Research Articles

Plastic Adaptation Toward Mutations in Proteins: Structural Comparison of Thymidylate Synthases

Kathy M. Perry,[1] Eric B. Fauman,[1] Janet S. Finer-Moore,[1] William R. Montfort,[1] Gladys F. Maley,[2] Frank Maley,[2] and Robert M. Stroud[1]

[1]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, California 94143-0448, and [2]Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, New York 12201.

ABSTRACT **The structure of thymidylate synthase (TS) from *Escherichia coli* was solved from cubic crystals with $a = 133$ Å grown under reducing conditions at pH 7.0, and refined to $R = 22\%$ at 2.1 Å resolution. The structure is compared with that from *Lactobacillus casei* solved to $R = 21\%$ at 2.3 Å resolution. The structures are compared using a difference distance matrix, which identifies a common core of residues that retains the same relationship to one another in both species. After subtraction of the effects of a 50 amino acid insert present in *Lactobacillus casei*, differences in position of atoms correlate with temperature factors and with distance from the nearest substituted residue. The dependence of structural difference on thermal factor is parameterized and reflects both errors in coordinates that correlate with thermal factor, and the increased width of the energy well in which atoms of high thermal factor lie. The dependence of structural difference on distance from the nearest substitution also depends on thermal factors and shows an exponential dependence with half maximal effect at 3.0 Å from the substitution. This represents the plastic accommodation of the protein which is parameterized in terms of thermal B factor and distance from a mutational change.**

Key words: **thymidylate synthase, plasticity, crystal structure, B factor, mutation**

INTRODUCTION

Thymidylate Synthase

Thymidylate synthase (TS, EC 2.1.1.45) plays an essential role in DNA synthesis. To understand how methyl group transfer takes place we sought to find one species where structures of both unliganded enzyme and a complex with a transition state analog bound could be determined and compared. A major conformational change accompanies binding of substrates and catalysis and this has so far prevented us from examining the structure of transition state analogs bound to the *L. casei* enzyme, whose unliganded* structure we determined first.[1] With the cloned and expressed *E. coli* enzyme we were able to crystallize both the unliganded enzyme and a ternary complex formed between enzyme, its substrate, and a folate analog.[2] These structures, solved independently from different crystal forms, provide insights into the reaction chemistry. Here we report the structure determined for the enzyme from *E. coli,* and compare it with that from *L. casei.* Through this comparison we aim to describe a general theory for the average degree of plastic adaptation in these protein structures in response to changes in sequence.

TS is among the most highly conserved enzymes; approximately 18% of the residues are absolutely conserved among the 17 known sequences. The unit evolutionary period for TS is 22 × 10^6 years, longer than for most other enzymes and about the same as for cytochromes.[3,4] Yet the cytochromes are small (i.e., the conserved functional component per unit volume is high) and much of their surface is involved in interaction with oxidase and reductase. Both factors are responsible for high conservation in cytochromes. While the high conservation of TS sequences is clearly related to the chemical reaction catalyzed, it may also be linked to interaction with other enzymes and to conserved conformational dynamics upon substrate binding.[5]

Four native enzymes, now including structures for

*All native or unliganded enzymes discussed have a phosphate ion (P_i) bound at the dUMP phosphate site.

Received November 27, 1989; revision accepted February 15, 1990.

Address reprint requests to Robert M. Stroud, S-960 Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143-0448.

Abbreviations: TS, thymidylate synthase; ECTS, TS from *Escherichia coli*; LCTS, TS from *Lactobacillus casei*; CH_2THF, methylenetetrahydrofolate; TES, N-tris[Hydroxymethyl] methyl-2-aminoethane sulfonic acid; βME, 2-mercaptoethanol; dUMP, 2′-deoxyuridine 5′-phosphate; KH_2PO_4, potassium phosphate; EDTA, ethylenediaminetetraacetic acid; DTT, dithiothreitol; CB3717, 10-propargyl-5,8-dideazafolate; LCR, largest constant region; Cα, α-carbon.

William R. Montfort's current address is: Department of Biochemistry, the University of Arizona, Tucson, Arizona 85721.

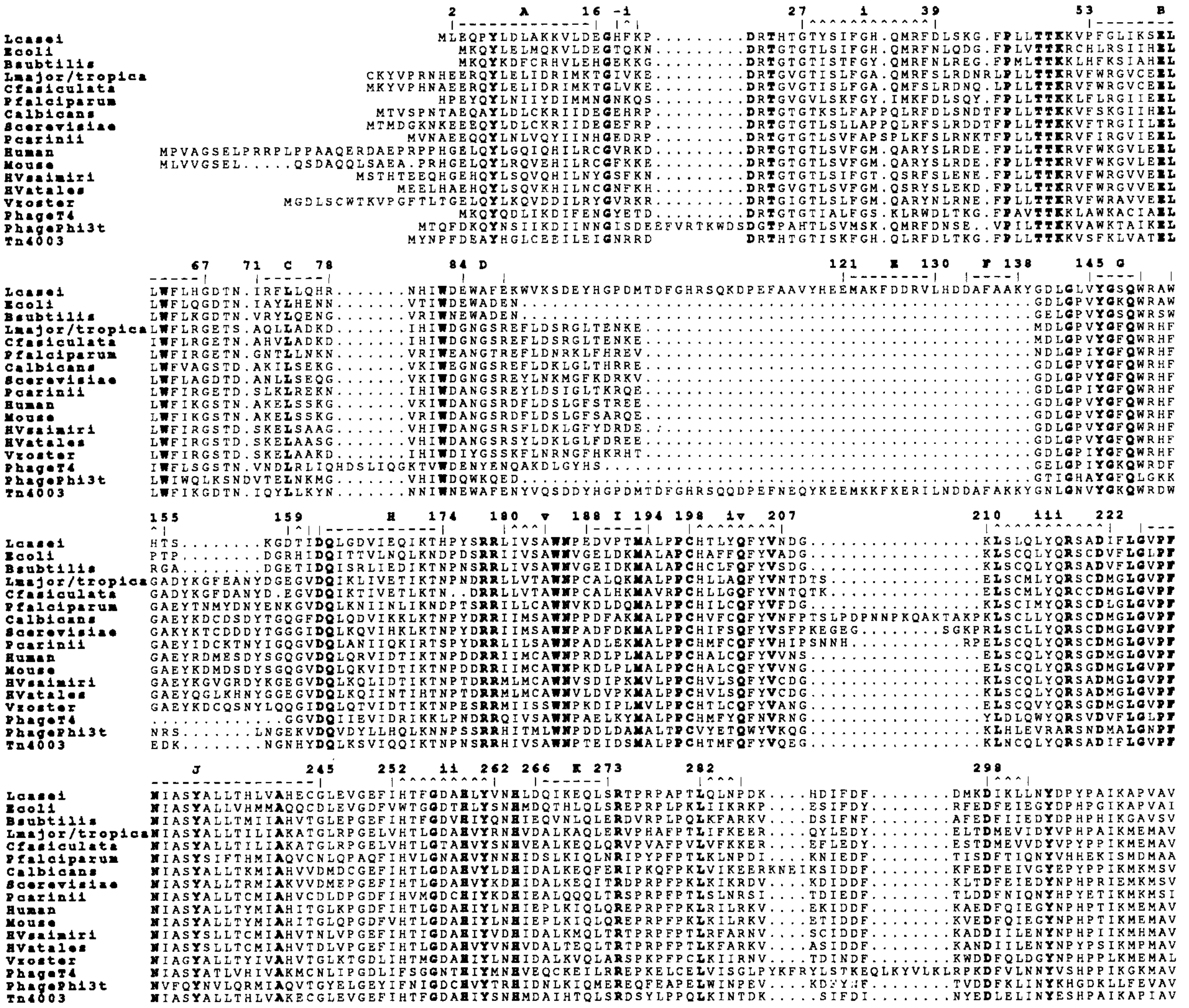


Fig. 1. The aligned sequences of 17 TS species. Invariant residues are in boldface. The numbering scheme and secondary structure (/\/\/\/\ = β-sheet; ⌐– – – = α-helix) marked above the sequences refer to TS from *L. casei*. Vertical bars mark the first and last residue of each secondary structural element. TS sequences shown include those for *Lactobacillus casei*,[40–42] *Escherichia coli*,[43] *Bacillus subtilis*,[44] human,[45] mouse,[46] *Herpesvirus saimiri*,[47] *Herpesvirus atales*,[48] *Leishmania major*,[49] *Leishmania tropica*,[50] *Plasmodium falciparum*,[51] *Varicella-Zoster* virus,[52] *Saccharomyces cerevisiae*,[53] *Pneumocystis carinii*,[54] *Candida albicans*,[55] coliphage T4,[56] *Bacillus subtilis* phage ϕ3T,[57] *Crithida fasiculata*,[58] *Staphylococcus aureus* transposon *Tn*4003.[59]

TS from T4 phage and human (in preparation), are perfectly symmetric dimers. However, there is much independent evidence for asymmetry in the catalyzed reaction.[6–8] The asymmetry is also represented in crystal structures of the ternary complexes of TS where the entire dimer becomes the asymmetric unit.[2] There may be as yet unrecognized functions of TS which require an allosteric change relayed from one active site to the other.

Of the known sequences of TS (Fig. 1) those from *E. coli* and *L. casei* are from among the most distantly related,[9] sharing 60% sequence identity overall. Therefore, they sample the most essential elements of the enzyme. Of the residues associated with ligand binding in the active site, 100% are conserved, presumably for functional reasons. However, the 122 internal residues are only 75% conserved, indicating that conservation of an internal residue depends on the degree to which structure change can be tolerated in response to a sequence substitution at that site. Analyzing the structural variation of buried residues in these related proteins, we derive a relationship between the plasticity of folded proteins, location in the protein structure, and the experimentally observed thermal vibration parameters.

Plasticity of Proteins in Response to Mutation

Site-directed mutations affect chemical properties of a protein in an as yet poorly understood manner, often with a variation due to accommodation of a substitution by readjustment of neighboring residues. This plasticity of structure toward substitution results in unexpected properties. Our goal here is to define quantitatively the expectation for structural plasticity in response to substitutions.

Several hundred site-directed mutations have been made in TS.[10–17] Highly conserved active site residues have been replaced by several other amino acids, though generally the protein remains active for several of the substitutions made at any position. Thus there is plasticity of function as well as structure. In subtilisin[18] and in α-lytic protease[19] substitutions that change specificity often lead to accommodation of a variety of substrates. Thus an overall estimate of the resilience of a protein is useful in predicting the expected latitude in chemical or thermodynamic properties of an altered protein.

Two approaches to parameterizing structural plasticity seem feasible. First is comparison of structures of many site-directed mutations of a protein. A second approach is to compare two structures that differ at many places in attempts to generalize the effects of substitutions. The first approach suffers because each mutation at a single locus is a special case; substitutions at a particular site may produce no change elsewhere, or large and distant changes. The required data base to give an overall expectation is therefore large and requires many structure determinations. Here we take the second approach. We compare two structures that differ at 105 residues out of 264, having 60% sequence identity. Further, they are 89% conserved in terms of corresponding atom sites, and share 3824 common atoms of 4300 (*E. coli* TS). From the data base we derive a parametric description of the difference in structures as a function of distance from the nearest substitution and thermal parameters of the atoms.

MATERIALS AND METHODS

Purification and Crystallization of *E. coli* TS

The *thyA* gene was cloned and expressed in *E. coli* K12 strain UC5826/pKTAH.[15,20] TS was purified by the method of Maley and Maley[15] yielding 100% activity (defined in terms of the best attainable specific activity of 5.0 units/mg). The enzyme was assayed spectroscopically by monitoring the normal enzyme catalyzed oxidation of cofactor, methylenetetrahydrofolate (CH_2-H_4folate), to dihydrofolate by its absorption at 340 nm[21] at 30°C. The assay buffer contained 50 mM TES, 25 mM $MgCl_2$, 6.5 mM formaldehyde, 1 mM ethylenediaminetetraacetic acid, disodium salt (EDTA), 75 mM 2-mercaptoethanol (βME), and at least a 10-fold excess of both 2′-deoxyuridine 5′-phosphate (dUMP) and CH_2-H_4folate at pH 7.4. The protein was stored as an ammonium sulfate precipitate in the absence of exogenous thiols at −70°C. TS is stable under the these storage conditions for at least 1 year as determined by specific activity after dialyzing against 50 mM potassium phosphate, 20 mM βME at pH 7.0.

