

Atomic Structure of Thymidylate Synthase: Target for Rational Drug Design

LARRY W. HARDY, JANET S. FINER-MOORE, WILLIAM R. MONTFORT,
MELVIN O. JONES, DANIEL V. SANTI, ROBERT M. STROUD*

The atomic structure of thymidylate synthase from *Lactobacillus casei* was determined at 3 angstrom resolution. The native enzyme is a dimer of identical subunits. The dimer interface is formed by an unusual association between five-stranded β sheets present in each monomer. Comparison of known sequences with the *Lactobacillus casei* structure suggests that they all have a common core structure around which loops are inserted or deleted in different sequences. Residues from both subunits contribute to each active site. Two arginine side chains can contribute to binding phosphate on the substrate. The side chains of several conserved amino acids can account for other determinants of substrate binding.

THY MIDYLATE SYNTHASE (TS) (E.C. 2.1.1.45) CATALYZES the conversion of deoxyuridine monophosphate (dUMP) and 5,10-methylenetetrahydrofolate ($\text{CH}_2\text{H}_4\text{folate}$) to deoxythymidine monophosphate (dTMP) and dihydrofolate. It provides the sole de novo pathway for biosynthesis of dTMP and is the only enzyme of folate metabolism in which the $\text{CH}_2\text{H}_4\text{folate}$ is oxidized during one-carbon transfer. Because of the central role of TS in the synthesis of an essential DNA precursor, and its importance as a chemotherapeutic target, the enzyme has been much studied (1).

As obtained from sources as varied as bacteria, bacteriophage, yeast, viruses and vertebrates, TS is a dimer of identical subunits of about 35 kilodaltons (kD) each. In protozoa, TS is a bifunctional protein with dihydrofolate reductase (DHFR) on the same polypeptide chain (2). Primary sequences of TS from eight different sources reveal that the protein is even more conserved than the cytochromes, with a unit evolutionary period of 22.9 million years (3), if we assume a constant rate of mutation and ignore insertions and deletions in the sequences.

The catalytic mechanism of TS involves the initial formation of a covalent bond between the 6-position of dUMP and a catalytic thiol of the enzyme (Fig. 1). Formation of the covalent adduct 1 serves to activate the 5-position of the heterocycle for condensation with the one-carbon unit of the cofactor. Subsequent to formation of the ternary covalent complex 2, β elimination and hydride-transfer reactions occur to generate the methylated pyrimidine, the oxidized cofactor, and the catalytically active enzyme. Thymidylate synthase is of additional importance because it serves as a paradigm for the mechanism and inhibition of several related enzymes that catalyze one-carbon transfer to pyrimidines (4).

Binding requirements of substrates and analogs have been studied extensively, and there are several extremely potent inhibitors of

TS. Of particular note is 5-fluorodeoxyuridine monophosphate (FdUMP), an active metabolite of the antineoplastic agent 5-fluorouracil, which is a mechanism-based inhibitor of TS and forms a stable covalent adduct analogous to the steady-state intermediate 2 (5).

In 1974 we reported the first diffraction patterns of TS crystals although those crystals were not suitable for high resolution structure analysis (6). Here, we describe the high resolution structure of *Lactobacillus casei* TS.

Crystallization. TS was isolated (7) from a methotrexate-resistant strain of *L. casei* (8). Crystals were grown by vapor diffusion against 20 mM potassium phosphate buffer (pH 6.8) at room temperature from TS at 10 mg/ml in the same buffer containing 15 to 20 mM ammonium sulfate with or without 5 mM dUMP. These crystals diffract to 2.4 Å resolution and are hexagonal with space group $P6_122$ and unit cell dimensions $a = 78.8$ Å and $c = 230.1$ Å. Neither dUMP nor dithiothreitol had any detectable effect on the diffraction patterns. The crystalline enzyme was fully active upon redissolution. Crystal density was measured by flotation in organic solvent (9) as $\rho = 1.16$ g/cm³, thus the asymmetric unit contains one monomer and the crystals contain 54 percent solvent. Crystals occasionally convert to a second form with an 11 Å longer c dimension ($a = 78.5$ Å and $c = 240.9$ Å), which correlates with exposure to temperatures above 23°C.

Structure solution. Initial 3.5 Å diffraction data of TS crystals in 20 mM inorganic phosphate (P_i) were measured on a four-circle diffractometer with the use of graphite monochromatized copper $\text{K}\alpha$ radiation (10) and by the scan technique of Wyckoff *et al.* (11). Of numerous heavy atom derivatives screened, only four were used for multiple isomorphous replacement (MIR) phase determination (Table 1). A 5 Å map calculated with MIR phases from the four derivatives showed the boundary of the two monomers as well as elements of secondary structure.

The data collected on the diffractometer were used only for solving derivatives and for calculation of the 5 Å map. For higher resolution phasing, 3 Å data were collected from TS crystals grown in 5 mM dUMP containing 20 mM P_i and from corresponding *p*-chloromercuribenzoate (pCMB) and $\text{Pt}(\text{CN})_4$ derivatives on a position-sensitive multiwire area detector (12). Independent reflections had average redundancies of 8.5, 4.3, and 3.5 times, respec-

L. W. Hardy and D. V. Santi are in the Departments of Pharmaceutical Chemistry and Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA 94143-0446. J. S. Finer-Moore, W. R. Montfort, M. O. Jones, and R. M. Stroud are in the Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA 94143-0448. The present address of L. W. Hardy is Department of Molecular Genetics and Microbiology, University of Massachusetts Medical Center, Worcester, MA 01605.

*To whom all correspondence should be addressed.

tively, for the TS, TS-pCMB, and TS-Pt(CN)₄ data sets. Friedel pairs for the derivatives were treated independently during data reduction for use of anomalous dispersion. Weighted

$$R_{\text{syms}} = \left\{ \sum_{\text{hkl}} \left(\frac{\sum_i [(I_{\text{avg}} - I_i)/\sigma(I_i)]^{**2}}{\sum_i [I_{\text{avg}}/\sigma(I_i)]^{**2}} \right)^{**1/2} \right\}$$

for the three data sets were 6.75, 7.11, and 6.38 percent, where I_i is a measured intensity, $\sigma(I_i)$ is the standard deviation of that measurement, and I_{avg} is the average over the redundant intensity measurements. Figure 2 summarizes the statistics for 3 Å MIR phase determination.

Although the map calculated with the 3 Å MIR phases had some interpretable features, the amino acid sequence (13) could not be traced in it. Single isomorphous replacement phases from the pCMB derivative were therefore refined with the constraints of flat solvent density and positive protein density. The protein boundary was determined by definition of regions of the 3 Å resolution map that had positive density in the 5 Å MIR map (14) and by the algorithm of Wang (15). The phases from the latter procedure gave the more interpretable 3 Å map. Seventy percent of the peptide backbone was fitted to this map. Subsequently, MIR phases were refined by the same solvent flattening procedure, and also combined in real space with the partial model. For the real space combination, spheres of density were substituted for MIR density at the atoms in the model and the hybrid density was back-transformed. Refined MIR phases and phases from density combination both produced maps used to build the model. Recognition of two key elements in the protein density helped to fit the amino acid sequence to the map. The first was the NH₂-terminal sequence, which was identified as an amphipathic helix (16), and was fitted to a surface helix in the electron density map. The second key element was the active site sulfhydryl of Cys¹⁹⁸ which was located adjacent to one of the heavy atom sites of the pCMB derivative.

