Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 9264–9268, September 1995
Biochemistry

Crystal structure of the large fragment of *Thermus aquaticus* DNA polymerase I at 2.5-Å resolution: Structural basis for thermostability

(x-ray crystal structure)

SERGEY KOROLEV, MURAD NAYAL, WAYNE M. BARNES, ENRICO DI CERA, AND GABRIEL WAKSMAN*

Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, MO 63110

Communicated by Charles C. Richardson, Harvard Medical School, Boston, MA, June 16, 1995

ABSTRACT **The crystal structure of the large fragment of the *Thermus aquaticus* DNA polymerase (Klentaq1), determined at 2.5-Å resolution, demonstrates a compact two-domain architecture. The C-terminal domain is identical in fold to the equivalent region of the Klenow fragment of *Escherichia coli* DNA polymerase I (Klenow pol I). Although the N-terminal domain of Klentaq1 differs greatly in sequence from its counterpart in Klenow pol I, it has clearly evolved from a common ancestor. The structure of Klentaq1 reveals the strategy utilized by this protein to maintain activity at high temperatures and provides the structural basis for future improvements of the enzyme.**

Amplification of DNA fragments by the polymerase chain reaction (PCR) has become an important and widespread tool of genetic analysis since the introduction of the thermostable DNA polymerase from *Thermus aquaticus* (*Taq*) (1–3). The enzyme, by enabling the amplification reaction to be performed at higher temperatures, allows the convenience of heat denaturation of DNA without enzyme inactivation. Purified *Taq* DNA polymerase, however, is devoid of 3′-5′ exonuclease activity and thus cannot excise misincorporated nucleotides (4, 5). Consequently, DNA amplification by the *Taq* DNA polymerase is an error-prone process. Enzymes with N-terminal deletions show a reduced tendency toward errors, as do some recently discovered thermostable DNA polymerases which have an integral editing exonuclease activity (6, 7). The latter enzymes, however, are unable to amplify sequences in excess of 5.0 to 7.0 kb that full-length *Taq* DNA polymerase (8) or N-terminally deleted enzyme (7) can amplify readily. The amplification of very large DNA fragments (up to 35 kb) was recently achieved by combining an N-terminally deleted *Taq* DNA polymerase called Klentaq1 with a low level of an archaebacterial thermostable DNA polymerase exhibiting 3′-5′ exonuclease activity (9, 10). *Taq* DNA polymerase or forms of the enzyme with N-terminal deletions are also used in DNA sequencing (10–12). However, the quality of the data has been limited and the expense kept high by the poor affinity of the enzyme for dideoxynucleotides. Mutants with increased affinity for chain terminators would be of considerable interest.

To understand the structural basis of thermostability and provide the foundation for the improvement of the *Taq* DNA polymerase, we present here the three-dimensional structure of Klentaq1.†

MATERIALS AND METHODS

Crystallization of Klentaq1. A modified version of Klentaq1 (10) (residues 281–832) with a 7 amino acid N-terminal extension (MGKRKST) was used and yielded crystals diffracting to beyond 2.5-Å resolution (Table 1). Crystals of Klentaq1 were obtained at room temperature by using vapor diffusion against a solution of 6% (wt/vol) polyethylene glycol 3350/50 mM $MgCl_2$/100 mM Tris·HCl, pH 9.0, starting with equal mixtures of protein and polyethylene glycol solutions (13). Klentaq1 crystals are in space group $P2_12_12$ (a = 109.4 Å, b = 136.8 Å, c = 45.6 Å) with one molecule in the asymmetric unit.