Crystals were grown in 3 days at room temperature by vapor diffusion of a solution initially containing 10 μl of 2.5 mg/ml TS in buffer (0.2 mM EDTA, 1 mM DTT, 20 mM potassium phosphate at pH 7.0) made 1.15 M in ammonium sulfate, against 2.3 M ammonium sulfate with the same buffer.

Data Collection

Intensities to 2.1 Å resolution were collected from a single 400 μm diameter crystal using a Xentronics area detector, on a graphite monochromated 200 μm focal spot CuK_α source. A total of 66,941 observations were recorded for 22,275 independent reflections. Intensities were processed using the program of Howard et al.,[22] and statistics between ∞ and 2.08 Å resolution are provided in Table I. The overall weighted Rsym for corrected intensities[†] was 7.23%.

Overlapping Structures: Difference Distance Matrices

In order to assess the structural changes observed between the two proteins, we determined a constant "core" of α-carbons (Cαs) atoms whose positions are unchanged relative to each other in the two molecules. These regions were selected by first calculating intramolecular distances between all pairs of α-carbons common to sequences in both the *L. casei* dimer ($\delta_{Lij} = |\mathbf{r}_i - \mathbf{r}_j|$ where $\mathbf{r}_i$ is the atomic position of the Cα of residue *i*) and in the *E. coli* dimer (δ_{Eij}), δ_{Lij} and δ_{Eij} were then used to calculate an intermolecular difference distance matrix, $\Delta_{ij} = |\delta_{Lij} - \delta_{Eij}|$ which compares the distance between Cαs *i* and *j* in the *L. casei* structure with that between atoms *i* and *j* in TS from *E. coli*.

The constant region or core was selected by the following algorithm: first, all Cαs within 10 Å of the first Cα are located. All atoms, *j*, with $\delta_{E1j} < 10$ Å and $\Delta_{1j} < 0.5$ Å become candidates for the growing core. In those cases where two or more atoms are candidates for the core but have difference distances of more than 0.5 Å from each other, the atom which deviates the least from the other core atoms is chosen. All Δ_{ij} greater than 0.5 Å are rejected. These atoms constitute an initial core. This process is reiterated beginning with the current core atoms to find all the Cαs of constant distance from them which lie within 10 Å of at least one atom in the core. When no other atoms can be added to this core, the search is reinitiated at the next Cα($i = 2$). The search is initiated at each Cα and the core containing the largest number of atoms of therefore constant relationship in space is reported. This largest

[†] $$R_{\text{sym}} = \left\{ \frac{\sum_{hkl} \sum_{i=1}^{N} w_i (I_{\text{avg}} - I_i)^2}{\sum_{hkl} \sum_{i=1}^{N} w_i (I_i)^2} \right\}^{1/2} \quad \text{where } I_{\text{avg}} = \frac{1}{N} \sum_{i=1}^{N} I_i \text{ and } w_i = \frac{1}{\sigma_i^2}$$

TABLE I. Statistics on Crystallographic Data for *E. coli* and *L. casei* TS

Resolution range (Å)	Number of reflections collected	% of total possible reflections collected	Total number of observations	$I/\sigma(I)$*	% R-factor†
		E. coli TS			
∞–3.78	4025	96.2	16058	37.15	21.23‡
3.78–3.00	4013	99.9	14688	11.02	23.08
3.00–2.62	3950	99.8	12481	3.54	26.93
2.62–2.38	3974	99.9	11049	1.98	29.42
2.38–2.21	3813	97.2	8462	1.34	31.70
2.21–2.08	2500	63.3	4203	1.18	32.64
∞–2.08	22275	92.8	66941	10.04	24.74‡
		L. casei TS			
∞–4.19	3438	93.9	12643	36.34	20.94‡
4.19–3.33	3395	99.5	21636	34.57	17.55
3.33–2.91	3288	97.9	19033	18.73	21.99
2.91–2.64	3249	97.6	14931	11.72	24.87
2.64–2.45	2994	90.7	7439	6.72	26.83
2.45–2.31	2047	62.2	3908	5.11	27.97
∞–2.31	18411	90.5	79590	20.24	21.40‡

*Intensities (I) correspond to the average over multiple observations for each reflection.
†The %R-factor is calculated with no sigma cutoff imposed.
‡The %R-factor has been calculated using data from 7 Å.

constant region (LCR) is used as the reference core to superimpose the two sets of dimer coordinates by minimizing the overall rms differences, S, between core atoms, k, with respect to rotation, described by matrix $\mathbf{A}$, and translation, $\mathbf{T}$

$$S = \sqrt{\frac{1}{N_{\mathrm{LCR}}} \sum_{k=1}^{N_{\mathrm{LCR}}} (\mathbf{A}(x_k,y_k,z_k)_{\mathrm{LC}} + \mathbf{T} - (x_k,y_k,z_k)_{\mathrm{EC}})^2}$$

The transformed coordinate set, $\mathbf{A}(x_i,y_i,z_i)_{\mathrm{LC}} + \mathbf{T}$, was used as the *L. casei* coordinate set for all further calculations.

Several small constant regions are identified in the difference distance matrix. Each of these, characterized by constant internal vector distances between species, represents a small set of residues that move as a unit (or microdomain) with respect to the LCR.

Solvent Accessibility

Solvent accessibility[23] was determined by van der Waals contact of a protein atom with a spherical solvent molecule of radius 1.4 Å, assessed using van der Waals radii provided by Rose et al.[24] The percent accessibility was defined, relative to a stochastic standard state derived by Shrake and Rupley,[25] as the mean surface accessibility of residue X in the ensemble of tripeptides Gly-X-Gly with conformations as observed in protein structures. We categorize residues with <11% accessibility as buried, and those with >20% accessibility as exposed residues. Using solvent accessibility, we defined the surface as the set of atoms with >11% of the van der Waals surface solvent accessible. This set includes 576 atoms out of 4300 present in the dimer. The distance of any atom from the surface is then the minimum distance of that atom from any of the surface atoms.

Shift of Secondary Structural Elements

To compare changes in Cα position of secondary structural elements, n, between *L. casei* and *E. coli* proteins, we defined a vector sum, $\mathbf{V}_n$, as

$$\mathbf{V}_n = \sum_{i=p}^{q} [(x_i,y_i,z_i)_{\mathrm{LC}} - (x_i,y_i,z_i)_{\mathrm{EC}}]$$

over Cαs $i=p$ to q in either a strand, helix, or loop. The $\mathbf{V}_n$ for pairs of secondary structural elements were compared in direction. The angle between the vectors $\mathbf{V}_n$ and $\mathbf{V}_m$ is given by

$$\theta_{\mathrm{n,m}} = \arccos \left(\frac{\mathbf{V}_n \cdot \mathbf{V}_m}{|\mathbf{V}_n|\,|\mathbf{V}_m|} \right)$$

Parallel (+), antiparallel (−), and perpendicular (·) vectors (Fig. 2) are classified by $\theta < 45°$, $\theta > 135°$, and $45° \leq \theta \leq 135°$, respectively. Helix A of the *E. coli* protein, for example, is translated in parallel with strand ii relative to *L. casei* TS.

RESULTS

Solution and Refinement of *E. coli* TS Structure

Crystals of *E. coli* TS grow as rhombic dodecahedra up to 400 μm in diameter. Unit cell dimensions are 133.0 ± 0.3 Å as measured by precession photography. From the systematic absences, the space group was determined to be either I_{23} or I_{2_13}. The crystals contain 60% solvent with one monomer per asymmetric unit, thus the molecular 2-fold axis coincides with a crystallographic symmetry axis. The

	A	A/ii	ii	ii/B	B	B/C	C	D	G	H	H/I	I	J	iii	K	K/C term	C term
A	+																
A/ii	·	+															
ii	+	·	+														
ii/B	+	−	·	+													
B	·	+	·	·	+												
B/C	·	·	·	·	+	+											
C	·	·	·	·	+	+	+										
D	−	+	·	−	·	·	·	+									
G	·	+	·	·	+	+	+	·	+								
H	−	·	−	·	·	−	−	·	·	+							
H/I	·	+	·	−	·	·	·	+	+	·	+						
I	−	·	−	·	·	−	·	·	·	+	·	+					
J	·	·	−	·	·	−	·	·	−	+	−	+	+				
iii	·	·	+	·	+	+	·	·	+	−	+	−	−	+			
K	+	·	+	·	·	+	·	·	·	−	·	−	−	+	+		
K/C term	·	·	·	·	+	+	+	·	·	−	·	·	·	·	·	+	
C term	+	·	·	+	·	·	·	−	·	·	·	·	·	·	·	·	+

Fig. 2. Displacement of units of structure for *E. coli* TS relative to the *L. casei* protein. +, −, and ·, indicate shifts in parallel, antiparallel, and perpendicular directions, respectively. The sum of all Cα vectors for each unit was calculated and compared to all other units in the structure.

structure was solved by molecular replacement[26] using the *L. casei* TS dimer, oriented such that the molecular 2-fold was coincident with the crystallographic 2-fold along the *a* axis, as a template. The *L. casei* search model was modified in two ways: (1) the 50 amino acid insert and residues 1–2 unique to the *L. casei* sequence were removed (see Fig. 1); and (2) of the remaining residues, all side chains which are nonidentical between the two sequences were reduced to alanine. Glycines 55, 165, and 245 of *L. casei* were not increased to alanines.

The correct solution was obtained using the Crowther rotation function[27] with data to 4.5 Å resolution and a radius of integration from 4.5 to 25 Å. The solution was expected to be around the a axial direction, which allowed for an estimation of the noise level elsewhere and facilitated the correct choice. The position of the rotated search molecule in the unit cell was found by a residual search along (*x*,0,0) in I_{23}, and (*x*,0,1/4) in I_{2_13} (using the program of E. Dodson and P. Evans). A minimum residual[‡] of $R = 47\%$ in a background of 53–57% was obtained along (*x*,0,1/4) identifying the correct space group as I_{2_13}.

Restrained least-squares refinement[28] of the modified *L. casei* model gave a crystallographic residual $R = 36\%$ after 12 cycles against the 4.5 Å data set. After several cycles of rebuilding using the graphics package FRODO[29] and further restrained refinement against all the data, the *R*-factor is 25% for all data, and 22% for intensities greater than 2 σ between 7.0 and 2.1 Å resolution. Fourteen water molecules per monomer have been modeled into the density. The standard deviation of bond distances and angles from standard values is ± 0.018 and ± 0.071 Å, respectively. In Figure 3, density corresponding to Arg-286, Lys-287, and Pro-288 of *E. coli* TS demonstrates the correctness of the structure which was derived from the modified *L. casei* protein containing alanine at these positions.

Refinement of *L. casei* Structure

The structure of *L. casei* TS was from a different crystal form than we previously described.[1] The previous form is hexagonal, $P6_122$, with cell dimensions $a = 78.5$ Å and $c = 230.0$ Å. These crystals undergo a transition in which the *c* cell dimension rapidly increases by 12.8 Å to $a = 78.5$ Å and $c = 242.8$ Å, while still preserving integrity of the crystal. This second crystal form has the same space group and was solved by rigid body refinement of the molecular orientation and placement, and refined to higher resolution, 2.3 Å. After extensive refinement $R = 21\%$ for all the nonzero data between 7.0 and 2.3 Å. Improved density in the new crystal form clearly indicates the registry of the last 15 residues in the sequence, where there were several disordered side chains in the previous crystal form (see Fig. 4).[1] Residues 90–139 of the small domain are an insertion present in only one other species of TS. Of these, residues 91–118 are not interpreted in the weak and broken density for them. Residues 22–25 are in broken density and some of these may have multiple conformations in the crystal structure. All other residues with the exception of the C-terminal valine 316 are clearly resolved. Two water molecules per monomer have been modeled in the density.