Maps from reciprocal-space phase combination (17), Fourier refinements (18), and least-squares refinements (19) were also used in the latter stages of model building. These maps clearly define the positions and orientation of 278 of the 316 residues in the structure; residues 90 to 120 and 301 to 307 seem to be disordered although parts of the disordered region may become interpretable during refinement. The crystallographic residual is $R = 31$ percent, and the standard deviations of bond distances and angles from standard values are 0.013 Å and 2.6° for our model.

Overall structure. TS is a dimer of identical subunits related by a crystallographic twofold symmetry axis (Fig. 3). A five-stranded β sheet within each monomer forms the dimer interface. Whereas the handedness of the twist of the β sheets in each monomer is normal, the right-handed twist of one sheet relative to the other is unprecedented (Fig. 4) (20). The dihedral angle between strand directions is +28°, and the observed value for other sheets is typically about -30°. Of the 29 residues per monomer which lie in the interface (21), half are from β strands, and of these all but two are completely solvent-inaccessible in the dimer. The remaining interface residues are from interconnecting loops, most of which are partially solvent-accessible. The total buried surface area, 1211 Å², in the interface, is typical for dimeric proteins (22). Eleven (40 percent) of the contact residues are very hydrophilic and include four glutamines and three arginines. So far, five probable hydrogen bonds across the interface have been identified.

The tertiary fold of the *L. casei* TS monomer is shown in Fig. 5. The subunit is composed of two domains. The larger domain, residues 1 to 69 and 140 to 316, has five α helices and two 3₁₀ helices (helices A, B, and G to K in Fig. 5), and the five-stranded β sheet. These secondary structure elements are arranged in three

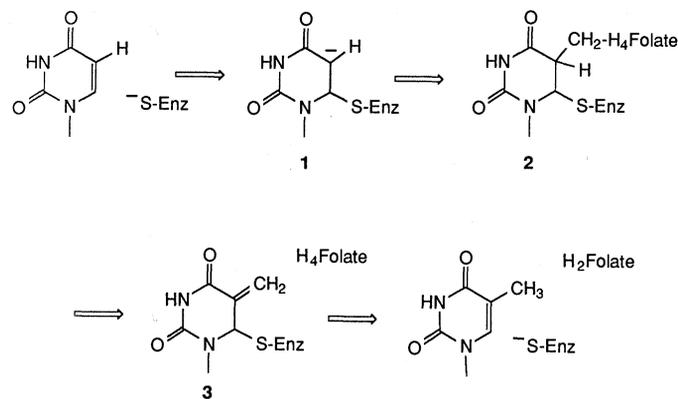


Fig. 1. The proposed mechanism for TS.

layers, but unlike most α/β domains, the central layer is composed of helices (23). The smaller “variable” domain consists of residues 70 to 139, which includes α helices C and D (α-C, D), a disordered region between α-D and α-E (residues 90 to 120) and α-E, F. The two domains are separated by a cleft (Fig. 3), the hinge of which is defined by residues of α-B. Contacts between α-B and Leu⁷⁴ of α-C define the only interface between the domains.

The larger domain has several notable features.

1) There is the five-stranded, predominantly antiparallel β sheet that forms the interface between monomers. Aligned discontinuities in strands i, ii, and iii (at Gly³³, Ala²⁵⁸, and Gln²¹⁷) twist the bottom left corner of the sheet outward (Fig. 4). Although distorted, hydrogen bonding between the three anomalous residues is preserved through the bend.

2) There is a six-turn unusually hydrophobic α helix, J, a core structural element around which the large domain folds. This helix is

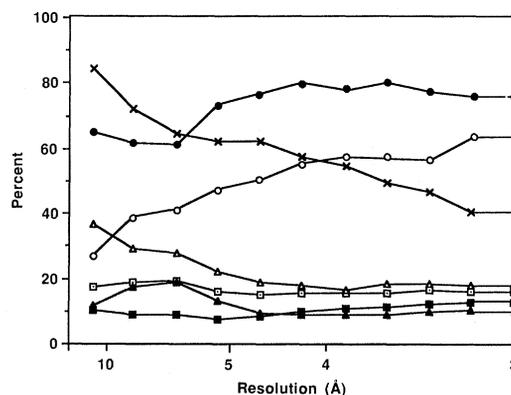


Fig. 2. MIR statistics for 3 Å phases of TS grown in a mixture of 5 mM dUMP and 20 mM P_i. The pCMB and Pt(CN)₄ derivatives, which had previously been solved with diffractometer data (Table 1), were refined with data collected on the area detector from crystals grown in the presence of dUMP. Four sites with occupancies of 0.20, 0.20, 0.14, and 0.12 were located for the Pt(CN)₄ derivative. Two major pCMB sites had occupancies of 0.49 and 0.43. Two minor pCMB sites (occupancies of 0.16 and 0.09) were used to account for residual density around the highest occupancy site. The residual density was probably the result of anisotropy or doubling of the site. Plotted are:

- pCMB $\langle |\Delta F| \rangle / \langle |F| \rangle$
- pCMB $\langle |\epsilon| \rangle / \langle |\Delta F| \rangle$
- Pt(CN)₄ $\langle |\Delta F| \rangle / \langle |F| \rangle$
- ▲ Pt(CN)₄ 0.1 * $\langle |f_h| \rangle / \langle |\epsilon| \rangle$
- x figure of merit (46),

where $\langle |f_h| \rangle$ is the average magnitude of the heavy atom structure factors, $\langle |F| \rangle$ is the average magnitude of the protein structure factors, $\langle |\Delta F| \rangle$ is the average difference in magnitude between the native and derivative structure factors, and $\langle |\epsilon| \rangle$ is the average lack of closure error, that is, the difference in magnitude between calculated and observed structure factors for the derivative. For the figure of merit calculation, the root-mean-square estimate of error in the derivative data was calculated from centric reflections only.