Structure Determination. Details of structure determination will be published elsewhere. In brief, heavy-atom derivatives were prepared by soaking the crystals with uranyl and platinum compounds as indicated in Table 1. Anomalous scattering measurements were included for all derivatives. Heavy-atom positions were obtained by inspection of the Patterson maps or use of difference Fourier techniques. Heavy-atom parameters were refined and initial phases were calculated by using the program MLPHARE (Z. Otwinowski). The MIR phases were further improved by solvent flattening using SQUASH (14). A partial model consisting of a polyalanine chain was built, using the program O and a data base of protein structure (15, 16). The map was improved by cycles of refinement using X-PLOR (17), phase combination using SIGMAA (18), and model building (Fig. 1). Residues 274–289 and residues 504–514 do not have interpretable density. Density for side chains is weak for residues 514–519 and therefore side chains have not been built in this region. The structure has been refined to an R factor of 22.5% without addition of solvent molecules. Restrained refinement of temperature factors resulted in an average B-factor of 25 Å^2.

RESULTS AND DISCUSSION

Structure of Klentaq1. Sequence homology between residues 420–832 of Klentaq1 and residues 516–928 of Klenow pol I is high (49.6% identity). This region corresponds to the large domain of the Klenow pol I structure (19). As expected, this region in Klentaq1 is similar in fold (Fig. 2*A*) and superimposes with the large domain of Klenow pol I with an rms deviation in C^α of 1.42 Å. As in Klenow pol I, the large domain of Klentaq1 consists of three subdomains (the thumb, the palm, and the fingers) forming a deep crevice of the appropriate size to accommodate double-stranded DNA (19, 21–26). Differences in fold between the large domains of Klentaq1 and Klenow pol I are primarily located in the fingers' tip region,

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: *Taq*, *Thermus aquaticus*; Klenow pol I, Klenow fragment of *Escherichia coli* DNA polymerase I; MIR, multiple isomorphous replacement.

*To whom reprint requests should be addressed at: Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, Box 8231, 660 South Euclid Avenue, St. Louis, MO 63110.

†The C^α coordinates of Klentaq1 have been deposited in the Protein Data Bank, Chemistry Department, Brookhaven National Laboratory, Upton, NY 11973 (1KTQ).

Table 1. Summary of crystallographic data

Measurement	Native	$UO_2(OAc)_2$ (2 mM, 13 d)	K_2PtCl_4 (1 mM, 20 h)	$UO_2(NO_3)_2$ (2 mM, 24 h)
Resolution, Å	2.5	3.0	3.0	3.0
Reflections (observed/unique)	66,183/22,333	37,980/13,065	38,377/12,983	29,652/11,622
Data coverage, %	90.6	90.7	89.4	81.0
R_{sym}, %	7.0	6.9	7.2	8.3
R_{iso}, %		13	18	20
MIR analysis (15–3.0 Å)				
Phasing power		1.70	0.90	0.70
Mean overall figure of merit	0.663			
Refinement (6–2.5 Å)				
R factor, %		22.5		
Reflections ($\|F\| > 2\sigma\|F\|$)		20,438 (90.0%)		
Total number of atoms		4239		
rms deviation in bond length, Å		0.011		
rms deviation in bond angle, °		2.80		

$R_{sym} = \Sigma |I - \langle I \rangle| / \Sigma I$, where I = observed intensity, and $\langle I \rangle$ = average intensity from multiple observations of symmetry-related reflections. $R_{iso} = \Sigma \| F_{PH} | - |F_P| / \Sigma |F_P|$, where $|F_P|$ = protein structure factor amplitude, and $|F_{PH}|$ = heavy-atom derivative structure factor amplitude. MIR, multiple isomorphous replacement. Phasing power = rms ($|F_H|/E$), where $|F_H|$ = heavy-atom structure factor amplitude and E = residual lack of closure. The rms deviations in bond lengths and angles are the deviations from ideal values. The 2.75- to 2.5-Å resolution outer shell for the native data set is 84% complete with $\langle F/\sigma F \rangle = 3.7$.

where an early termination of helix O1 is observed (see ref. 19 for notation of secondary structures). Also, helices H and I differ from their counterparts in apo Klenow pol I in that they are tilted toward the N-terminal domain so that residues in the N terminus of helix I have moved about 4 Å.