Overall Structural Comparison

The difference distance matrix identifies a common core of 95 α-carbon atoms (out of 528 total) per dimer of TS as bearing a constant relationship ±0.5 Å to each other in both species of TS. This core was used in superimposing the structure of the *L. casei* dimer onto the *E. coli* dimer. The Cα backbone structures of the *E. coli* (white) and *L. casei* (black) monomers after overlap are shown in Figure 4. The tertiary fold is largely identical as expected for two proteins with 60% sequence identity. The dimer interface (Fig. 5) formed by the surface to surface association of β-sheets contained within each monomer has a unique right handed twist between the sheets with a dihedral of +28°, as first described in the *L. casei* native structure.

Each monomer in *L. casei* TS is composed of two domains. The larger domain, common to all species,

[‡] $$R = \frac{\sum|(|F_o| - |F_c|)|}{\sum|F_o|}$$

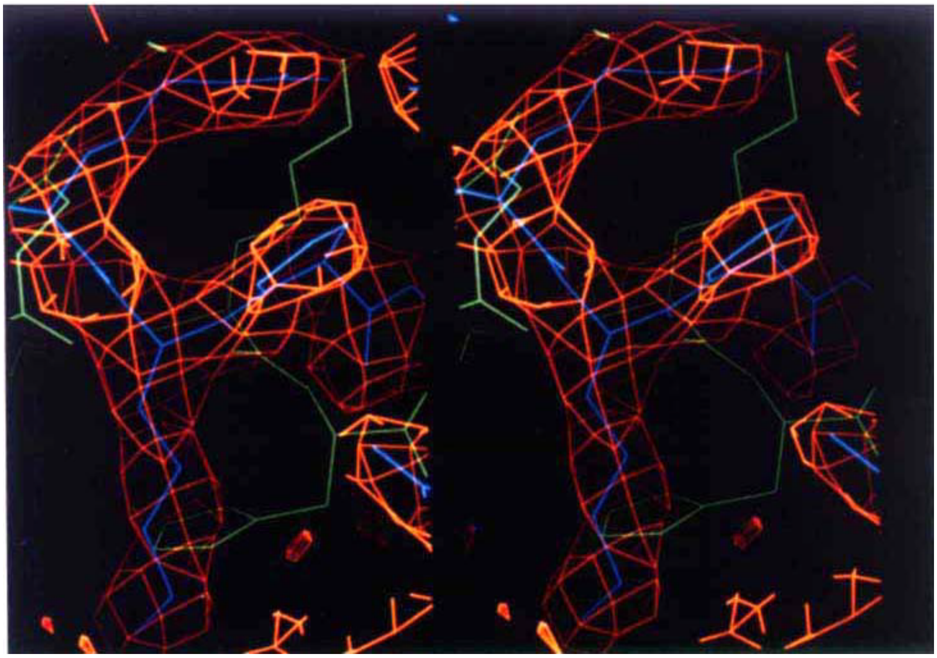


Fig. 3. $2F_o - F_c$ map (orange) corresponding to Arg-286, Lys-287, and Pro-288 (blue) of *E. coli* TS. Lys-288 and His-289 from the *L. casei* structure are superimposed in green.

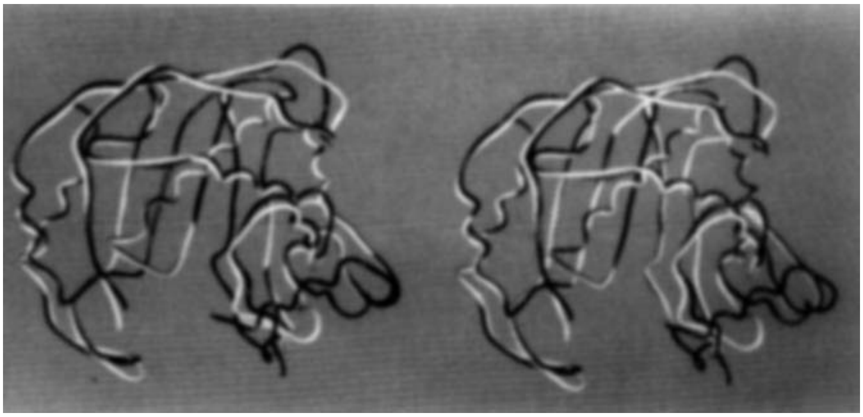
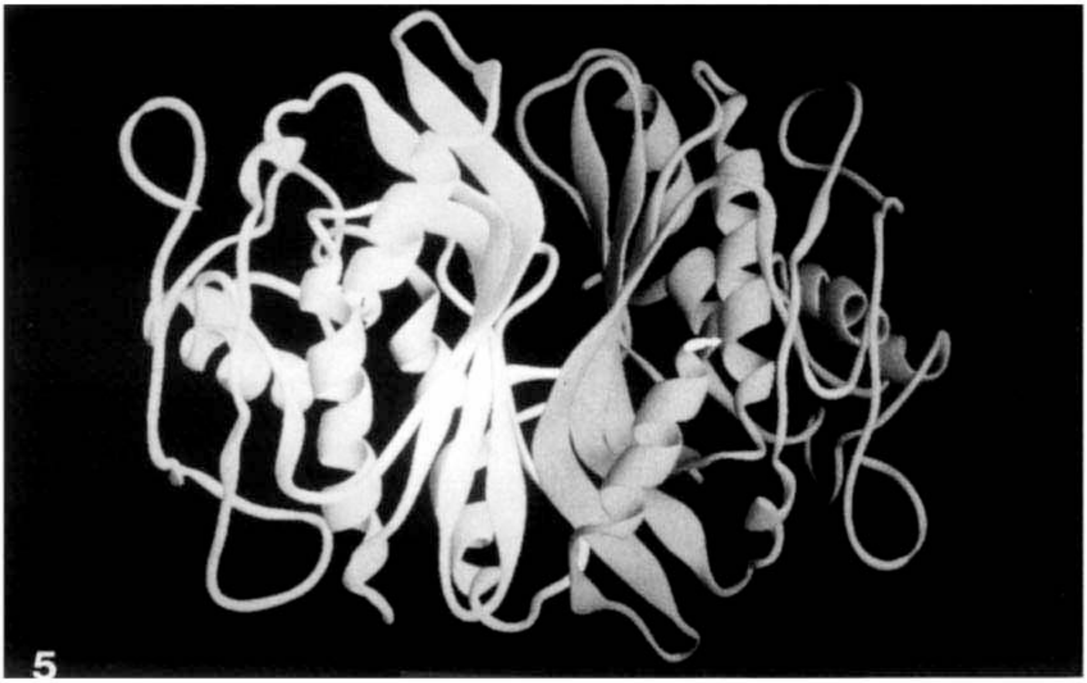


Fig. 4. Superimposed Cα atom tracings for the *L. casei* (black) and *E. coli* (white) proteins.

contains the β-sheet and helices[§] A, B and G to K (Figs. 6 and 4). The small domain includes residues 71–144 (see Figs. 1 and 4). Seventy percent of this small domain, residues 89 to 140, is absent in *E. coli* TS. The carbonyl carbon of residue 89 is 10.6 Å from the amide nitrogen of 140 in the *L. casei* structure. In the *E. coli* structure, a peptide bond links the two residues between conserved flanking regions, leaving just 18 residues in the small domain of *E. coli* TS. In both structures, helices A, B, C, G, H, and J are α-helices, helix I is a 3_{10} helix, and helix K is mixed, with one turn of α-helix followed by 2 turns of 3_{10} helix. Helix D is a loosely helical region in *L. casei,* with no regular secondary structure in *E. coli.*

Of the 528 residues in the TS dimers 206 are buried in both proteins. Of these, 75% are identical between the *L. casei* and *E. coli* sequences and 48% are nonpolar. In Figure 6, the buried region (white) consists of almost the entire β-sheet and much of helices A, B, G, and J. Portions of helices H, I, and K are also included. Large segments of external loops are exposed to solvent with the exception of the loop

[§] Nomenclature for secondary structural elements, and sequence numbering is as we first defined for *L. casei* TS.[1] The β-strands are numbered i–v for those in the β-sheet interface, and i, −i for the few paired residues around residue 20. The sequence numbers, *n,* for *E. coli* TS can be calculated from those for *L. casei, m,* as $n = m-2$ for *m* less than or equal to 89, $n = m-52$ for *m* greater than 89.



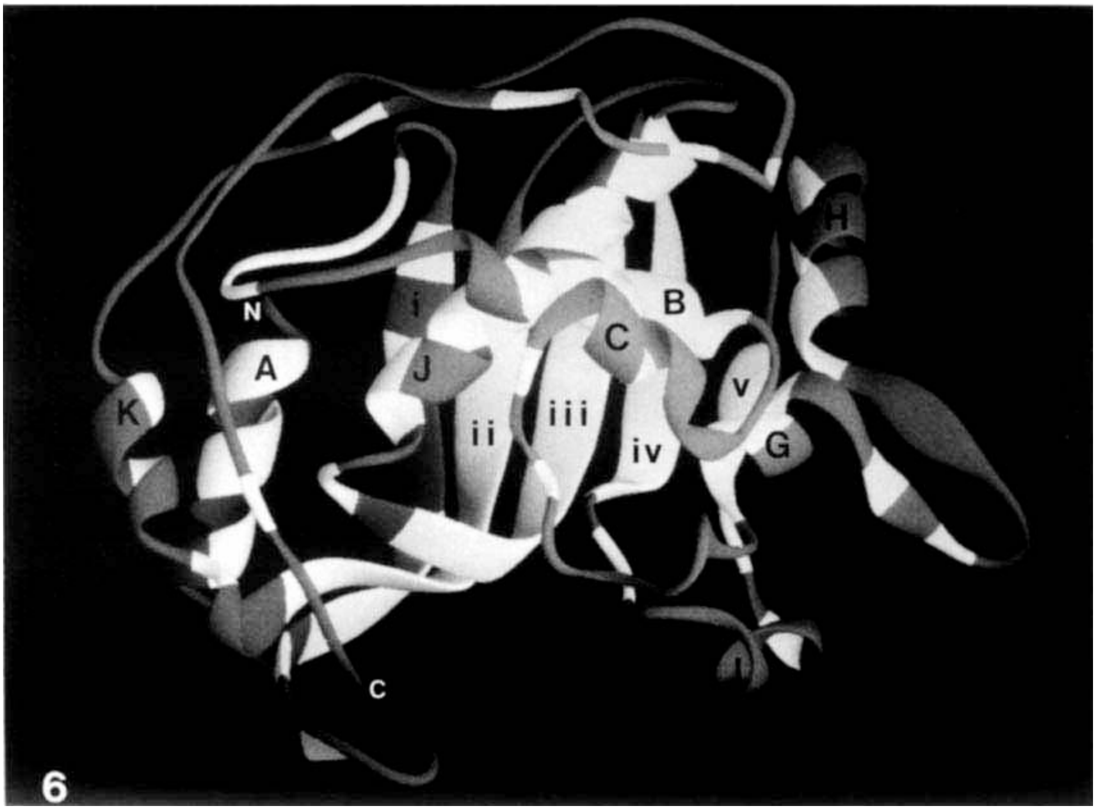


Fig. 5. Cα tracing of *E. coli* TS dimer showing the β-sheet interface and viewed approximately down the 2-fold axis (side view). The orientation is approximately 90° away from the that shown in Figures 4 and 6. The individual monomers are colored in white and gray.

Fig. 6. *E. coli* TS monomer (front view) indicating positions of all solvent inaccessible residues in white. The α-helices are labeled A through K and β-strands in order from left to right i through vi.