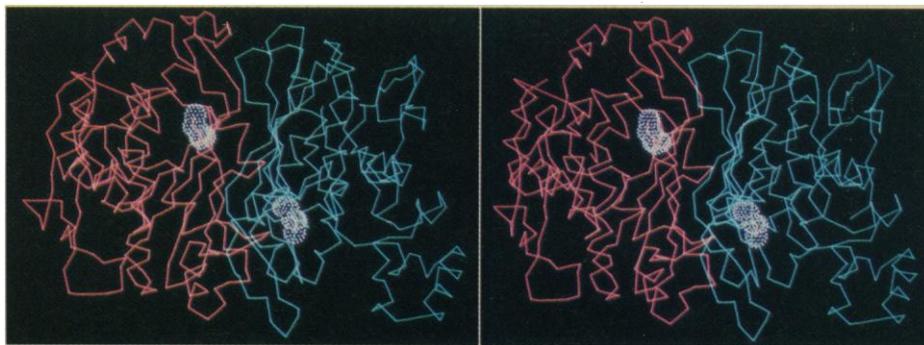


Fig. 3. Convergent (crossed-eye) stereoscopic view of the α -carbon tracing of the TS dimer in which each monomer is colored differently. Such stereo pictures can be viewed without stereo glasses by first looking at a pencil point held midway between the eyes and the figure, then moving the pencil toward and away from the eyes until the images of structure fuse together, when the pencil can be removed. The view is approximately 90° away from the schematic drawings shown in Figs. 5 and 7, and is viewed approximately down the twofold axis. The small domain is clearly visible to the right of the right side monomer. For a divergent stereoscopic view, the type viewed through stereo glasses, the two sides of the figure can be cut out and their positions switched.

even more hydrophobic than some membrane spanning helices (24). It is inaccessible to water except on one face of the NH_2 -terminal end. The helix packs diagonally across the β sheet, and it is flanked by amphipathic α -A and α -H, which are on the edge of the middle layer, and by α -B which packs on top of α -J on the side opposite from the β sheet.

3) There is a shallow cavity embedded in the large domain and this cavity contains the active site of the enzyme. The left wall of the cavity is formed by the sharply turning part of β strands ii and iii (β -ii and β -iii), and the loop between α -A and β -i. The ceiling is composed of residues from the exposed NH_2 -terminus of α -J, and the right wall by the end of β strand iv. The back of the cavity is formed by four residues of the β sheet of the other monomer, one of which (Arg^{179}) (25) protrudes directly into the cavity. Between α -B and the β sheet there is an unusual cluster of seven aromatic side chains adjacent to the cavity. Although α -G and α -H are not linked by hydrogen bonds, they are nearly aligned and appear as a single kinked helix interrupted by a loop, which is stabilized by a salt bridge between Arg^{151} and Asp^{162} . Trp^{150} protrudes into the back of the aromatic cluster, pointed toward the active site cavity. Continuation of α -G would put this residue in steric conflict with α -B; thus it may initiate loop G-H. Finally, the variable domain lies on the edge of the cavity, but does not directly contribute to its structure.

Structural equivalence. The sequences of eight different species of TS (Fig. 6) are among the most highly conserved of all proteins,

which implies that they have similar tertiary structures. The eight sequences have 55 invariant residues and 38 residues that show only conservative changes (26) between all species; if the $\phi 3\text{T}$ sequence is omitted, there are 91 invariants (27). None of the secondary structure elements have a corresponding sequence with an insertion in another species of TS. Therefore, there is a common "core" structure which, we expect, will be found in other TS's.

Invariant residues are likely to be those most crucial for the structure and function of the enzyme. Sixteen invariants line the active site cavity and may be important for substrate binding and catalysis. Ten invariant residues are glycine or proline, which because of their ability to adopt different conformational angles from other amino acids are often structural determinants. Conservation of the remaining residues in the sequence generally correlates with the degree to which each is involved in packing within the structure, and is sequestered from solvent. This correlation is expected if packing within a common tertiary structure is preserved in all TS's. The most conserved regions make up the hydrophobic protein core. These include α -G, α -J, the COOH -terminal half of α -B, and the hydrophobic surface of the amphipathic α -H. The sequences of each of these regions show 86 to 100 percent conservation. The five strands of the β sheet are somewhat less conserved than the hydrophobic core helices, although residues involved in the dimer interface are very well conserved: 11 of 15 are either invariant or have a single difference among the eight known

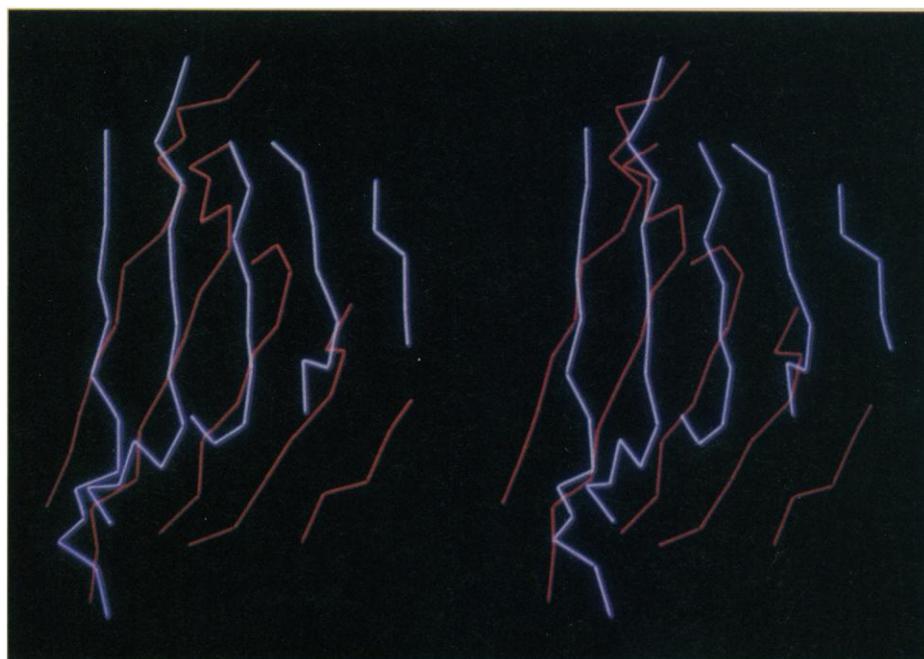


Fig. 4. Convergent stereo view of the two β sheets forming the dimer interface. The back sheet (in red) is rotated clockwise with respect to the front sheet, an orientation opposite to that normally seen in β sheet interfaces. Breaks in β strands i, ii, and iii (on the left side of the figure) bend one corner of each sheet sharply away from the interface. The orientation is essentially the same as in Fig. 5.

sequences. Also highly conserved are loops on either side of β strand v which lie in the dimer interface, and the hydrophobic loop (residues 44 to 51) preceding α -B, implicated in binding folyl polyglutamate. Least conserved in the large domain are the external surface residues of the amphipathic α -A, α -H, and α -K, and the "northern" loops [40 to 43 (57 percent), 206 to 211 (73 percent), 274 to 316 (62 percent conserved)] at the top of Fig. 5, which by virtue of the position of the twofold axis are external in the dimer. The "southern" loops, at the bottom of Fig. 5 (22 to 26, 81 to 83, 185 to 188, 195 to 198), are more intimately involved in the dimer interface and the surface of the active site cavity and are highly conserved (88 to 100 percent conserved).

The invariance of residues that are not packed within the core, tight turns, or the active site cavity suggests that they may have functions other than the enzymatic reaction that have not yet been recognized. These include Thr⁴⁹, Lys¹⁷², Tyr³⁰⁴, and Ala³¹⁵ within the large domain. The variable domain is poorly conserved yet, remarkably, the loop immediately following α -C contains three almost invariant residues 81 to 83, -Ile-Trp-Asp-, that precede α -D. Residues of α -D fall into two families; they are almost invariant (83 percent) in eukaryotes, but are entirely different in the bacterial and phage sequences that are 66 percent homologous among themselves. Herpesvirus saimiri (HVS) is not a eukaryote, but infects a eukaryotic species, and its sequence is more homologous to the eukaryotic sequences than to the prokaryotic and phage sequences of TS.