The lack of a 3′-5′ proofreading exonuclease activity in Klentaq1, together with the apparent lack of sequence homology between the N-terminal regions of Klentaq1 and Klenow pol I, suggested that their structures might be different. Yet Klentaq1 conserves the two-domain structure observed in Klenow pol I with a distinct N-terminal domain (Fig. 2*A*; ref. 19), and there is considerable topological homology between the N-terminal domains of Klenow pol I and Klentaq1 (Fig. 2 *B* and *C*). The major sheet composed of strands 1, 2, 3, and 4 is conserved, as are helices B, C, D, E, and F.

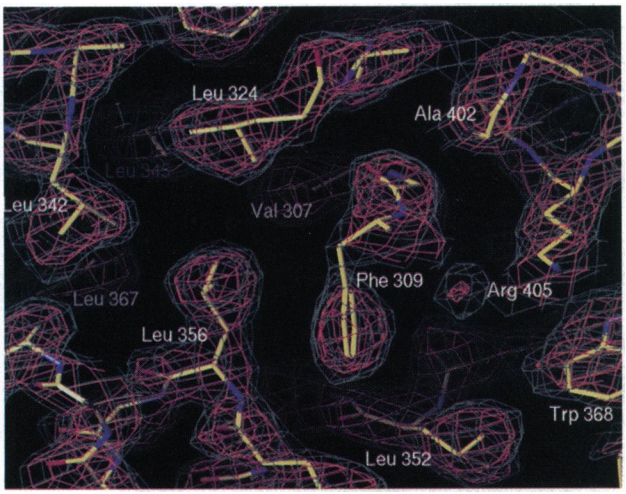

FIG. 1. Electron density at 2.5-Å resolution of a representative region of the Klentaq1 structure. The overlying stick figures represent the refined atomic coordinates. The electron density was calculated by using coefficient $(2|F_{obs}| - |F_{calc}|)\exp(-i\alpha_c)$, where $|F_{obs}|$ is the observed factor amplitude, and $|F_{calc}|$ and α_c are the amplitudes and phases calculated from the model. Green contour lines indicate electron density at 1.2σ, and orange, at 1.8σ above the mean density. The represented region corresponds in the Klenow fragment of *Escherichia coli* DNA polymerase I (Klenow pol I) to the dTMP-binding site. In Klentaq1, this region is integrated to the hydrophobic core of the domain, and consequently, it has lost the ability to bind nucleoside phosphates.

There are also remarkable differences between the N-terminal regions of Klentaq1 and Klenow pol I. Helix A of Klenow pol I is missing and is replaced by a proline-rich loop. Major deletions in the loop structures are also observed. For instance, an extensive loop seen between helices E and F has been deleted in Klentaq1. Helix F itself is much shorter in Klentaq1 than in Klenow pol I. The loop between strand 4 and helix B is 10 amino acids shorter, and helix B itself is 6 amino acids shorter than its counterpart in Klenow pol I. Clearly, this domain has undergone major rearrangements during evolution. As a consequence, the N-terminal domain in Klentaq1 is much more compact (40 × 40 × 29 Å in Klentaq1 against 50 × 45 × 37 Å in Klenow pol I). This domain now presents a surface that extends smoothly toward the palm region of the large domain and merges at the same angle with a similar surface in the latter region of the protein to form a vast flat area covering almost the entire length of the protein (Fig. 3). Overall, the structure resembles a wishbone, the handles of which consist of the tips of the thumb and fingers domains.

The structure presented here provides a clear explanation as to why a proofreading activity is lacking in Klentaq1. dTMP in Klenow pol I binds in a cavity of the small domain that is formed by residues of strand 2 (residues 355, 357, and 358), helix C (residue 424), and helix F (residues 497 and 501) (19, 29). In Klentaq1, these residues have either disappeared (Glu-357 and Thr-358) or have been replaced (Fig. 1). The same cavity in Klentaq1 is now filled with the hydrophobic side chains of Phe-309 (substituting for Asp-355 of pol I), Leu-356 (substituting for Asp-424), Leu-345, and Val-307. These residues form a hydrophobic region contributing to the core of the protein. The architecture of the protein in this region of the small domain is not dramatically affected by any of the deletions or truncations that seem to have affected the N-terminal domain during evolution. It is therefore a distinct possibility that the region could be reengineered to recover nucleotide-binding affinity.