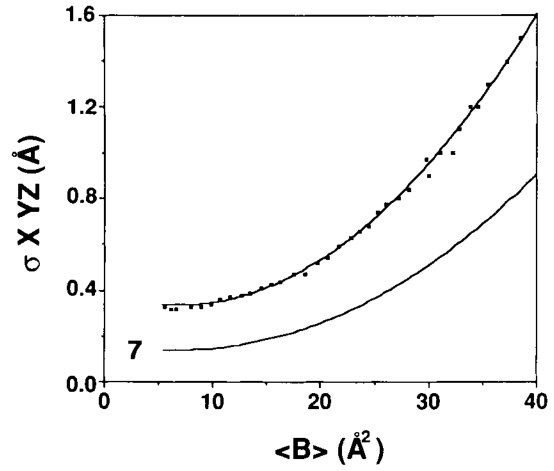


Fig. 7. Filled squares correspond to the standard deviation, σ_{xyz}, of a Gaussian distribution fitted to a histogram of shifts (Δx, Δy, and Δz) of atoms grouped as a function of B-factor for the TS structures. The curve is the best fit function of the form $\sigma_{xyz} = K_0B^2 + K_iB + K_2$. The B-factor used is that determined for the atoms in the *E. coli* structure. The lower curve is the adjusted trypsin curve [Eq. (7)] which shows the magnitude of shift between the two TS proteins which can be accounted by errors in the structures. The scale factor applied to the trypsin curve is described in Eq. (7).

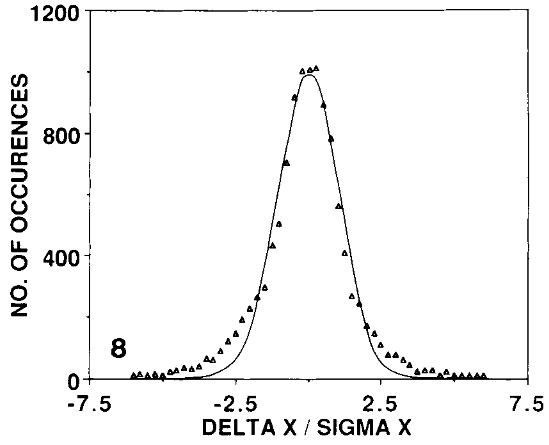


Fig. 8. Distribution of structural differences (Δx, Δy, and Δz) after division by $\sigma_{xyz}(B)$. The solid line is a Gaussian function with a standard deviation of 1.0, as expected if all differences were randomly distributed.

connecting strand i and helix B, which is in contact with the β-sheet interface.

Phosphate is a weak competitive inhibitor of TS.[8,30] Three ligand-free species we have solved to date (*E. coli, L. casei,* and phage T4) contain a phosphate ion situated between the guanidinium groups of three arginines 218, 178′, and 179′** at the same site the phosphate group of dUMP is bound.[2] All three arginines and one serine form hydrogen bonds to the phosphate oxygens of dUMP, however, the hydrogen bonds formed in the phosphate bound structures are longer and less regular. Arg-23 does not appear to hydrogen bond to the phosphate in the native forms of the enzyme (see also Ref. 5).

Separation of Significant Changes in Atomic Position From Residual Oscillations: The Snapshot Effect

The difference in atomic position between *E. coli* TS and *L. casei* TS include real differences and random variation due to termination of refinement at a given cycle which we will call the snapshot effect. These can be described by the vector shift, $\Delta\mathbf{r}_i$, or by Δx_i Δy_i Δz_i, defined as

$$\Delta x_i = x_{\mathrm{Ec},i} - x_{\mathrm{Lc},i},\ \Delta y_i = y_{\mathrm{Ec},i} - y_{\mathrm{Lc},i},$$
$$\Delta z_i = z_{\mathrm{Ec},i} - z_{\mathrm{Lc},i}$$

where $(x_i,y_i,z_i)_{\mathrm{Ec}}$ is an atomic coordinate in an orthogonalized axial set listed in Ångstroms. The atoms, i, are for all atoms that are common to both structures. For substituted side chains between species, common atoms within side chains are included, different atoms are omitted.

To extract the true positional variations, $\Delta\mathbf{s}_i$ or $\Delta\mathbf{x}'_i$, from the observed $\Delta\mathbf{r}_i$ requires estimation of the distribution of probable errors in position, $P(\Delta\mathbf{e}_i)$, associated with a given atom in both structures being compared. Then the observed differences $\Delta\mathbf{r}_i = \Delta\mathbf{s}_i + \Delta\mathbf{e}_i$ will be described by the integral

$$P(\Delta\mathbf{r}_i) = \int P(\Delta\mathbf{e}_i) \text{ over the surface } \Delta\mathbf{s}_i + \Delta\mathbf{e}_i = \Delta\mathbf{r}_i \quad (1)$$

The errors $\Delta\mathbf{e}_i$ include random errors for which we derive a statistical description of expectation, real differences due, for example, to the extra domain in *L. casei* TS, which we eliminate from the comparison, and systematic errors due to misplacement which are minimized by direct reassessment of density maps to see whether structures from both species could be identical at such sites.

The apparent random error component is a necessary consequence of the snapshot represented in a single coordinate set after one of a number of refinement cycles. Several cycles of constrained refinement necessarily lead to different coordinate sets which may nonetheless have identical R-factors.[31] The resulting random variations were found to be correlated with thermal vibrational factors, B_i, in well-refined structures by Chambers and Stroud.[31] Thus B-factors can be considered as a consequence of the width of the energy well in which each atom lies. The width of the energy well, represented by the amplitude of vibration u_i (usually assumed to be isotropic), is represented in the refined crystallographic coordinate list as a thermal vibrational factor B_i, related according to

**Residues whose side chains enter the active site, that derive from the monomer other than the one which provides the majority of the site, are designated by a prime.

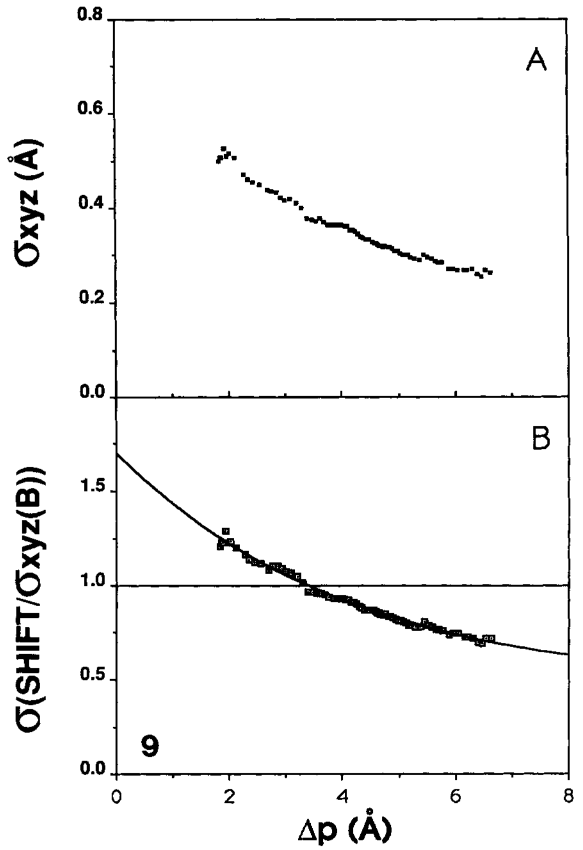


Fig. 9. (A) The standard deviation of a Gaussian distribution σ_{xyz} fitted to a histogram of shifts (Δx, Δy, and Δz) of atoms grouped as a function of distance from the geometrical center of the nearest amino acid substitution (Δp), plotted versus Δp. Only atoms closer to a buried substitution than to either a surface substitution, or to a Cα involved in a significant displacement (greater than 2 σ_r) were considered. (B) The standard deviation, σ_{ratio}, of a Gaussian distribution fitted to a histogram of atom shifts (Δx, Δy, and Δz) after correcting for dependence on B-factors, plotted as a function of distance from the nearest buried amino acid substitution (Δp). Only atoms closer to a buried substitution than to either a surface substitution, or to a Cα involved in a significant displacement (greater than 2 σ_r) were considered. The shift for each atom i was corrected for B-factor by dividing by the expected shift $\sigma_{xyz}(B_i)$. The curve is the best fit function of the form $\sigma_{\text{ratio}} = K_0 + K_1 \exp(-\Delta p/\tau)$.

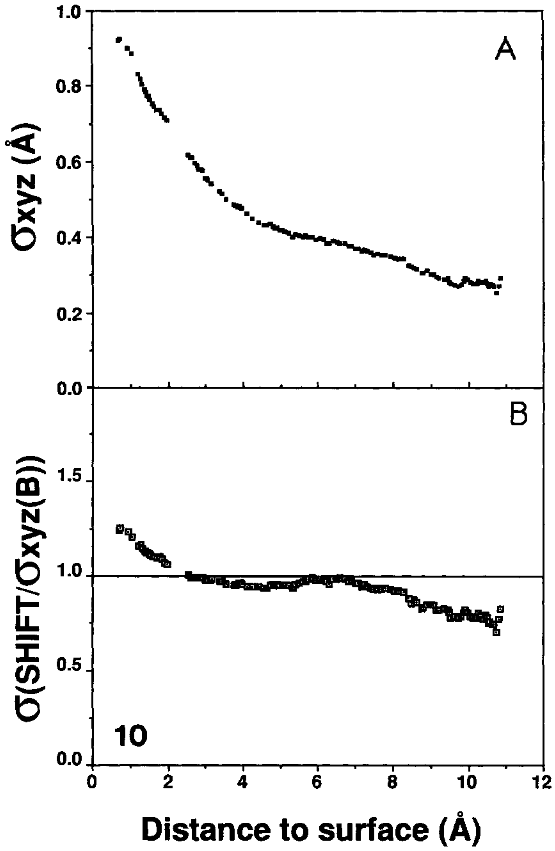


Fig. 10. (A) The standard deviation of a Gaussian distribution σ_{xyz} fitted to a histogram of shifts (Δx, Δy, and Δz) of atoms grouped as a function of distance from the surface (see Materials and Methods: Solvent Accessibility). Shifts are larger close to the surface. (B) The standard deviation, σ_{ratio}, of a Gaussian distribution fitted to a histogram of atom shifts (Δx, Δy, and Δz) after correcting for dependence on B-factors, plotted as a function of distance from the surface (see Materials and Methods: Solvent Accessibility). The shift for each atom i was corrected for B-factor by dividing by the expected shift $\sigma_{xyz}(B_i)$.

$$B_i = 8\pi^2 \langle u_i^2 \rangle \qquad (2)$$

where $\langle u_i^2 \rangle$ is the mean square displacement of the atom from its equilibrium position.

The sources of systematic error in coordinates include incomplete refinement, limited resolution of the data, and inhomogeneity of the crystal. Systematic differences in structure were reassessed by reference to the electron density maps, and in our case with inspection of the correspondence between TS from all four independently refined species. Thus we expect systematic errors to be much smaller than usually found in a crystal structure analysis since the same error would have to be consistent with all four density maps. The diffuse density which may indicate multiple positions or disordering of the structure is also represented in the thermal factors B_i. However, we use here only internal residues where such effects are rare.

The histogram of differences in structure between *E. coli* and *L. casei* TS yields a distribution with most components following a normal error distribution, and a few others that show significant deviations indicating a structural change. Without the latter, for any given range of $\langle B_i \rangle$ factor, the differences in position follow a normal or Gaussian distribution, and the width of the distribution, represented by its standard deviation (σ) increases with increasing $\langle B_i \rangle$. By plotting differences Δx_i, Δy_i, Δz_i versus B_i for all atoms in the dimer, we find that the standard deviation of positional variations, σ_{xyz} (including all Δx_i, Δy_i, Δz_i as independent observations in the same distribution) is a quadratic function of B (see Fig. 7) described by

$$\sigma_{xyz}(B) = 0.000983\, B^2 - 0.0102B + 0.325 \quad (3)$$

The standard deviation, σ_r, of the overall magnitude Δr is related to the standard deviation in the x direction, σ_{xyz}, by $\sigma_r = (\sqrt{3})\, \sigma_{xyz}$. The distribution of Δr values follows a Maxwellian distribution, whereas Δx, Δy, Δz are normally distributed. Thus, for ease of handling we use σ_{xyz} here. We derive an empirical relation for the probability of a given difference in position from those that follow a Gaussian distribution. This distribution includes both plastic shifts in position, caused by the changes in the sequence, and differences due to B_i related variations in position.