Inserts and heterologous regions. There are six sites where large inserts are found around the TS core (Figs. 5 and 6). They are all in surface loops or in the variable domain and probably do not influence the core structure. Since these are not present in all species, they cannot be relevant to fundamental aspects of structure or function, but may serve roles specific to individual organisms.

The insertion of residues 90 to 139 (70 percent of the variable domain) occurs only in *L. casei* TS, and it includes the disordered stretch 90 to 120. However, even when the entire *L. casei* insert is deleted, as it is in *Escherichia coli* or ϕ 3T TS, the remaining COOH-

Table 1. Statistics for four derivatives of TS crystals grown in 20 mM P_i. Heavy atom sites were determined from difference Patterson and difference Fourier maps. The number of sites, highest resolution data collected, number of crystals used to collect the data, and agreement with native data are reported. The Pt(NH₃)₂Cl₂ derivative was unstable in the x-ray beam. The heavy atom positions for the pCMB and MeHg⁺ derivatives were the same although the sites had different relative occupancies.

Derivative	Crystals (No.)	Highest resolution data (Å)	R (%)	Sites (No.)
pCMB	3	3.5	19.2	2
<i>cis</i> -Pt(NH ₃) ₂ Cl ₂	1	7.0	13.3	1
Pt(CN) ₄	2	4.0	12.1	3
MeHg ⁺	1	6.3 (partial)	17.6	2

and NH₂-termini (positions 89 and 140) are in close proximity so that a connection could be made without major perturbations in the core structure.

Eukaryotic and HVS TS's have inserts of 12 and 8 amino acids at 90 to 101 and 156 to 157, respectively. Both are too peripheral to be involved in substrate binding or catalysis (Fig. 5). These inserts are invariant in size, and show quite high sequence conservation (94 and 88 percent) suggesting that they may have an as yet undiscovered eukaryote-specific function. For example, *Leishmania* TS undergoes rapid and specific proteolysis at the COOH-terminus of the loop 90 to 101 (28). Cleavage of this external loop inactivates TS but not DHFR. Perhaps this is important for the regulation of *Leishmania* TS. In the other eukaryotes, such "suicide" sequences could have similar roles in controlling enzyme activity.

The two bacteriophage sequences have insertions of 7 to 12 amino acids at peripheral sites, for ϕ 3T between 21 and 22, and for T4 between 74 and 75, 90 to 97, and 294 and 295. There are also small insertions between 43 and 44 and 209 and 210, specific only to *Leishmania*.

Splice junctions. The genomic sequences for mouse TS contains

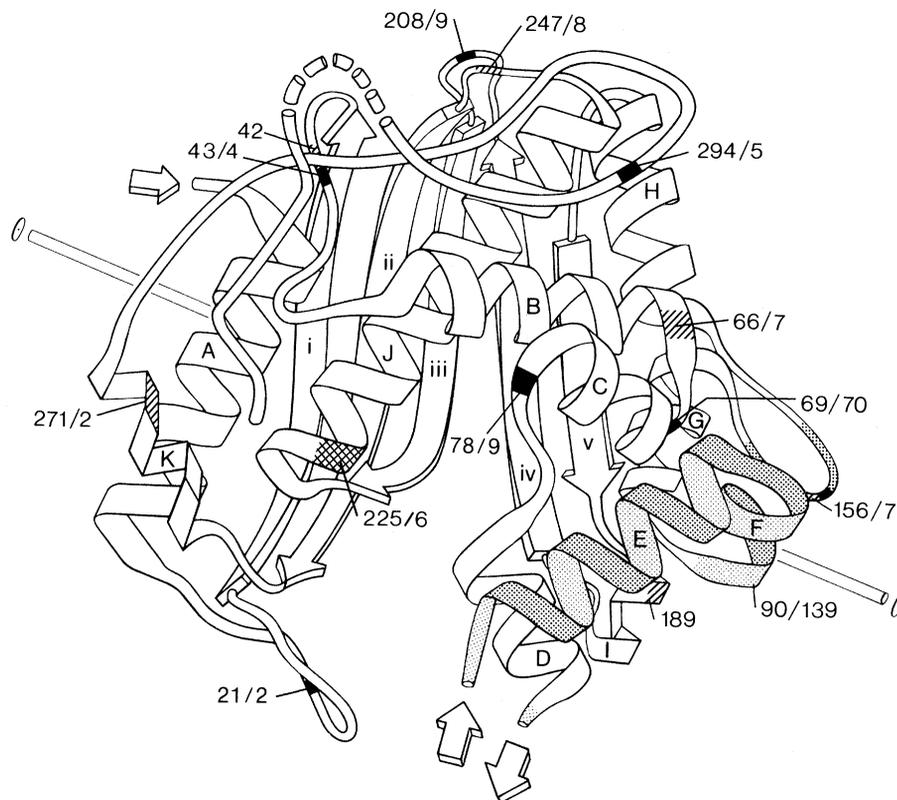


Fig. 5. A schematic of the TS structure drawn according to Richardson (47) showing the arrangement of helices and extended chain within one monomer. The position of the crystallographic twofold axis is indicated by the diagonal rod behind the monomer. The NH₂-terminal residue 1 is indicated by an arrow at top left. The helices are labeled in alphabetical order from NH₂- to COOH-terminus and the strands of β sheet are numbered in order from left to right. Positions of the six introns in the structural gene for mouse TS (dashed line), and the noncoding region (cross-hatched) within the T4 genome are indicated on the schematic with sequence numbers indicating the position of the insert. Locations of amino acid insertions (solid black) and deletions (stippled) in other sequences relative to that of *L. casei* and flanking residue numbers are indicated. The broken chain at 301 to 307 and the break between 89 and 121 indicated by arrows correspond to disordered regions in the structure.

six introns that separate the coding sequence into seven exons (29) (Fig. 5). All amino acids at the exon splice junctions map to the surface of the protein. Nine of the ten splice junction residues are conserved between the only two mammalian TS sequences known although they lie in regions that are often charged and have substitutions in other species. Thus, the splice junction positions occur within regions where insertions or deletions in the protein sequence are perhaps more easily accommodated (30).

Only one of the splice junctions, that between 156 and 157, coincides with a variable length region. It lies within the hairpin loop which connects α -G and α -H. Variability at this splice site may have contributed to tolerable deletions and insertions in the protein sequence.