Significant differences between Klentaq1 and Klenow pol I can be found at the interface between the small N-terminal domain and the large C-terminal domain. Residues forming the interface in both enzymes are essentially contributed by helices C and D, and the loop connecting helices D and E in the small domain, and by residues in helix G, strand 7, and the loop connecting strands 7 and 8 in the large domain. However, helix C in Klenow pol I interacts with helix G only at its N-terminal tip. In contrast, the two helices in Klentaq1 lie

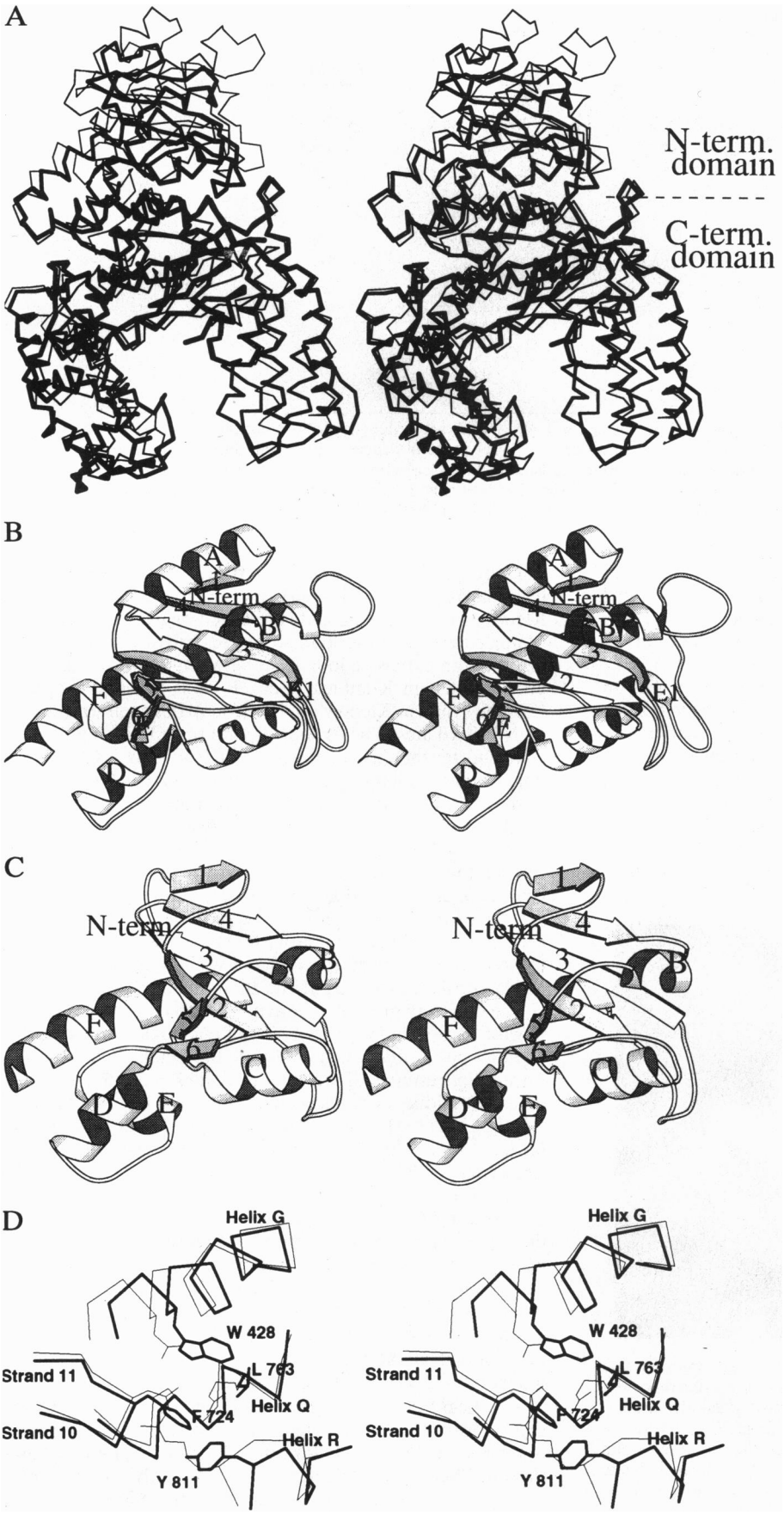

FIG. 2. Comparison of the structures of Klentaq1 and Klenow pol I (20). (*A*) Superimposed stereodiagrams of the C^{α} tracing of the structures of the two enzymes. Klentaq1 and Klenow pol I are shown with thick and thin lines, respectively. (*B*) Stereo ribbon diagram of the small domain of Klenow pol I. The notation used to label secondary structures is according to ref. 19. (*C*) Stereo diagram of the small domain of Klentaq1. (*D*) Superimposed stereodiagrams of a cluster of aromatic residues in Klentaq1 (thick lines) replacing a cluster of charged residues in Klenow pol I (thin lines). Only residues in Klentaq1 