain flags which affect some of the GEM functions. The flags are as follows:

```
SET HIST          ! BCOUNT should plot full histograms
```

```
SET NOHIST      ! BCOUNT should only plot the peak for each rectangle
SET ONE         ! LIST SEQ should write 1-letter amino acid codes
SET THREE       ! LIST SEQ should use the residue type given in the file
SET XPLOR       ! WRITE should include REMARK and END
SET NOXPLOR     ! WRITE omits REMAKR and END records
SET SEGID       ! comparisons will examine Segment ID
SET NOSEG       ! comparisons will ignore SEGID
SET CHAIN       ! comparisons will examine Chain ID
SET NOCHAIN     ! comparisons will ignore Chain ID
```

See the named functions for further information.

4. WINDOW

The WINDOW command sets the window over which averaging should be performed for the RSEQ and BSEQ functions. For example,

```
WINDOW 5
```

Selects a window of 5 residues, two before and two after each residue. The default is a window of 1, which means average over only one residue. A window with an even number N will average over N/2 residues before and (N/2-1) residues after each residue.

Appendix 2-A. PDB Files

PDB stands for protein databank. A PDB file contains information about a macromolecular structure in a (relatively) standard format. Here is a typical record from a PDB file:

ATOM 100 NE1 TRP A 14 20.564 14.102 -8.265 7.00 33.70 2HHB 305

One record is one line in a PDB file. An "ATOM" record, the only kind of record GEM uses, contains information about a single atom in the structure. The first field, "ATOM", is a record identifier and specifies what sort of record this is. The 14 fields of an ATOM record are listed below.

1. Record Type columns 1 to 6

ATOM ----------

Currently GEM only uses ATOM type records. All other record types are ignored.

2. Atom Number columns 7 to 11

----- 100----------

The sequential number of this atom in the structure. GEM ignores this field as well.

3. Atom Type columns 12 to 17

---------- NE1 ----------

The name of the atom, with Romanized Greek letters. For example, the name NE1 indicates the epsilon nitrogen. In an input GEM file, the name can start in any of the 6 columns. In output, Gem left justifies the name starting at column 14, unless the name is 4 letters long, in which case it starts at

columns 13.

4. Residue Type columns 18 to 21

----------TRP ----------

A 3-letter code name for the residue or base.

5. Chain Identifier column 22

----------A----------

This column is often used to indicate separate monomers of a dimer, trimer or tetramer. The chain ID is used in the PDB and is recognized by FRODO, but is ignored by X-PLOR.

6. Residue Number columns 23 to 26

---------- 14----------

Either the sequential number of this residue in this chain or the position of the residue in some line-up; for example, when a serine protease is numbered with respect to chymotrypsin. Each residue must have a unique residue number (perhaps with the use of an insertion code, see below) in the given chain. Residue numbers over 999 are problematic. Frodo uses only 3 digit residue numbers, and X-PLOR uses columns 24 to 27 for a 4 digit residue number, which is incorrect. However, GEM can interpret this arrangement if encountered in an input PDB file. It is preferable to make use of Chain Identifiers in Frodo and Segment IDs in X-PLOR.

7. Insertion Code column 27

---------- ----------

If a given sequence is numbered relative to a different sequence, gaps and insertions may occur. If an insertion is encountered, the residues

of the insertion are given the number of the residue before the insertion and then given sequential insertion codes starting with the letter A.

8. X coordinate columns 31 to 38

---------- 20.564----------

PDB files contain coordinates in real-space orthogonalized axes on an Angstrom scale. This field is 8 characters long, containing 3 digits following the decimal point.

9. Y coordinate columns 39 to 46

---------- 14.102----------

10. Z coordinate columns 47 to 54

---------- -8.265----------

11. Occupancy columns 55 to 60

---------- 7.00----------

The occupancy field contains either the number of electrons at the given atom position or the fractional occupancy at the position (only one or the other in any given PDB file). FRODO ignores this field. X-PLOR expects the fractional occupancy which will most often be 1.0.

12. B factor columns 61 to 66

---------- 33.70----------

The B factor is the isotropic temperature factor, which has units of square Angstroms. This most often will have a value between 2.00 and

40.00. Occasionally, in the PDB, an atom will be given a B factor of 0.00 if it couldn't be seen in the electron density. On the other hand, a B factor of above 50 really means that atom wasn't in the density anyway.

13. Segment ID columns 73 to 76

----------2HHB---

A four character identifier, used extensively by X-PLOR. In the PDB this field contains the name of the PDB file. In X-PLOR, each separate "groupings" of atoms is typically given its own Segment ID, or SEGID, for short. The SEGID is ignored by FRODO.

14. Line Number columns 77 to 80

---------- 305

The sequential number of this line in the file. Used almost exclusively in the PDB itself and ignored by all other programs.