All of the residues at the splice junctions map to loops between secondary structural elements or lie at the termini of helices, a result generally consistent with the proposal that separate structural units

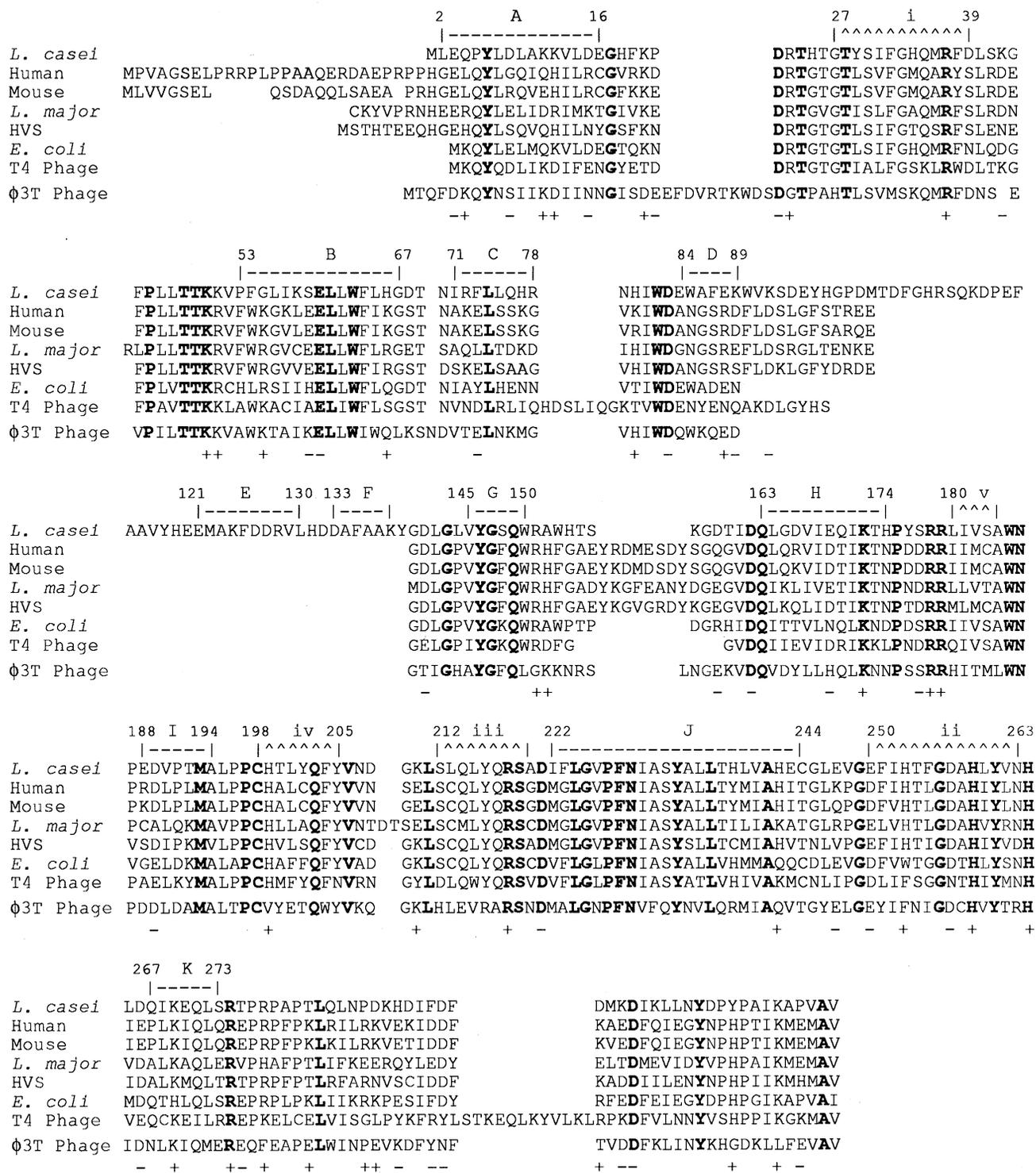
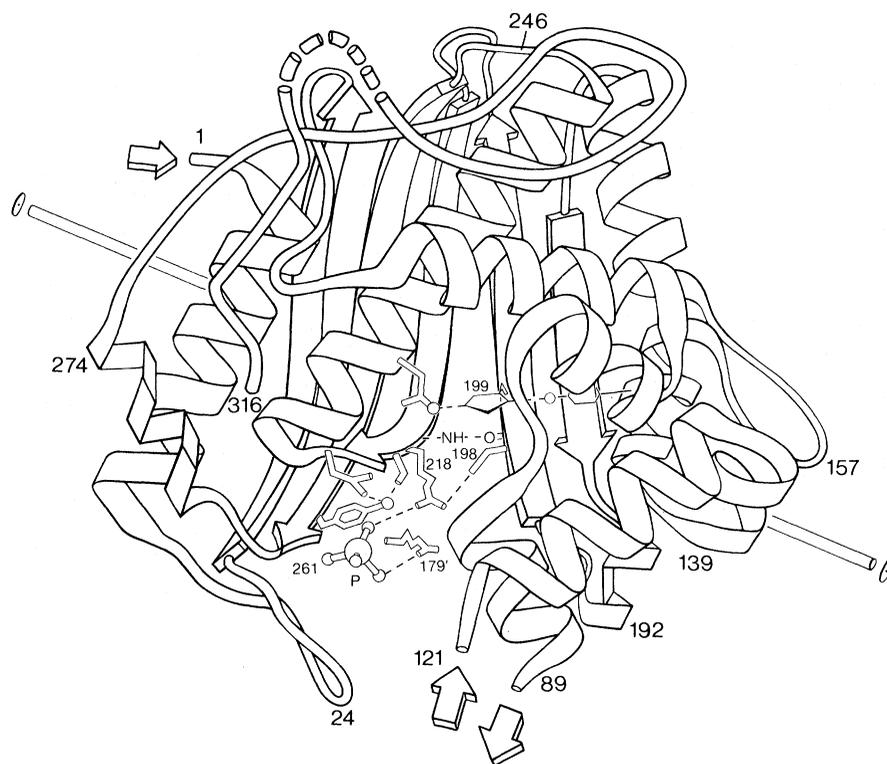


Fig. 6. The aligned sequences of eight TS species. Invariant residues are in boldface. Charged groups (+, -) in more than four species are indicated beneath the sequences. At positions where both positively and negatively charged groups are present, the charge which appears most often is used. The numbering scheme and secondary structure designations ($\wedge/\wedge/\wedge = \beta$

sheet, ----- = helix) marked above the sequences refer to TS from *L. casei*. Labeling of secondary structure elements follows the convention of Fig. 5. TS sequences shown include those for *L. casei* (13), human (48), mouse (49), the parasite *Leishmania major* (50), Herpesvirus saimiri (51), *E. coli* (52), coliphage T4 (T4) (32), and *Bacillus subtilis* phage ϕ 3T (ϕ 3T) (53).

Fig. 7. Schematic of the TS structure showing several conserved residues in the active center of the monomer, and Arg^{179'} of the neighboring monomer.



lie in separate exons (31). The first exon (1 to 42) encodes the amphipathic helix α -A and β -i. The second exon (43 to 66) encodes the α helical polypeptide that traverses the molecule and includes the polyglutamate binding regions, α -B and the loop preceding α -B. The third exon (67 to 156) encodes α -G of the large domain, and the entire variable domain. Exon four begins in the variable G-H loop and codes the right-hand flanking α -H and β -v segment (156 to 189); it contains amino acids that impinge on the other subunit, including Arg^{179'}, which contributes to phosphate binding. The fifth exon (189 to 247) encodes the central hydrophobic α helix J and two adjacent antiparallel β strands, and contains many of the residues in the active site cavity including Cys¹⁹⁸ and Arg²¹⁸. The sixth exon (248 to 271) encodes β -ii and β -iii helix K, and the seventh exon codes for the extended COOH-terminal loop structure.