are labeled. Amino acids at positions equivalent to Trp-428, Phe-724, Leu-763, and Tyr-811 of Klentaq1 are Asn-524, Asp-819, Arg-858, and Arg-909, respectively, in Klenow pol I. Asp-819 forms favorable ion pairs with Arg-858 and Arg-909. However, unfavorable contacts between the two arginine residues are eliminated in Klentaq1 by substitutions to leucine and tyrosine.

almost parallel to each other, with the result that contacts between helices C and G are more extensive. A consequence of the repositioning of helix C is an expanded hydrophobic core. Klentaq1 buries 2960 Å^2 of its surface or 21% of the total surface of the small domain at the interface compared to 2730 Å^2 and 14.7%, respectively, for Klenow pol I. This may

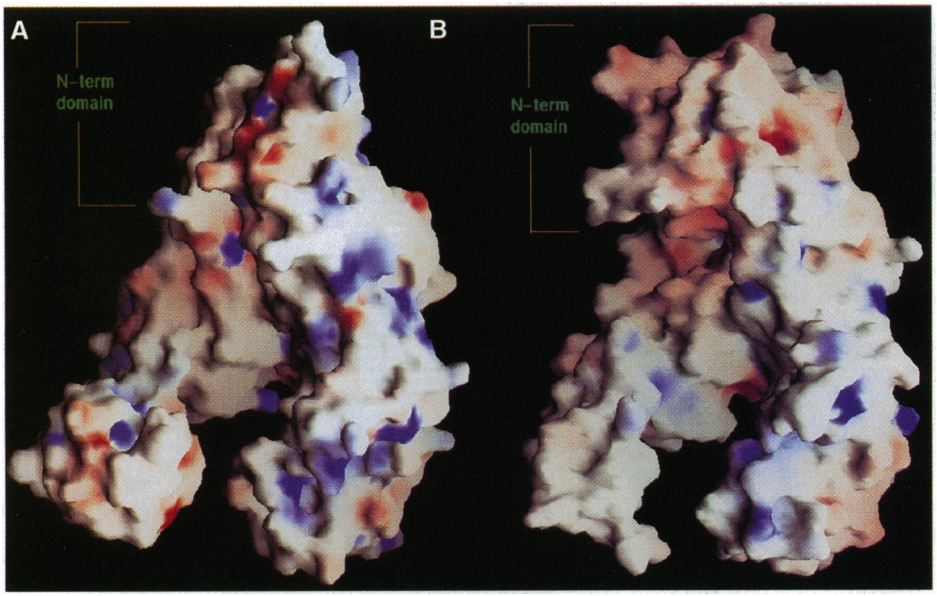

FIG. 3. Molecular surfaces of Klentaq1 and Klenow pol I. (*A*) Molecular surface of Klentaq1 (27), calculated and displayed by using GRASP (28). The surface is colored according to the local electrostatic potential and is deep blue in the most positive regions and deep red in the most negative, with linear interpolation for values in between. (*B*) Molecular surface of Klenow pol I. Color definitions are the same as for *A*.