Appendix 2-B. Glossary of GEM commands and functions

?	Same as Help
APPLY	Applies record modification instructions
ATOM	Atom selection on atom name
AVGB	Atom selection on average B between primary/secondary set
AXES	Calculates principal axes
BCOUNT	Histogram of B factor distribution and average B factor
BFACTOR	Atom selection on B factor
BSEQ	Plot of (delta) B factor versus sequence position

BSIGMA	standard deviation of delta B values between two sets
BVSB	plot of B factor vs B factor for two sets, and linear fit
CHAIN	Atom selection on Chain ID label
DO	Selects which function to perform next
DUMP	Displays coordinate records to the screen
GO	Performs the currently selected function
IN	Reads in PDB FILES
LIMITS	Give min, max for numeric fields
LIST ATOMS	Generates alphabetic list of atom name occurences
LIST DIFF	Displays PRIMARY atoms with no match in SECONDARY set
LIST RES	Generates alphabetic list of residue name occurences
LIST SEQ	Lists sequence of residues in the structure
NEWBFACTOR	Modifies B factor field
NEWCHAIN	Modifies Chain Identifier field
NEWOCC	Modifies Occupancy field
NEWRANGE	Modifies residue number field
NEWS	Lists latest improvements to GEM
NEWSEGID	Modifies Segment Identifier field
OCCUPANCY	Atom selection on occupancy
OUT	Specifies name of output file
PDBRMS	RMS deviation of selected atoms
PRIMARY	Specifies which coordinate slot is the PRIMARY set
RANGE	Atom selection on residue number
RESET	Sets commands back to their default values
RESIDUE	Atom selection on residue name
REVIEW	Lists commands not currenty set to default values

RGYR	Calculates radius of gyration
RSEQ	RMS deviation per residue versus residue number
SECONDARY	Specifies which coordinate slot is the SECONDARY set
SEGID	Atom selection on segment ID
SET	Sets specific flags
SORT	Creates new, sorted, coordinate set
SUPER	Calculates matrix to superimpose PRIMARY on SECONDARY
WINDOW	Window-size for BSEQ, RSEQ functions
WRITE	Writes coordinates to a PDB file

Appendix 3. RamPlus

RamPlus - a program for calculating and evaluating dihedral angles in a PDB file

written by Eric Fauman July 1990

This program evaluates main chain and side chain dihedral or torsional angles. One substantial feature of RamPlus is the ability to select subsets of residues on the basis of residue number or residue type. The primary limitation of RamPlus is that only "minimal" PDB files can be used: i.e. the input PDB files should contain only ATOM records, the residue numbers should be integers (for example, not 65A) and there should be no columns past the B factor column (for example, segid information). However, any PDB file can be edited to these standards by the program GEM, described in Appendix 2 of this thesis.

RamPlus by itself can be used for 3 main functions, which are described below. Also, a list of commands is available from within RamPlus by entering the word 'help'.

Making a Ramachandran scatter plot (with or without contours)

The following will create a text file containing the phi-psi angles of a selected set of residues:

```
-----------------
$run ramplus
in  mypdbfile.pdb
out mypdbfile.gly
res gly
curvy
go
```

```
$run ramplus
in  mypdbfile.pdb
out  mypdbfile.nogly
res -gly
curvy
contour
go
```

The file mypdbfile.gly contains phi-psi values for glycines only, while the file mypdbfile.nogly contains all the other phi-psi values, and also will contain the familiar Ramachandran contour lines.

Evaluating main chain dihedrals

The following will create a text file detailing phi, psi, and omega, and in addition will assign secondary structure and evaluate how close a residue's phi-psi is to one of the allowed regions:

```
$run ramplus
in mypdbfile.pdb
out mypdbfile.dih
secondary
go
```

Note that the omega calculated for residue i is that torsional angle between the C of i-1 and the N of i. Thus a cis-proline at position i will be have an omega of (near) 0 at position i.

Evaluating side chain dihedrals

The following will create a text file which lists all side-chain dihedral angles and will assign each residue to one of the Ponder-Richards rotamers, if possible. The user enters a sigma cutoff, where the sigma is the standard deviation for each angle in each rotamer as given in Ponder and Richards. If

a residue is too far from any rotamer, RamPlus reports the closest rotamer in angle space, giving the greatest weight to chi1, then to chi2 and so on. This is not always the closest rotamer in coordinate space, and a better algorithm, not available in RamPlus, is to pick the rotamer which moves the center of mass of the sidechain the least, as is done by another program of mine, called FitRot.

```
$run ramplus
in mypdbfile.pdb
out mypdbfile.side
sidechain
sigma 2.0
go
```

---

note that a sigma cutoff of 2.0 is the default, so the sigma command above is actually not needed.