Those shifts that fall outside the Gaussian distribution represent more obvious differences in structure (see Fig. 8). Twenty percent of the atoms have $|\Delta r|$ that are greater than $2{*}\sigma_r(B)$. Most of these atoms are located in the B, C, and G helices and reflect the adjustment of the *L. casei* enzyme to the acquisition of the 50 residue insert in the small domain. Cαs of these residues are generally excluded from the plasticity calculation by the criteria used for selection (see below).

Protein Plasticity

In separating protein plasticity from energy well size (B-factor) related variations we analyzed differences in position versus proximity to an amino acid substitution. There are 210 amino acid substitutions in the common region of the TS dimer from *E. coli* and *L. casei*, of which 54 are completely internal and solvent inaccessible. Thus we analyze the dependence of structure change on distance from the nearest substitution, initially, just in internal residues in order to address effects propagated within the protein matrix. Shifts of atoms that are closer to one of the 54 buried substitutions than to either the surface or to a Cα involved in the systematic small-domain changes which fall outside the Gaussian distribution were included. There are 1629 such atoms in the dimer. The position of the substitution was defined as the centroid of the altered side chain atoms.

The average shift per atom is greatest closest to a substitution (see Fig. 9A). Atoms greater than 7 Å from a buried substitution are not represented in this figure since such atoms are necessarily closer to the surface than they are to any substition. As shown in Figure 10A, such atoms show larger positional differences between the structures, even slightly greater than that expected from their higher B-factors (see Fig. 10B). This is undoubtedly due to the crystal packing and ionic strength differences in the two crystals.

The effects on the magnitude of the shift due to B-factor alone are eliminated by computing the ratio of the observed shift, Δx_i, to the shift expected from the observed B_i value, $\sigma_{xyz}(B_i)$. Shown in Figure 9B is the standard deviation of this ratio $\sigma(\Delta x_i/\sigma_{xyz}(B_i))$ versus distance, Δp, from the nearest internal substitution. The fall off of this thermally corrected shift versus distance, we describe as due to plastic accommodation. It can be described by an exponential of the form

$$\sigma\left(\frac{\Delta x_i}{\sigma_{xyz}\,(B_i)}\right) = K_0 + K_1 \exp\,(-\Delta p/\tau) \qquad (4)$$

The characteristic distance, τ, is 4 Å, and effects of a substitution persist out from the nearest substitution through interatomic distances of ~7 Å. These shifts also include expected random errors in the structure which will contribute primarily to K_0. This exponential form parallels the kind of effects expected from incorporation of a defect into an arrangement of closely packed spheres.

The plot of the B-factor corrected shift as a function of distance to the surface (Fig. 10B) shows that greater energy well sizes account for 90% of this positional freedom at the surface. In contrast, the sharp decrease in $\sigma(\Delta x_i/\sigma_{xyz}(B_i))$ between 8 and 10 Å from the surface indicates a smaller than predicted set of shifts corresponding to a more rigidly structured core of 20 residues in the protein interior.

Lack of a Dependence on Size of Substitution

Intrinsically it seemed that the magnitude of structural perturbation might be a function of change in type or size, ΔV, of the substitution. To test this, ΔV was defined as the change in the number of heavy atoms (carbon, nitrogen, oxygen, sulfur) that occurs in a given substitution. Each thermally independent shift $(\Delta x_i/\sigma_{xyz}(B_i))$ was divided by the expected shift based on plasticity, $K_0 + K_1\exp(-\Delta p_i/\tau)$ which should now be independent of both thermal factor, and distance from a substitution, Δp_i. These expected shifts were grouped according to ΔV for the nearest substitution to assess any correlation between shift and magnitude of the nearby perturbation. No such dependence was found for all sizes of ΔV atom equivalents, which extend up to $\Delta V = 7$ (data not shown). Thus the corrected shifts all fall in the overall predicted distribution based on B_i and Δp_i alone. The size of the nearest substitutional change is uncorrelated with the size of the shift.

One reason why magnitude of a substitution has no effect on nearby shifts in position could be that larger mutations are accompanied by second site mutations nearby. To probe the relationship between structural plasticity and covariant adaptation, the distance between a given mutation and the closest second site mutation was plotted as a function of ΔV. As shown in Figure 11, we find that the larger the magnitude of a given buried mutation, the closer, on average, will be a second buried mutation. Since mutations with $\Delta V \leq 2$ are randomly

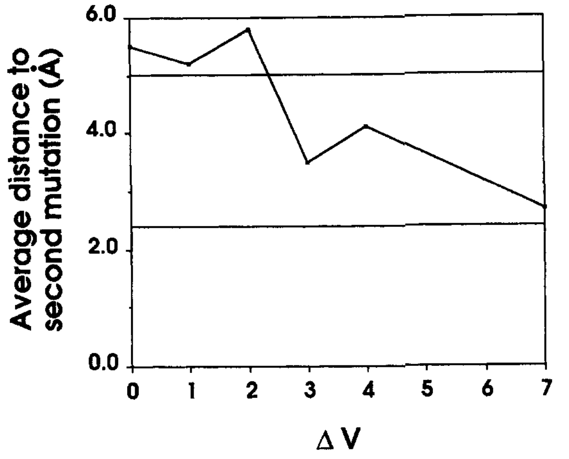


Fig. 11. Distance to the nearest second buried substitution as a function of the magnitude (ΔV) of the first buried substitution. The top horizontal bar indicates the average distance any atom is from a buried substitution. The bottom horizontal bar indicates the minimum distance between any two buried substitutions in the *E. coli* TS dimer.

positioned with respect to the other mutations, we conclude that in TS mutations of $\Delta V = 3$ or higher are more likely to be accompanied by a covariant change nearby.

Random Positional Variations in Protein Structures

To extract a general expression for expected plasticity requires subtraction of the *B*-factor related variance from the observations described in Eq. (4). Thus the variance is subtracted from the variance of total observed shifts, to give the variance in the general case.

The shifts between two different species, *E. coli* and *L. casei* TS, are expected to be larger than for identical structures refined independently (i.e., errors). On average the shifts are four times larger than the differences between the two chemically identical bovine trypsin structures, crystallized in the same space group, but solved independently, as compared by Chambers and Stroud.[31] Part of this increase is plastic accommodation to substituted amino acids and part is due to larger random errors and positional variation expected in structures of lower resolution and higher *R*-factor. The plastic alterations fall off at larger distance from the substitution. The asymptotic value at large distances from substitutions is one estimate of the overall random error in position. Alternatively it could represent an overall level of strain distributed throughout the protein, or simply be due to differences in crystal packing. Rather than make assumptions about which of these is true, we used an independently determined empirical error estimation.

Comparing pairs of structures for the same protein solved in different space groups, or for the same structures solved in different laboratories, and cases of independently refined structures in one asymmetric unit, we developed an empirical relation which estimates the average overall random errors in a structure as a function of resolution, and level of refinement:

$$\text{rms } \Delta\mathbf{e}_i(\text{C}\alpha\text{s}) = 3/4\, rR$$

where r is the resolution in Å, R is the crystallographic R-factor ($>2\sigma$ data), and the relation applies to Cαs in regions of the structure that are visible in the density map ($B_i<40$). For our trypsin structure refined to a residual $R = 15.7\%$ at 1.5 Å resolution in 1978,[31] this estimates the overall rms $\Delta\mathbf{e}_i$ as 0.18 Å. By comparison with an independently refined trypsin structure (the second with estimated rms $\Delta\mathbf{e}_i = 0.31$ Å at that time) a quadratic dependence of differences in position upon thermal parameters B_i was described.[31] The same relationship holds for the current better refined coordinates listed for the second structure,[32] with estimated rms $\Delta\mathbf{e}_i = 0.24$ Å. For any particular B_i, the differences were normally distributed with respect to x, y, and z, with a standard deviation of the distribution that depends on B_i. These differences, regarded as inevitable random errors of sampling in a crystal structure, can be separated between respective structures; thus they can be parameterized for each structure as

$$\sigma_{xyz,\text{error}}(B) = \tfrac{3}{4}rR\ (aB^2 + bB + c) \qquad (5)$$

where the quadratic component was derived from the trypsin comparison, to describe errors in each structure, and gives $a = 0.0015$, $b = -0.0203$, $c = 0.359$. This relationship can be used to estimate the magnitude of the random errors in either trypsin structure. The estimate of error for each TS structure, calculated based on the same formula would be rms $\Delta\mathbf{e}_i \approx 0.33$ Å for *L. casei*, 0.35 Å for *E. coli*. Thus the overall expected variation in the comparison of two TS structures would then be given by

$$\Delta\mathbf{e}_i\text{TS} = \sqrt{}(\text{rms } \Delta\mathbf{e}_i\ E.\ coli)^2 + (\text{rms } \Delta\mathbf{e}_i\ L.\ casei)^2 = 0.48\ \text{Å} \qquad (6)$$

Equations (5) and (6) also provide an estimate of the standard deviation of the overall error distribution, $\Delta\mathbf{e}_i$, in the TS comparison, as a function of B_i:

$$\sigma_{\text{TS,error}}(B) = \Delta\mathbf{e}_i\text{TS}\ (aB^2 + bB + c) \qquad (7)$$

Analyzed as a function of Δp_i, the shifts due to plastic deformation Δx_i expected in the general case for an atom with temperature factor B_i, versus expected errors can be expressed as a ratio, and described by a decaying exponential:

$$\sigma\left(\frac{\Delta x_i}{\sigma_{\text{TS,error}}(B_i)}\right) = K_0 + K_1 \exp(-\Delta p/\tau) \qquad (8)$$

In this case, for TS, the characteristic distance, τ, is 3.9 Å, K_1 is 3.1.

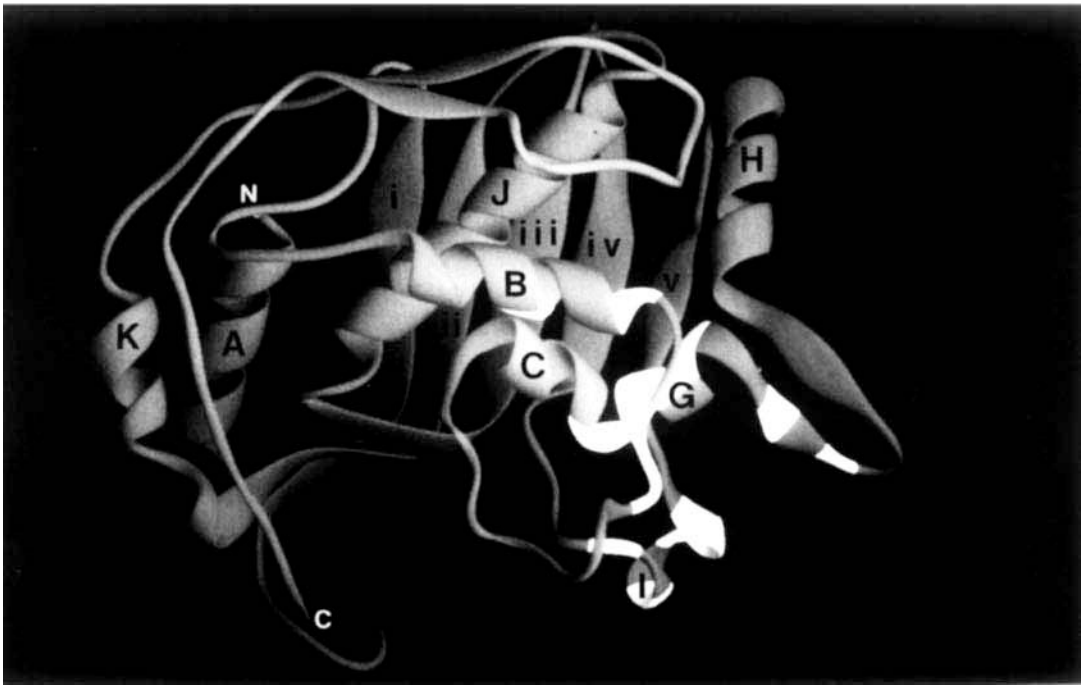


Fig. 12. *E. coli* TS monomer indicating positions of all buried Cα atoms (in white) that are also members of the subset of core residues used for overlapping (see Materials and Methods: Overlapping Structures).