significantly increase the stability of Klentaq1. A survey of ion-pairing interactions involved at the interface in both polymerases shows that three additional ion pairs are formed in Klentaq1, also contributing to overall stability (30, 31).

Thermostability. A classical approach to address the problem of thermostability in proteins has been to compare equivalent protein species in thermophilic and mesophilic organisms (31–33). The structure of Klentaq1 can be readily compared with its mesophilic counterpart from *E. coli*. The large domains in the two enzymes can be superimposed, and, in spite of a striking lack of sequence similarity, the N-terminal domains show a very similar fold.

Original treatments of thermostability have emphasized the role of amino acid substitutions (31, 33). For instance, these studies point toward substitutions of lysine to arginine and of aspartic to glutamic acids as possible contributors to thermostability. Inspection of substitutions involving these amino acids in the sequence alignment of Klentaq1 and Klenow pol I indicates a large number of nonconserved (charged to uncharged polar, charged to hydrophobic, or charged to oppositely charged) amino acid substitutions (79 in total). Interestingly, charged to oppositely charged amino acid substitutions occur 19 times and are spread out over the entire structure. Only six substitutions conserve charge (lysine to arginine or vice versa) with only four lysine residues replaced in Klentaq1 with arginine residues. This extensive pattern of opposite-charge substitutions clearly indicates a global rearrangement of the charge distribution.

To further address this observation, the ion-pairing patterns in Klentaq1 and Klenow pol I were examined. Contradictory results were obtained. For instance, the C-terminal domain of Klenow pol I exhibited a larger number of ion pairs than that of Klentaq1 (41 against 32). However, the number of unfavorable ion-pairing interactions was also larger (12 against 8). To evaluate the net result of these opposing effects, the electrostatic contribution to the free energy of folding was calculated by using the continuum electrostatic method (34–36). These calculations show that the electrostatic energy for the process of assembling the protein from individual amino acids in solution is significantly more favorable in Klentaq1 than in Klenow pol I (Fig. 4). The largest differences occur in the N-terminal domain, where most unfavorable electrostatic interactions found in Klenow pol I have been eliminated in Klentaq1 (Fig. 4). In the large domain, the global energy profile is slightly in favor of Klentaq1 with higher numbers of unfavorable interactions in Klenow pol I (Fig. 4). We conclude from this study that the structural basis for thermostability may lie in part in a reorganization of the N-terminal and C-terminal domains during evolution that has resulted in the optimization of the electrostatic residue–residue and residue–solvent interactions in the folded state.