The TS gene from T4 is the only prokaryotic gene thus far sequenced that contains an intron (32). The position of the T4 splice junction does not coincide with any of those in the mouse TS gene; it maps to a residue on the protein surface within the NH₂-terminus of the hydrophobic core element α -J, one of the most conserved and important structural elements of TS (Fig. 5). Either the intervening sequences in each gene originated independently, or a primordial TS gene lost introns differentially in the two evolutionary lines leading to mammals and T4 bacteriophage.

The active site. The surface of the cavity in the large domain that contains the active site is lined by the side chains of 25 residues. Of these, 16 (64 percent) are invariant and 4 others are found in seven of the eight sequences. There are 16 charged or polar side chains that are candidates for functional groups in catalysis or binding (or both); of these, eight are invariant. Four of the residues in the cavity are contributed by the other subunit. This explains why the native enzyme has never been dissociated into catalytically active monomers and suggests how the active sites of the two subunits might communicate in the putative negative-cooperativity of ligand binding (33).

The thiol of the invariant Cys¹⁹⁸ is the nucleophile that transiently adds to the 6-position of dUMP during catalysis, and that covalently

binds to mechanism-based inhibitors such as FdUMP (1, 34). It lies at the NH₂-terminus of β -iv, close to the dimer interface, with the thiol exposed on the right wall of the active site cavity (Fig. 7). The catalytic thiol forms a hydrogen bond or ion pair with Arg²¹⁸, which is located on β -iii on the left wall of the cavity. Such interactions may lower the pK_a of the catalytic thiol and enhance its nucleophilicity for attack at the 6-position of dUMP. Antiparallel β sheet cross-links the backbone chains of strands iii and iv between the C=O of residue 198 and the NH of residue 218. The two sides of the active site are also cross-linked by hydrogen bonds between side chains of the highly conserved His¹⁹⁹ and Asn²²⁹; the side chain of the former is also hydrogen-bonded to Tyr¹⁴⁶. This hydrogen-bonding scheme is tentative since the details of the active site geometry may change when the protein is refined at higher resolution. The β bulge at residue 200 and prolines 196 and 197 restrict the conformation of Cys¹⁹⁸.

Chemical modification and ¹³C-NMR (nuclear magnetic resonance) studies of TS implicate at least one arginine residue as an essential cationic locus for binding the 5'-phosphate of dUMP (35, 36). Within the active site cavity are two closely approximated guanidinium groups of the invariant Arg²¹⁸ and Arg^{179'}, whose guanidinium carbons are 5.5 Å apart and could provide binding for the 5'-phosphate. The next closest guanidinium groups, those of Arg²³ and invariant 178' are not in the cavity, and appear too far (10 and 20 Å, respectively) from the catalytic thiol to serve as a binding site. However, the guanidinium of Arg^{178'} freely extends into solvent and could rotate to within 7 Å of the SH of Cys¹⁹⁸. Thus, it is possible that Arg^{178'} may play a role in binding the nucleotide.

We were unable to grow crystals containing dUMP even though TS-dUMP complexes readily form in solution. However, it is well established that P_i binds tightly to TS and effectively competes with dUMP (37), and that there is a density peak that spans and is approximately equidistant from the guanidinium groups of Arg²¹⁸ and Arg^{179'}, which is likely to be P_i. The distances between the centers of the peak and the guanidinium carbons are 5.0 and 4.0 Å. The P_i density is close (5 Å) to the catalytic thiol and, through steric effects or reduction of nucleophilicity, may explain why P_i decreases

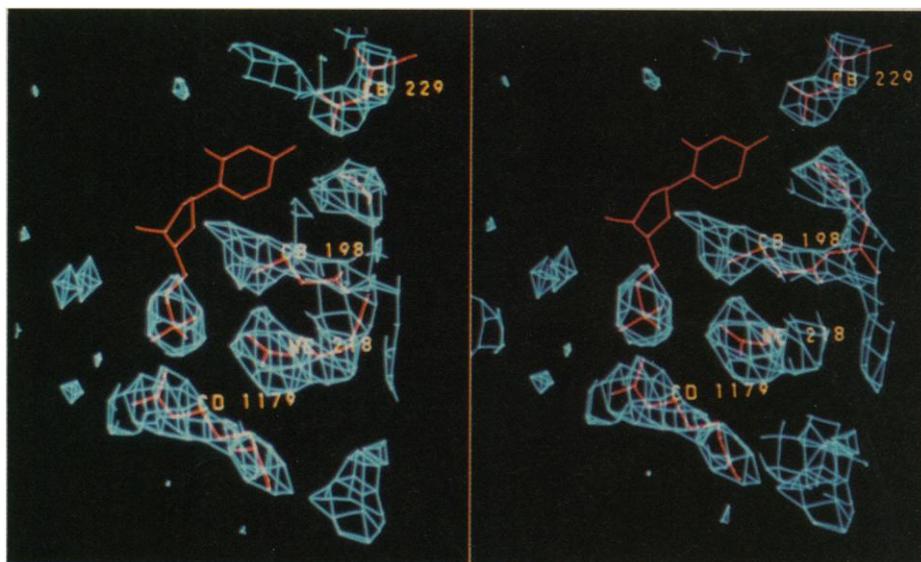


Fig. 8. Convergent stereoscopic view of the 3 Å resolution electron density and the model for TS in the region of the active site with dUMP modeled into the expected binding site. The phosphate of dUMP is placed in the density interpreted as a phosphate-binding site, and dUMP is docked into the active site of TS in the anti conformation. Arg²¹⁸ and Arg^{179'} from the neighboring subunit (labeled 1179) coordinate the phosphate site.

the rate of enzyme inactivation by thiol reagents (36). Thus, Arg²¹⁸ and Arg^{179'} probably provide a bidentate locus for binding the 5'-phosphate dianion of the substrate dUMP.

Binding of substrate. Molecular modeling techniques (38) were used to simulate the structure of the TS-dUMP complex (Fig. 8). The 5'-phosphate of dUMP was placed in the P_i binding site, in close contact with the guanidinium groups of Arg²¹⁸ and Arg^{179'}. We placed the 6-position of dUMP as close to the catalytic thiol of Cys¹⁹⁸ as possible without incurring intolerable steric interactions with other groups. With these constraints, the most acceptable models require that the nucleotide be in the anti conformation and that nucleophilic attack by thiol occur from over the ribosyl ring oxygen. The anti conformation is the favored conformation of dUMP in solution (39), and our models do not support the proposal that TS preferentially binds to nucleotides in their syn conformation (40). A basic residue seems essential to assist in removal of the poorly acidic proton at C-5 of the ternary covalent complex during dTMP formation, and to catalyze 5-H exchange at C-5 of the binary TS-dUMP complex (41). His¹⁹⁹ is close enough (3.5 Å) to the 5-position of dUMP in the putative structure of the binary complex to act as the base; however, it is not an invariant since it is substituted by Val in the ϕ 3T enzyme. The side chain amides of conserved residues Asn²²⁹ and Gln²¹⁷ are in the vicinity for possible hydrogen bonding to O-4 and N-3 of dUMP. The phenolic oxygen of Tyr⁶¹, which is another conserved residue, is only 6 Å from the catalytic thiol and is in close proximity to the 3'-hydroxyl of dUMP, a known determinant of binding specificity (1).