The effects decrease exponentially toward an asymptotic value of K_0 (= 1.2) as the distance from substitutions increases. This value represents the sum of random error [Eq. (7)] and possible strain distributed evenly throughout the protein. K_0 reveals that even at large distances from any insult, shifts are 20% larger than expected from Eq. (7), indicating either that it underestimates the errors, or that the entire protein suffers some strain throughout.

Plastic Accommodation Among Buried Residues

Structural differences between *L. casei* and *E. coli* TS are conservative since both enzymes accommodate the same reaction chemistry, though the two species are only distantly related. Yet the 3-D structure of interior Cαs between the two proteins has an rms deviation of 0.9 Å as opposed to 1.3 Å for the entire protein, well above the thermal factor related shifts. The most structurally conserved region of TS is found in the β-sheet where most of strands ii, iii, iv, and v lie in the common core (Fig. 12). The 32 core residues which are in this β-sheet have an rms deviation among Cαs of 0.3 Å, and only 7 of these are nonidentical between the two species. Some of the substitutions show covariance in which multiple sequence changes combine to minimize distortion of the main chain and conserve the volume of packed side chains. Distortions are also expressed differently in different secondary structural elements, providing examples of segmental accommodation in which helices move as a unit. Such structural alterations must also occur during the enzymatic reaction which involves several major stereochemical changes in the substrate, cofactor, and enzyme,[5] and are observed in ternary complexes.[2]

Covariant accommodation

The most nonconservative substitution among buried residues is located in the β sheet at the center of strand iii, where Phe-255 in *L. casei* TS is replaced by Gly in *E. coli* (Fig. 13). The surrounding space in the *E. coli* structure is filled by side chains of Trp-253 and Thr-258, versus His-253 and Ala-258 in *L. casei,* for a net change of only one atom in these three residues. Thus while all the neighboring main chain is constant and lies in the LCR (255 is flanked by 4 residues contained within the common core on each side, 251–254 and 258–261), none of the adjacent residues shows shifts in Cα position greater than 2σ.

A second covariant pair is Thr-200, Leu-201 of *L. casei* TS, replaced by Ala-200, Phe-201 in *E. coli* with structural conservation of the surrounding main chain which lies on the 2-fold symmetry axis of the dimer, flanked by residues of the core (Fig. 14).

Segmental accommodation

Helices especially seem to move as a unit (Fig. 15). Further, helices B, C, G are displaced in the same

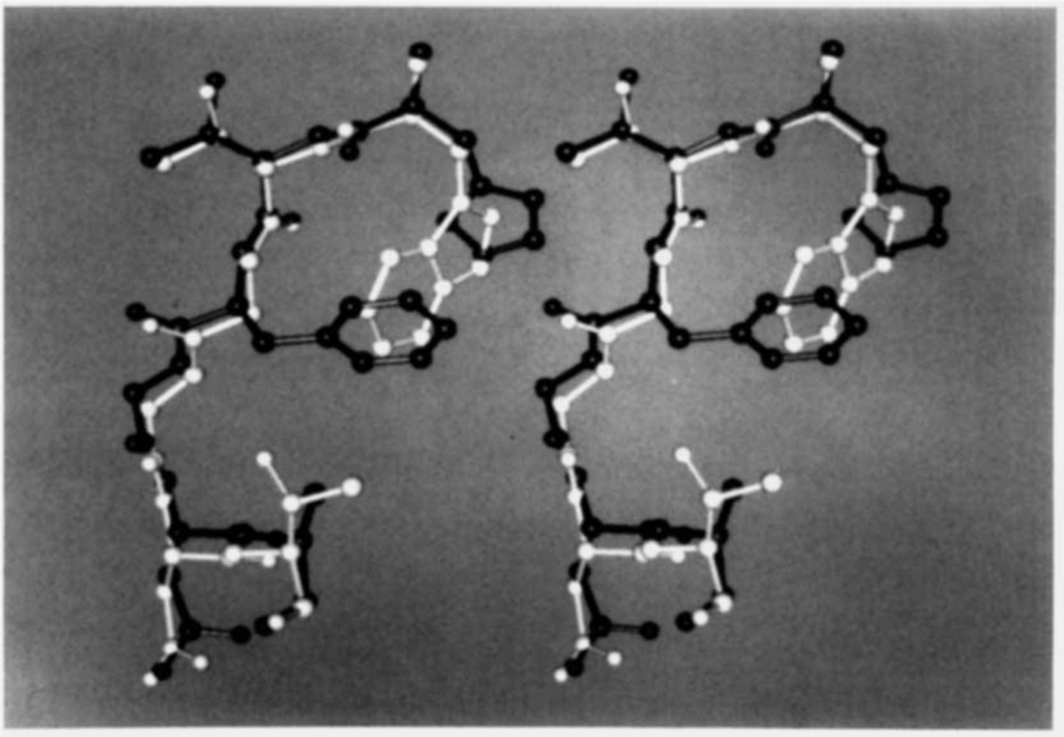


Fig. 13. Segment of β-strand iii depicting the covariant residues Trp 253, Gly 255, and Thr 258 for the *E. coli* (white) and His 253, Phe 255, and Ala 258 for the *L. casei* (black) proteins. (Crosseyed stereo view)

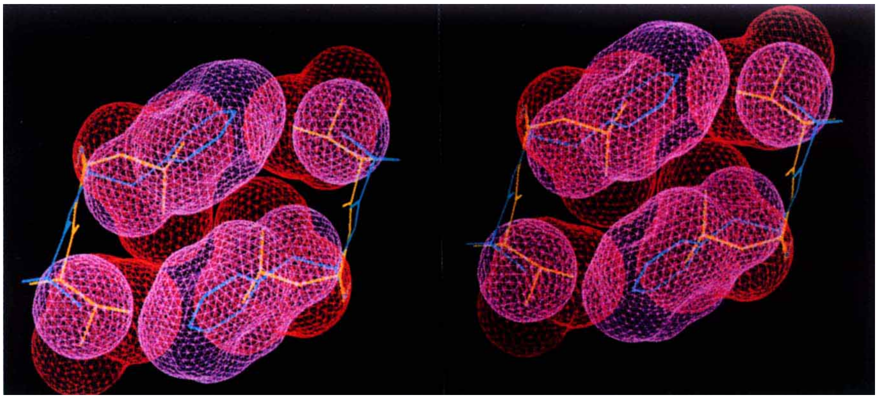


Fig. 14. Segment of strand v depicting covariant residues 200 (Thr and Ala) and 201 (Phe and Leu) for the *L. casei* (yellow) and *E. coli* (blue) proteins. van der Waals surfaces corresponding to *L. casei* and *E. coli* TS structures are colored in red and pink, respectively. (Crosseyed stereo view)

direction while residues 188–194 (helix I) are shifted in a perpendicular direction to that of the B, C, G helix region. This deformation probably reflects interactions of these regions with the 50 amino acid insert in *L. casei* TS, where helix E packs against helices C and D of the small domain (see Fig. 4).

The close packing of highly conserved, mostly hydrophobic buried residues is preserved in these different structures, as demonstrated by the van der Waals surfaces for helices B, C, and G. A representative example of the interdigitation of side chains in this region (Fig. 16) lies in the altered backbone and side chain of Tyr-146 (helix G) which accommodates shifts of Phe-64 (helix B). Many other conserved buried residues including Glu-60, Leu-65, Thr-69, Asn-70, Ile-71, Gly-143, Leu-144, Val-145, Gly-147, Gln-149, and Trp-150 are altered in the

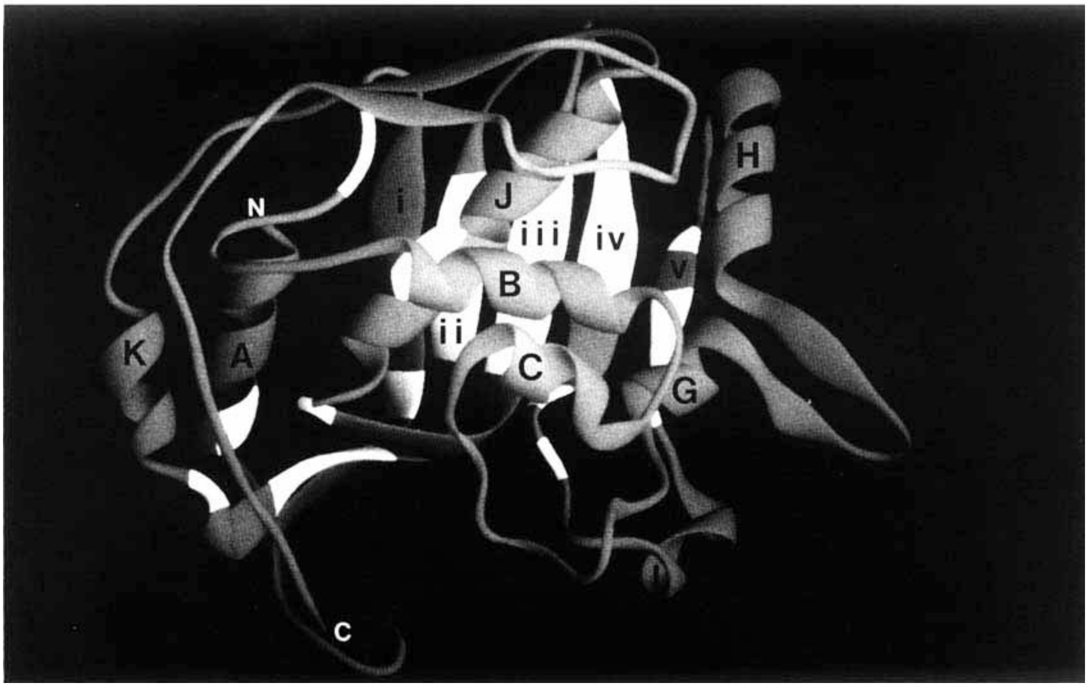


Fig. 15. *E. coli* TS monomer (front view) indicating positions of all structurally nonconserved Cαs (in white) between the *L. casei* and *E. coli* proteins.

plastic response. (Leu-144 is the only nonconserved residue of these, being proline in *E. coli* TS.) Several of these residues (60, 146, 147, and 149) are absolutely conserved among all 17 known TS sequences. Thus it is expected that different species will tolerate structural changes of very similar nature without significant loss of function.

DISCUSSION

Plasticity, the way in which protein molecules accommodate changes in primary structure, is the key to predicting the effect of mutations in proteins. The species comparison of TS has served to map the tolerance of this protein toward mutation. The buried residues demonstrate three mechanisms by which mutations are accommodated:

1. those due to covarions in the common core, in which Cα positions are conserved as multiple side chains combine to fill the same volume,
2. local change whereby large effects observed close to the site of mutation dissipate as a function of distance,
3. systematic shifts, especially in helices, where clusters of tightly packed, nonpolar residues change position in a concerted fashion.

Covariant changes are seen in residues of the β-sheet in the dimer especially, which forms a framework with little change in backbone configuration. However, the Phe-255 to Gly substitution demonstrates that larger changes in sequence can be accommodated by concomitant mutations of His-253 to Trp and Ala-258 to Thr. In contrast, the immunoglobulins, where only three residues are absolutely conserved, demonstrate large changes between β-sheet interfaces as well as local changes in conformation.[33,34] Complementarity among adjacent mutations is rarely seen, though variability of structure may be part of the function in immunoglobulins. TS catalyzes an intricate series of stereochemical changes during the reaction of large substrates. Thus covariant changes provide for an alternate way of preserving the functional degree of dynamic complementarity in TS.