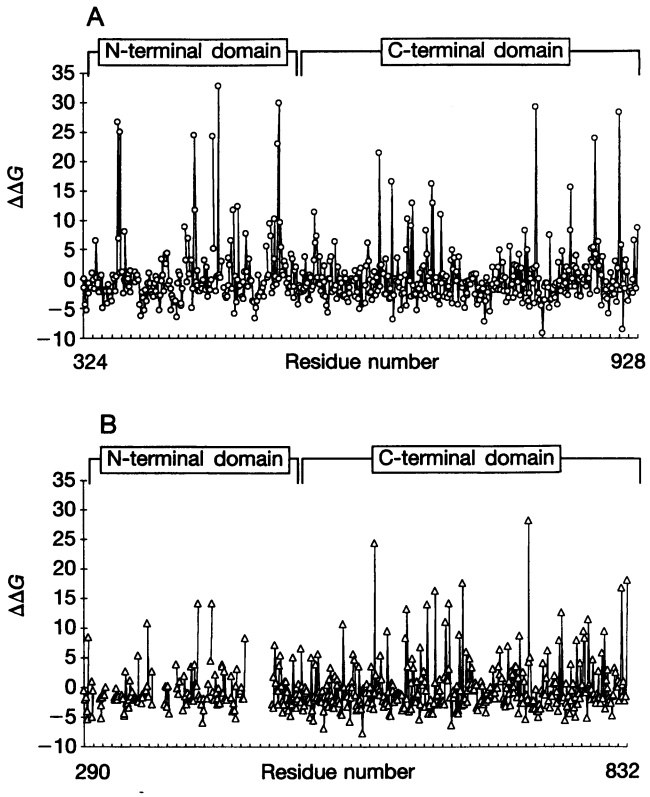

FIG. 4. Difference in the electrostatic component of the folding free energy ($\Delta\Delta G$) for each residue. (*A*) Klenow pol I. (*B*) Klentaq1. $\Delta\Delta G$ is equal to the sum of $\Delta\Delta G_{solv}$ and $\Delta\Delta G_{protein}$, where $\Delta\Delta G_{solv}$ is the electrostatic free energy of desolvating each residue in the process of protein folding and $\Delta\Delta G_{protein}$ is the electrostatic free energy of interaction of each residue with the rest of protein in the folded state. Units are kcal/mol. These calculations were carried out using finite-difference numerical methods to solve the Poisson–Boltzmann equation, as implemented in the program DELPHI (34–36). The sequences of the N-terminal domains have been aligned; gaps in plot B indicate regions present in Klenow pol I but absent in Klentaq1. The N-terminal and C-terminal residue numbers are indicated.

Residues having unfavorable electrostatic interactions in Klenow pol I are often replaced with hydrophobic or aromatic clusters in Klentaq1 (Fig. 2*D*; ref. 37). Sequence alignment based on secondary structural elements shows a large number of substitutions involving hydrophobic or aromatic residues (a net gain of 17 such residues for the large domain of Klentaq1 alone). A survey of the buried surfaces in the small and large domains of Klenow pol I and Klentaq1 shows that there is no difference in the total buried area per residue (126 Å^2). However, the hydrophobic component of this surface area is larger in Klentaq1 than in Klenow pol I (63.7% and 62.2%, respectively, in the large domain, and 65.2% and 64.8%, respectively, in the small domain). This increase is achieved through an equivalent reduction in uncharged polar buried surface area. Charged buried areas remain the same. These results suggest that thermostability in Klentaq1 may therefore also be partly achieved through an enhanced hydrophobic core (38–42).

Additional potential stabilizing features include a large number of substitutions involving proline residues. Alanine-to-proline substitutions have been used to evaluate the theory of entropy-dependent enhancement of protein stability (43) and have often resulted in increased stability (43, 44). The N-terminal domain of Klentaq1 contains a large number of proline residues (13 prolines or 10% of the amino acids that constitute this domain, against 6 and 3%, respectively, in Klenow pol I). Interestingly, helix A in Klenow pol I is replaced by a proline-rich loop structure.

The structure of Klentaq1 reveals the strategy utilized by this enzyme to maintain activity at high temperatures. Clearly, dramatic differences from Klenow pol I can be observed. In particular, the N-terminal domain has undergone extensive sequence and structural rearrangements that resulted in an enhanced hydrophobic core, a systematic elimination of unfavorable electrostatic interactions, and an increased size of its interface with the large domain. Similar observations, although to a lesser extent, apply to the large domain.

We thank F. S. Matthews for useful suggestions and support, A. B. Herr for help in crystallization, R. Jones for protein purification, and D. Vassilyev for help with phasing. This work was supported in part by funds from Washington University School of Medicine (to G.W. and W.M.B.), by National Institutes of Health Research Grant HL49413 (to E.D.C.), and by National Science Foundation Research Grant MCB94-06103 (to E.D.C.).