The normal physiological forms of folate cofactors possess multiple γ -linked polyglutamates, which increase their affinity for TS (42). Lysines at 50, 51, and 58 of *L. casei* TS, implicated in binding of carboxylate moieties of folyl polyglutamates (43), are on the polyglutamate-binding loop and helix, extending over the highly conserved residues 44 to 67. Lys⁵⁰ and Lys⁵¹ lie on an exposed part of the loop immediately preceding α -B. Lys⁵⁸ on α -B is only found in the *L. casei* sequence. However, in all other TS's there is a Lys or Arg in the position corresponding to residue 55; the side chains of these residues would emanate from the same side of α -B as Lys⁵⁸ and could serve the same role in folyl polyglutamate binding. Although none of the putative polyglutamate-binding groups of TS are within the active site cavity, the polyglutamate moieties of the cofactor are also far removed from the site of catalysis. The distances between the ϵ -amino groups of Lys⁵⁰ and Lys⁵¹, and between those of Lys⁵¹ and Lys⁵⁸, are 11 to 12 Å, suitably placed for charge-charge interactions

with the α -carboxylates of consecutive residues of polyglutamate.

Conformational differences. We could not optimize docking of dUMP in the enzyme structure solved here; thus, there must be conformational differences between the structure reported here and the structure of the protein in the TS-dUMP complex. Bound nucleotides greatly enhance the affinity of TS for folates. However, we could not model binding of CH₂H₄folate to the TS-dUMP complex so as to place the one carbon unit in a position suitable for transfer. Thus, there must also be a significant conformational change after dUMP binding to accommodate formation of ternary complexes and position groups appropriately for catalysis. Other studies have demonstrated conformational changes on formation of inhibitory complexes between TS, nucleotide, and folate cofactor or analogs (36, 44). The structure we have determined does not reveal certain salient aspects of catalysis, and conformational changes must occur to accommodate productive dUMP and folate binding, position the one carbon unit of the cofactor for condensation, and juxtapose a basic residue of the protein with the 5-H of covalent dUMP adducts. Identification of these may derive from solution of structures of complexes which reveal various conformational states of the enzyme.

Prospects. The work described here consolidates more than a decade of effort directed at understanding the structure of TS. We can now provide a vista of future directions in studies of this enzyme.

1) The high conservation of primary sequences of TS in elements important to higher order structure defines a core structure of the protein that must be common in other TS's. Thus, with the present data and current methods of energy minimization, structures of other TS's can be modeled prior to crystallographic analysis.

2) Provided that suitable crystals can be formed, it will be a straightforward matter to solve the crystal structures of TS's from other sources. Since several TS's are available, it should not be long before a number of such structures are available.

3) We and others have shown that several conformational states of the enzyme occur during catalysis and upon ligand binding. To understand the structure, function, and inhibition of TS, we must understand these alternative conformers. Since ligands are known which, when bound to the enzyme, freeze such conformational states, we have an opportunity to take snapshots of conformations that occur during catalysis. Thus far, we have crystallized a binary complex containing a nucleotide analog and a ternary complex containing dUMP and a folate analog; crystallization of the ternary

FdUMP-CH₂H₄folate-TS has also been recently reported (45).

4) Site-directed mutagenesis of TS can now be profitably directed toward answering questions on structure, binding, and catalysis.

5) Computer modeling, interfaced with more classical methods of drug design, has the potential of providing truly predictive and rational approaches to the design of potent inhibitors for any given enzyme, and inhibitors that discriminate between subtly different enzymes from different sources. Because TS is such an important chemotherapeutic target, we foresee a considerable investment of effort directed toward such goals.