Local flexibility accommodates internal substitutions. This permits relief of strain at the protein surface. Nevertheless the magnitude of differences between these two related sequences is correlated with temperature factors (Fig. 7). The temperature factors parallel the degree of flexibility inside the protein with one exception. In the center of the protein there is a small region for which the shifts are very significantly smaller than predicted by the function of observed B_i [Eq. (3), Fig. 10b]. Thus there is a small nugget of immobilized residues in the central region. This includes 66 atoms from 20 residues, of which 29 are side chain atoms contributed from 13

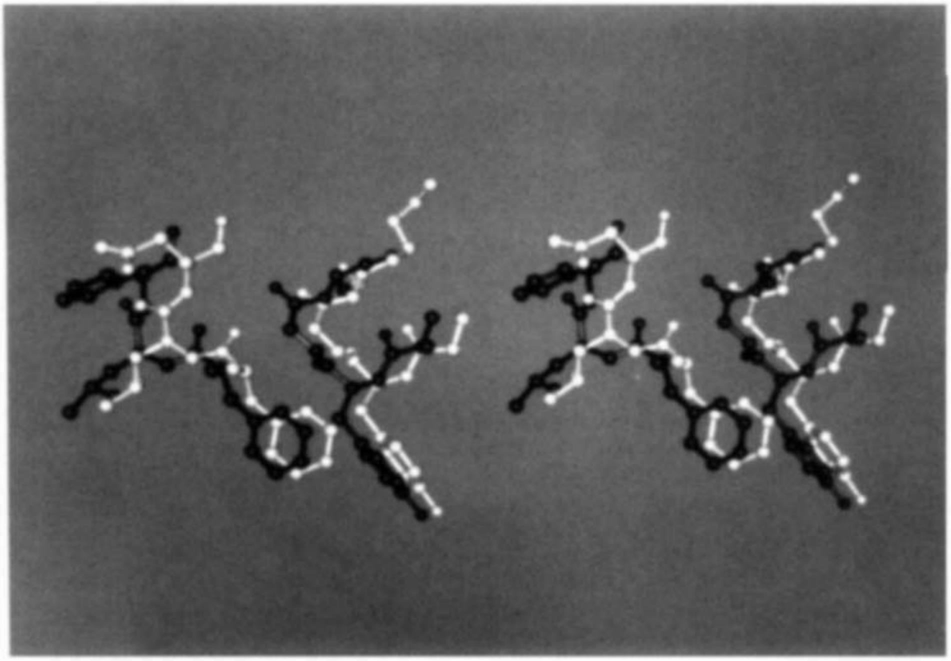


Fig. 16. Shift in positions of Tyr-146 and Phe-64 for the *E. coli* (white) and *L. casei* (black) TS proteins. The structures were overlapped using the difference distance matrix method described in Materials and Methods. (Crosseyed stereo view)

amino acids. The remaining 37 are backbone atoms. It is possible that such three-dimensional nuclei play a role as a template for tertiary folding. This notion can be tested by mutagenesis of residues in this central nucleus. At the protein surface, observed shifts are greater than predicted solely by *B*-factor due to interactions with solvent, crystal contacts, etc.

Concerted alterations are seen in the helices and external loops in TS which contain sites of the largest differences between the *E. coli* and *L. casei* structures. The ensemble movement of the B, C, and G helix region is the most prominent. The maintenance of hydrophobic interactions in this helical region between the two TS proteins represents the preservation of packing constraints during evolution.

In the active center, most of the residues that form hydrogen bonds or contacts with the reactants are structurally conserved and have shifts that lie within the $2\sigma_{xyz}(B_i)$ of the distribution. They were not generally considered in our analysis as they are solvent accessible. However in the reaction these residues undergo a large concerted shift. Presumably part of the high conservation of sequence in TS is to correctly mediate these changes. In other families of proteins, two scenarios concerning active site geometry have been noted. For the globins,[35] plastocyanin and azurin,[36] the binding site and the geometry of the surrounding secondary structure are conserved. In the many enzymes whose structures are based on the α/β barrel found in triose phosphate isomerase for example (about 17 are currently known),[37] and in immunoglobulins the diverse binding sites, comprised of loops rather than elements of secondary structure, are the most variable portion of the molecule,[36,37] as expected.

Changes in sequence are accommodated in unforseen ways as evidenced by unexpected binding modes for different substrates in α-lytic protease[19] and subtilisin,[18] two other cases where high resolution structures of mutants have been compared. Likewise, during the evolution of proteins, the degree to which structural changes can be accommodated in response to single sequence changes affects their rate of acceptance, and the necessity for a second covariant change. However protein plasticity is presumably not an intrinsic property of the amino acids per se; rather it too must have evolved to match the required tolerance in an evolving protein. Thus each protein, or domain may have its own strain coefficients, as determined by evolved instability in the core. TS is one of the most highly conserved proteins. Thus, with the exception of proteins containing one or more disulfide bonds, the strain coefficients in TS represent a lower limit expected for plasticity.

Our goal here is to understand accommodation by deformation (or strain) rather than repacking of side chains, i.e., adjustment within a single potential-well, rather than shifts between two different conformational states in which neighboring residue pairs are different. The latter case is exemplified in multimeric proteins such as the viral coat proteins or hemoglobin that evolved to switch between different packing at subunit interfaces during their function. The magnitude of deformation is related to temperature factors, and also has a clear dependence on distance from a mutation. The expected deformation is described by Eq. (8).

Extensive mutagenesis in the hydrophobic core of λ repressor shows that a net change of $> +2$ or < -3 methylene groups results in nonfunctional protein.[38] Similarly, for TS we have observed that replacements of ≥ 3 heavy atoms (i.e., carbon, nitrogen, and sulfur) must be accompanied by a covariant change nearby (see Fig. 11). We have also shown that the magnitude of shift is uncorrelated with the size of a substitutional change. This implies that both covariant change and plasticity are mechanisms by which substitutions are accommodated through the course of evolution.

Relationship Between Crystallographic Results and Expected Errors

In order to compare related structures it was necessary to correctly predict errors in positional coordinates of the atoms in both crystal structures. There is a clear relationship between thermal factors that are determined in any refined protein structure analysis, and the random differences in position that are inevitable in any particular set of refined coordinates. Such differences will be found, for example, between different refinement pathways for, or different refinemęnt cycles in the same structure. The empirical relationship found for such errors can be useful in analyzing any structure. The relationship between thermal factors and energy well size within which the atom moves, and between energy well size and random errors is however also supported by experimental observations. We refer to these in discussing a general quantitative basis for error prediction.

Connection between thermal factors and energy well size

In highly refined crystal structures for trypsin, Stroud and Chambers found that bonded atoms have B-factors that are related. This is rationalized, modeled, tested, and refined as being due to the rigid body motion of chemical groups, rings, amide planes, etc. (Currently B-factors for connected atoms are sometimes restrained within refinement schemes, urging caution in their direct interpretation from a data bank set.) More importantly this observation directly supports the physical connection between B-factor and energy well sizes in well refined protein X-ray crystal structures.

Connection between energy well size and errors

The B-factor also accommodates another component of random error, since the density map around an atom will be increasingly broad and shallow for an atom in a large energy well. As the width becomes larger and the peak of density drops to the noise level of the map, the placement of the refined atom will be less constrained by the data. Chambers and Stroud[31] derive a smooth empirical relationship between B-factor and positional error, in which the random errors in atomic position generally follow a normal error distribution [the probability of a given error occurring, $P(\delta)$, is proportional to the exponential $\exp(-\delta^2/2\sigma^2)$]. However the mean positional errors E were found to have a dependence generally proportional to B^2 rather than a dependence on the square root of B for each atom, as expected for the mean square amplitude of atomic vibration:

$$<u_i^2> = \frac{B_i}{8\pi^2}$$

Connection between thermal factors and errors

An empirical relationship between probable error in a given crystal analysis and B-factor, similar to that obtained by Chambers and Stroud,[31] pertains to the comparisons made with TS, and with other proteins. The B_i are normally refined along with spatial coordinates (x_i,y_i,z_i), and are a valuable component of the listing of coordinates such as those found in the Brookhaven data bank. If one assumes a behavior similar to those found for TS, these can be translated directly into the expected random errors at any given atom of a highly constrained structure analysis using the empirical relationship of Eq. (5). This provides the expected standard deviation of a normal error distribution, describing positional accuracy of an atom of given thermal factor, B_i, in a structure analysis with given resolution and residual. The dependence on B_i seems to be faithfully reproduced in different structures. The overall scale term may depend on other aspects of an analysis, including such aspects as the constraints applied during refinement, or the presence of disordered regions in the structure. Nevertheless it provides a valuable estimate of the kinds of accuracies to be expected in well-refined, highly constrained structures. This is in contrast to the method of Luzatti,[39] which estimates coordinate errors from the dependence of average R-factor on resolution for the structure. According to the Luzatti analysis, the predicted dependence of R-factor on resolution does not follow that found for the highly refined bovine trypsin structure (1.35 Å resolution: R-factor = 15%, unpublished results). This is due to the assumption that errors for all atoms can be described by a single Gaussian distribution, whereas, as shown here, the width of the gaussian error function is a function of B-factor. Thus, as $\sin\theta$ increases, the relative contribution of atoms with high B-factors decreases and the terms contributing to F_{calc} have smaller errors.

CONCLUSION

The structures of TS from several species have now been solved. Two highly refined structures, one for *E. coli* TS, the other for *L. casei* TS whose solution in a new crystal form, and refinement to higher resolution are reported here, are compared. These structures are from the two most distantly related species of TS whose structures we have solved.

In TS, differences in structure between products of high conservation during evolution reveal plasticity in their common cores, alongside some covariant changes in which substitutions are accompanied by nearby compensatory changes. The plastic changes or examples of molecular tolerance, determined within buried regions, are described by Eq. (8). By implication, the B-factor, which to a harmonic approximation represents the curvature/peak height of electron density, represents the potential energy well in which atoms lie. The disturbance in position of an atom depends exponentially on distance from the site of a change in protein sequence, and is described by a Gaussian distribution whose width increases as a quadratic function of B_i, increasing with B_i^2. Based on a separate comparison of two structure analyses of the identical protein, we conclude that this structural change is over and above expected errors in position in a particular structure analysis, that have a quadratic dependence on B_i of lower magnitude but similar form.

The comparison of shifts after correction for expectation based on distance from the nearest substitution and on B_i shows that they do not depend on the size change of the nearby substitution, though for size changes greater than 3 there is a clear trend to find second covariant substitutions close by. This suggests that in evolution, a certain degree of variation in sequence can be tolerated by plastic adaptation in structure, but that larger substitutions are generally compensated by covariant changes, a contention illustrated by the covariant change Phe-255 to Gly, His-253 to Trp, and Ala-258 to Thr (Fig. 13). Thus among species, based on the TS comparison, the larger the magnitude of an amino acid change, the closer one expects to find a second substitution (Fig. 11).

ACKNOWLEDGMENTS

Supported by National Institutes of Health Grant R01-CA-41323 to J.F.M. and R.M.S. GM24485 to R.M.S., and postdoctoral fellowships from the NIH, GM-11846 (K.P.) and The American Cancer Society S-49-87 (W.M.). Eric B. Fauman is a Howard Hughes Medical Institute Predoctoral Fellow. F. Maley and G.F. Maley are supported by National Cancer Institute CA44355 and National Science Foundation DMB86-16273.

Black and white images were created by Julie Newdoll using software written by Conrad Huang, Eric Petterson, and Greg Couch at the University of California, San Francisco Computer Graphics Lab, and Mike Carson of the University of Alabama, Birmingham.

REFERENCES

1. Hardy, L.W., Finer-Moore, J.S., Montfort, W.R., Jones, M.O., Santi, D.V., Stroud, R.M. Atomic structure of thymidylate synthase: Target for rational drug design. Science 235:448–455, 1987.
2. Montfort, W.R., Perry, K.M., Fauman, E.B., Finer-Moore, J.S., Maley, G.F., Maley, F., Stroud, R.M. Structure, multiple site binding, and segmental accommodation in thymidylate synthase on binding dUMP and an antifolate, CB3717. Biochemistry, in press.
3. Nolan, C., Margoliash, E. Comparative aspects of primary structures of proteins. Annu. Rev. Biochem. 37:727–790, 1968.
4. Dickerson, R.E. The structure of cytochrome c and the rates of molecular evolution. J. Mol. Evol. 1:26–45, 1971.
5. Finer-Moore, J.S., Montfort, W.R., Stroud, R.M. Pairwise specificity and sequential binding in enzyme catalysis; Thymidylate synthase. Biochemistry, in press.
6. Danenberg, K.D., Danenberg, P.V. Evidence for a sequential interaction of the subunits of thymidylate synthetase. J. Biol. Chem. 254:4345–4348, 1979.
7. Donato, H., Jr., Aull, J.L., Lyon, J.A., Reinsch, J.W., Dunlap, R.B. Formation of ternary complexes of thymidylate synthetase as followed by absorbance, fluorescence, and circular dichroic spectra and gel electrophoresis. J. Biol. Chem. 251:1303–1310, 1976.
8. Galivan, J.H., Maley, G.F., Maley, F. The effect of substrate analogs on the circular dichroic spectra of thymidylate synthetase from *Lactobacillus casei*. Biochemistry 15: 356–362, 1976.
9. Ochman, H., Wilson, A.C. Evolutionary history of enteric bacteria. In: *"Escherichia coli* and *Salmonella typhimurium,"* Vol. 2 (Neidhardt, F.C. ed.). Washington, D.C.: American Society for Microbiology, 1987:1649–1654.
10. Maley, G., Maley F. Properties of a defined mutant of *Escherichia coli* thymidylate synthase. J. Biol. Chem. 263: 7620–7627, 1988.
11. Dev, I.K., Yates, B.B., Leong, J., Dallas, W.S. Functional role of cysteine-146 in *Escherichia coli* thymidylate synthase. Proc. Natl. Acad. Sci. U.S.A. 85:1472–1476, 1988.
12. Dev, I.K., Yates, B.B., Atashi, J., Dallas, W.S. Catalytic role of histidine 147 in *Escherichia coli* thymidylate synthase. J. Biol. Chem. 264(32):19132–19137, 1989.
13. Michaels, M.L., Matthews, D.A., Miller, J.H., *Escherichia coli* thymidylate synthase: amino acid substitutions by suppression of amber nonsense mutations. Proc. Natl. Acad. Sci. U.S.A. 87(10):3957–3961, 1990.
14. Climie, S., Ruiz-Perez, L., Gonzalez-Pacanowska, D., Prapunwattana, P., Stroud, R., Santi, D.V. Saturation site-directed mutagenesis of thymidylate synthase. In preparation.
15. LaPat-Polasko, L., Maley, G.F., Maley, F. Properties of T4-phage thymidylate synthase following mutagenic

changes in the folate and phosphate binding regions. J. Biol. Chem. Submitted.

16. LaPat-Polasko, L., Maley, G.F., Maley, F. Properties of T4-phage thymidylate synthase following mutagenic changes in the active site region. Biochemistry. Submitted.
17. Frasca, V., Maley, G.F., Perry, K.M., Stroud, R.M., Maley, F. Properties of a Cys 50 to Phe mutant of *E. coli* thymidylate synthase. In preparation.
18. Wells, J.A., Cunningham, B.C., Graycar, T.P., Estell, D.A. Recruitment of substrate-specificity properties from one enzyme into a related one by protein engineering. Proc. Natl. Acad. Sci. U.S.A. 84:5167–5171, 1987.
19. Bone, R., Silen, J.L., Agard, D.A. Structural plasticity broadens the specificity of an engineered protease. Nature (London) 339:191–195, 1989.
20. Belfort, M., Maley, G., Maley, F. Characterization of the *Escherichia coli thyA* gene and its amplified thymidylate synthase product. Proc. Natl. Acad. Sci. U.S.A. 80:1858–1861, 1983.
21. Wahba, A.L., Friedkin, M. Direct spectrophotometric evidence for the oxidation of tetrahydrofolate during the enzymatic assay. J. Biol. Chem. 236(2):PC11–12, 1961.
22. Howard, A.J., Neilsen, C., Xuong, Ng H. Software for a diffractometer with multiwire area detector. Methods Enzymol. 114:452–472, 1985.
23. Lee, B.K., Richards, F.M. The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. 55:379–400, 1971.
24. Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., Zehfus, M.H. Hydrophobicity of amino acid residues in globular proteins. Science 229:834–838, 1985.
25. Shrake, A., Rupley, J.A. Environment and exposure to a solvent of protein atoms. Lysozyme and insulin. J. Mol. Biol. 79, 351, 1973.
26. Rossmann, M.G. (ed.). "The Molecular Replacement Method." New York: Gordon and Breach, 1972.
27. Crowther, R.A. The fast rotation function. In: "The Molecular Replacement Method" (Rossmann, M.G., ed.). New York: Gordon and Breach, 1972:173–185.
28. Hendrickson, W.A., Konnert, J. In: "Biomolecular Structure, Conformation, Function, and Evolution," Vol. 1 (R. Srinivasan, ed.). Oxford: Pergamon Press, 1981:43–47.
29. Jones, T.A. Interactive computer graphics: FRODO. Methods Enzymol. 115:157–171, 1985.
30. Lewis, C.A., Jr., Munroe, W.A., Dunlap, R.B. Effects of polyoxyanions on sulfhydryl group modification of thymidylate synthetase. Biochemistry 17(4):5382–5387, 1978.
31. Chambers, J.L., Stroud, R.M. The accuracy of refined protein structures: Comparison of two independently refined models of bovine trypsin. Acta Crystallogr. B35: 1861–1874, 1979.
32. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
33. Lesk, A.M., Chothia, C. Evolution of proteins formed by β-sheets II. The core of the immunoglobin domains. J. Mol. Biol. 160:325–342, 1982.
34. Huber, R., Bennett, W.S., Jr. Functional significance of flexibility in proteins. Biopolymers 22:261–279, 1983.
35. Lesk, A.M., Chothia, C. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. J. Mol. Biol. 136:225–270, 1980.
36. Chothia, C., Lesk, A.M. Evolution of proteins formed by β-sheets: I. Plastocyanin and azurin. J. Mol. Biol. 160:309–323, 1982.
37. Lesk, A.M., Branden, C.I., Chothia, C. Structural principles of alpha/beta barrel proteins: The packing of the interior of the sheet. Proteins. 5(2):139–148, 1989.
38. Lim, W.A., Sauer, R.T. Alternative packing arrangements in the hydrophobic core of λ repressor. Nature (London) 339:31–36, 1989.
39. Luzatti, V. Statistical treatment of errors in the determination of crystal structures. Acta Crystallogr. 5:802–810, 1952.
40. Bellisario, R.L., Maley, G.F., Guarino, D.U., and Maley, F. The primary structure of *Lactobacillus casei* thymidylate synthetase: II. The complete amino acid sequence of the active site peptide, CNBr 4. J. Biol. Chem. 254:1296–1300, 1979.
41. Maley, G.F., Bellisario, R.L., Guarino, D.U., Maley, F. The primary structure of *Lactobacillus casei* thymidylate synthetase: I. The isolation of cyanogen bromide peptides 1 through 5 and the complete amino acid sequence of CNBr 1, 2, 3, and 5. J. Biol. Chem. 254:1288–1295, 1979a.
42. Maley, G.F., Bellisario, R.L., Guarino, D.U., Maley, F. The primary structure of *Lactobacillus casei* thymidylate synthetase: III. The use of 2-(2-nitrophenyl sulfenyl)-3-bromoindolenine and limited tryptic peptides to establish the complete amino acid sequence of the enzyme. J. Biol. Chem. 254:1301–1304, 1979b.
43. Belfort, M., Maley, G., Pedersen-Lane, J., Maley, F. Primary structure of the *Escherichia coli thyA* gene and its thymidylate synthase product. Proc. Natl. Acad. Sci. U.S.A. 80:4914–4918, 1983.
44. Iwakura, M., Dawata, M., Tsuda, K., Tanaka, T. Nucleotide sequence of the thymidylate synthase B and dihydrofolate reductase genes contained in one *Bacillus subtilis* operon. Gene 64:9–20, 1988.
45. Takeishi, K., Kaneda, S., Ayusawa, D., Shimizu, K., Gotoh, O., Seno, T. Nucleotide sequence of a functional cDNA for thymidylate synthase. Nucleic Acids Res. 13: 2035–2043, 1985.
46. Perryman, S.M., Rossana, C., Deng, T., Vanin, E.F., Johnson, L.F. Sequence of a cDNA for mouse thymidylate synthase reveals striking similarity with the prokaryotic enzyme. Mol Biol Evol. 3:313–321, 1986.
47. Honess, R.W., Bodemer, W., Cameron, K.R., Niller, H.-H., Fleckenstein, B., Randall, R.E. 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, 1986.
48. Richter, J., Puchtler, I., Fleckenstein, B. Thymidylate synthase gene of herpesvirus ateles. J. Virol. 62:3530–3535, 1988.
49. Beverley, S.M., Ellenberger, T.E., Cordingley, J.S. Primary structure of the gene encoding the bifunctional dihydrofolate reductase-thymidylate synthase of *Leishmania major*. Proc. Natl. Acad. Sci. U.S.A. 83:2584–2588, 1986.
50. Grumont, R., Washtein, W.L., Santi, D.V. 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, 1986.
51. Bzik, D.J., Li, Wu-bo, Horii, T., Inselburg, J. Molecular cloning and sequence analysis of the *Plasmodium falciparum* dihydrofolate reductase-thymidylate synthase gene. Proc. Natl. Acad. Sci. U.S.A. 84:8360–8364, 1987.
52. Thompson, R., Honess, R.W., Taylor, L., Morran, J., Davison, A.J. Varicella-zoster virus specifies a thymidylate synthase. J. Gen. Virol. 68:1449–1455, 1987.
53. Taylor, G.R., Lagosky, P.A., Storms, R.K., Haynes, R.H. Molecular characterization of the cell cycle-regulated thymidylate synthase gene of *Saccharomyces cerevisiae*. J. Biol. Chem. 262:5298–5307, 1987.
54. Edman, U., Edman, J.C., Lundgren, B., Santi, D.V. Isolation and expression of the *Pneumocystis carinii* thymidylate synthase gene. Proc. Natl. Acad. Sci. U.S.A. 86(17): 6503–6507, 1989.
55. Singer, S.C., Richards, C.A., Ferone, R., Benedict, D., Ray, P. Cloning, purification, and properties of *Candida albicans* thymidylate synthase. J. Bacteriol. 171:1372–1378, 1989.
56. Chu, F.K., Maley, G.F., Maley, F., Belfort, M. Intervening sequence in the thymidylate synthase gene of bacteriophage T4. Proc. Natl. Acad. Sci. U.S.A. 81:3049–3053, 1984.
57. Kenny, E., Atkinson, T., Hartley, B.S. Nucleotide sequence of the thymidylate xynthetase gene (ThyP^3) from the *Bacillus subtilis* phage Φ3T. Gene 34:335–342, 1985.
58. Hughes, D.E., Shonekan, O.A., Simpson, L. Structure, genomic organization and transcription of the bifunctional

dihydrofolate reductase-thymidylate synthase gene from *Crithidia fasciculata*. Mol. Biochem. Parisitol. 34:155–166, 1989.

59. Rouch, D.A., Messerotti, L.J., Loo, L.S.L., Jackson, C.A., Skurry, R.A. Trimethoprim resistance transposon *Tn*4003 from *Staphylococcus aureus* encodes genes for a dihydrofolate reductase and thymidylate synthetase flanked by three copies of IS257. Mol. Microbiol. 3(2):161–175, 1989.