REFERENCES AND NOTES

1. See D. V. Santi and P. V. Danenberg, in *Folates and Pterins*, R. L. Blakley and S. J. Benkovic, Eds. (Wiley, New York, 1984), vol. 1, pp. 345–398.
2. C. E. Garrett *et al.*, *Mol. Biochem. Parasitol.* **11**, 257 (1984).
3. E. Margolish and W. M. Fitch, *Ann. N.Y. Acad. Sci.* **151**, 359 (1968).
4. D. V. Santi, Y. Wataya, A. Matsuda, in *Enzyme-Activated Irreversible Inhibitors*, N. Seiler, M. J. Jung, J. Koch-Weser, Eds. (Elsevier, New York, 1978), pp. 291–303; J. Wu and D. V. Santi, *J. Biol. Chem.*, in press.
5. C. Heidelberger, P. V. Danenberg, R. G. Moran, in *Advances in Enzymology and Related Areas in Molecular Biology*, A. Meister, Ed. (Wiley, New York, 1983), pp. 57–119.
6. R. E. Koeppe, R. M. Stroud, V. A. Pena, D. V. Santi, *J. Mol. Biol.* **98**, 155 (1974).
7. Y. Wataya and D. V. Santi, *Methods Enzymol.* **46**, 307 (1977).
8. T. C. Crusberg, R. Leary, R. L. Kisliuk, *J. Biol. Chem.* **245**, 5292 (1970).
9. B. W. Low and F. M. Richards, *J. Am. Chem. Soc.* **74**, 1660 (1952).
10. R. M. Stroud, L. M. Kay, R. E. Dickerson, *J. Mol. Biol.* **83**, 185 (1974).
11. H. W. Wyckoff *et al.*, *ibid.* **27**, 563 (1967).
12. Ng. H. Xuong, S. T. Freer, R. Hamlin, C. Nielsen, W. Vernon, *Acta Crystallogr. Sect. A* **34**, 289 (1978); R. Hamlin *et al.*, *J. Appl. Crystallogr.* **14**, 85 (1981); R. Hamlin, *Methods Enzymol.* **114**, 416 (1985); A. J. Howard, C. Nielsen, Ng. H. Xuong, *ibid.*, p. 452.
13. G. F. Maley, R. L. Bellisario, D. U. Guarino, F. Maley, *J. Biol. Chem.* **254**, 1288 (1979); R. L. Bellisario, G. F. Maley, D. U. Guarino, F. Maley, *ibid.*, p. 1296; G. F. Maley, R. L. Bellisario, D. U. Guarino, F. Maley, *ibid.*, p. 1301.
14. This idea was suggested to us by J. Herriott.
15. B. C. Wang, *Methods Enzymol.* **115**, 90 (1985).
16. J. Finer-Moore and R. M. Stroud, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 155 (1984).
17. G. A. Sim, *Acta Crystallogr.* **12**, 813 (1959); M. G. Rossmann and D. M. Blow, *ibid.* **14**, 641 (1961); W. A. Hendrickson and E. E. Lattman, *ibid. Sect. B* **26**, 136 (1970); G. Bricogne, *ibid. Sect. A* **32**, 832 (1976).
18. J. L. Chambers and R. M. Stroud, *ibid. Sect. B* **33**, 1824 (1977).
19. J. H. Kennert, *ibid. Sect. A* **32**, 614 (1976); W. A. Hendrickson and J. H. Kennert, *Biomolecular Structure, Function Conformation and Evolution*, R. Srinivasan, Ed. (Pergamon, Oxford, 1980), vol. 1, p. 43.
20. C. Chothia, *J. Mol. Biol.* **75**, 295 (1973); F. E. Cohen, M. J. E. Sternberg, I. W. R. Taylor, *ibid.* **148**, 253 (1981).
21. The solvent-accessible surface area of the dimer and each monomer in isolation was calculated by the method of B. Lee and F. M. Richards [*J. Mol. Biol.* **55**, 379 (1971)]. Residues at the dimer interface are more accessible to solvent in the monomer than in the dimer.
22. C. Chothia and J. Janin, *Nature (London)* **256**, 705 (1975).
23. J. S. Richardson, *Adv. Protein Chem.* **34**, 167 (1981).
24. Yu. A. Ovchinnikov, N. G. Abdulaev, M. Yu. Feigina, A. V. Kiselev, N. A. Lobanov, *FEBS Lett.* **100**, 219 (1979); H. G. Khorana *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 5046 (1979); A. Michel *et al.*, *EMBO J.* **5**, 1149 (1986).
25. Residues from the two subunits are distinguished by a prime (') following the residue numbers of one monomer.
26. M. O. Dayhoff, R. M. Schwartz, B. Orcutt, in *Atlas of Protein Sequence and Structure*, M. O. Dayhoff, Ed. (National Biomedical Research Foundation, Washington, DC, 1978), vol. 5, suppl. 3, pp. 345–352. Our sequence numbers are always referred to those of *L. casei* TS. The percentage numbers listed for conservation or invariance are calculated by adding up the maximum number of homologous (by the criteria of Dayhoff *et al.*) or invariant residues, less one, at each site, divided by the maximum possible number (the number of sequences less one).
27. The $\phi 3T$ sequence is the most diverged from the most conserved consensus sequence and may give new insights into substitutions and their effects on functional determinants.
28. E. P. Garvey and D. V. Santi, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 7188 (1985).
29. T. Deng, D. Li, C.-H. Jenh, L. F. Johnson, *J. Biol. Chem.* **261**, 16000 (1986).
30. C. S. Craik, W. J. Rutter, R. J. Fletterick, *Science* **220**, 1125 (1983).
31. W. Gilbert, *Nature (London)* **271**, 501 (1978); C. C. F. Blake, *ibid.* **273**, 267 (1978).
32. F. K. Chu, G. F. Maley, F. Maley, M. Belfort, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 3049 (1984).
33. J. L. Aull, J. A. Lyon, R. B. Dunlap, *Microchem. J.* **19**, 210 (1974); J. H. Galivan, G. F. Maley, F. Maley, *Biochemistry* **15**, 356 (1976); K. D. Danenberg and P. V. Danenberg, *J. Biol. Chem.* **254**, 4345 (1979).
34. R. L. Bellisario, G. F. Maley, J. H. Galivan, F. Maley, *Proc. Natl. Acad. Sci. U.S.A.* **73**, 1848 (1976); A. L. Poglotti, K. M. Ivanetich, H. Sommer, D. V. Santi, *Biochem. Biophys. Res. Commun.* **70**, 972 (1976).
35. M. Belfort, G. F. Maley, F. Maley, *Arch. Biochem. Biophys.* **204**, 340 (1980).
36. C. A. Lewis and R. B. Dunlap, in *Topics in Molecular Pharmacology*, A. S. V. Brugen and G. C. K. Roberts, Eds. (Elsevier/North-Holland, New York, 1981), pp. 169–219.
37. J. H. Galivan, G. F. Maley, F. Maley, *Biochemistry* **15**, 356 (1976); C. A. Lewis, Jr., W. A. Munroe, R. B. Dunlap, *ibid.* **17**, 5382 (1978).
38. T. A. Jones, in *Computational Crystallography*, D. Sayre, Ed. (Oxford Univ. Press, Oxford, 1982), pp. 303–317.
39. T. R. Emerson, R. J. Swan, T. L. V. Ulbricht, *Biochemistry* **6**, 843 (1967).
40. R. P. Leary, N. Beaudette, R. L. Kisliuk, *J. Biol. Chem.* **250**, 4869 (1975).
41. A. L. Poglotti, Jr., C. Weill, D. V. Santi, *Biochemistry* **13**, 2794 (1979).
42. J. J. McGuire and J. K. Coward, in *Folates and Pterins*, R. L. Blakley and S. J. Benkovic, Eds. (Wiley, New York, 1984), vol. 1, pp. 135–190.
43. G. F. Maley, F. Maley, C. M. Baugh, *Arch. Biochem. Biophys.* **216**, 551 (1982).
44. T. W. Bruce and D. V. Santi, *Biochemistry* **21**, 6703 (1982); A. Lockshin, K. Mondal, P. V. Danenberg, *J. Biol. Chem.* **259**, 11346 (1984).
45. E. Tykarska, L. Lebioda, T. P. Bradshaw, R. B. Dunlap, *J. Mol. Biol.* **191**, 147 (1986).
46. R. E. Dickerson, J. C. Kendrew, B. E. Strandberg, *Acta Crystallogr.* **14**, 1188 (1961).
47. J. S. Richardson, *Methods Enzymol.* **115**, 341 (1985).
48. K. Takeishi *et al.*, *Nucleic Acids Res.* **13**, 2035 (1985).
49. S. M. Perryman, C. Rossana, T. Deng, E. F. Vanin, L. F. Johnson, *Mol. Biol. Evol.* **3** (no. 4), 313 (1986).
50. S. M. Beverley, T. E. Ellenberger, J. S. Cordingley, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 2584 (1986); R. Grumont, W. L. Washtien, D. Caput, D. V. Santi, *ibid.*, p. 5387.
51. R. W. Honess *et al.*, *ibid.*, p. 3604.
52. M. Belfort, G. Maley, J. Pedersen-Lanc, F. Maley, *ibid.* **80**, 4914 (1983).
53. E. Kenny, T. Atkinson, B. S. Hartley, *Gene* **34**, 335 (1985).
54. Supported by PHS grant number CA41323, by NIH grants GM 24485 (R.M.S.) and AI 19358 (D.V.S.), and by a Damon Runyon-Walter Winchell cancer research fellowship (L.W.H.). We thank Ng. H. Xuong for providing time on his area-detector-based data collection system and R. Hamlin, D. Anderson, and C. Nielsen for assistance with the high resolution data collection; C. Craik and F. Cohen for their intellectual contribution to developing ideas; C. Schiffer and W. Den for leading us in the art of science and science of illustration, respectively; M. Sternberg for help with calculating solvent-accessible surfaces; and L. Johnson for providing the positions of the mouse gene introns before publication of his data. When sufficiently refined, coordinates will be deposited in the Brookhaven Data Bank.

28 October 1986; accepted 23 December 1986