1. Mullis, K. B. & Faloona, F. (1987) *Methods Enzymol.* **55,** 335–350.
2. Saiki, R. K., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A. & Arnheim, N. (1985) *Science* **230,** 1350–1354.
3. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1988) *Science* **239,** 487–491.
4. Brutlag, D. & Kornberg, A. (1972) *J. Biol. Chem.* **247,** 241–248.
5. Eckert, K. & Kunkel, T. A. (1990) *Nucleic Acids Res.* **18,** 3739–3744.
6. Lundberg, K. S., Shoemaker, D. D., Adams, M. W. W., Short, J. M., Sorge, J. A. & Mathur, E. J. (1991) *Gene* **108,** 1–6.
7. Barnes, W. M. (1992) *Gene* **112,** 29–35.
8. Lawyer, F. C., Stoffel, S., Saiki, R. K., Chang, S.-Y., Landre, P. A., Abramson, R. D. & Gelfand, D. H. (1993) *PCR Methods Appl.* **2,** 275–287.
9. Barnes, W. M. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 2216–2220.
10. Barnes, W. M. (1995) U.S. Patent 5,436,149.
11. Innis, M. A., Myambo, K. B., Gelfand, D. H. & Brow, M. A. D. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 9436–9440.
12. Craxton, M. (1991) *Methods Companion Methods Enzymol.* **3,** 20–26.
13. McPherson, A. (1990) *Eur. J. Biochem.* **189,** 1–23.
14. Zhang, K. Y. & Main, P. (1990) *Acta Crystallogr. A* **46,** 41–46.
15. Jones, T. A. & Thirup, S. (1986) *EMBO J.* **5,** 819–822.
16. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991) *Acta Crystallogr. A* **47,** 110–119.
17. Brünger, A. T. (1988) X-PLOR Manual (Yale Univ., New Haven, CT), Version 2.2.
18. Read, R. (1986) *Acta Crystallogr. A* **42,** 140–149.
19. Ollis, D. L., Brick, P., Hamlin, R., Xuong, N. G. & Steitz, T. A. (1985) *Nature (London)* **313,** 762–766.
20. Kraulis, P. J. (1991) *J. Appl. Crystallogr.* **24,** 946–950.
21. Beese, L. S., Derbyshire, V. & Steitz, T. A. (1993) *Science* **260,** 352–355.
22. Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A. & Steitz, T. A. (1992) *Science* **256,** 1783–1790.
23. Sousa, R., Chung, Y. J., Rose, J. P. & Wang, B.-C. (1993) *Nature (London)* **364,** 593–599.
24. Jacobo-Molina, A., Ding, J., Nanni, R. G., Clark, A. D., Lu, X., Tantillo, C., Williams, R. L., Kamer, G., Ferris, A. L., Clark, P., Hizi, A., Hughes, S. & Arnold, E. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 6320–6324.
25. Sawaya, M. R., Pelletier, H., Kumar, A., Wilson, S. H. & Kraut, J. (1994) *Science* **264,** 1930–1935.
26. Pelletier, H., Sawaya, M. R., Kumar, A., Wilson, S. H. & Kraut, J. (1994) *Science* **264,** 1891–1903.
27. Richards, F. M. (1977) *Annu. Rev. Biophys. Bioeng.* **6,** 151–176.
28. Nicholls, A., Sharp, K. A. & Honig, B. (1991) *Proteins Struct. Funct. Genet.* **11,** 281–296.
29. Beese, L. S. & Steitz, T. A. (1991) *EMBO J.* **10,** 25–33.
30. Perutz, M. F. (1978) *Science* **201,** 1187–1191.
31. Raidt, H. & Perutz, M. F. (1975) *Nature (London)* **255,** 256–259.
32. Menendez-Arias, L. & Argos, P. (1989) *J. Mol. Biol.* **206,** 397–406.
33. Argos, P., Rossman, M. G., Grau, U. M., Zuber, H., Frank, G. & Tratshin, J. D. (1979) *Biochemistry* **18,** 5698–5703.
34. Gilson, M. K., Sharp, K. A. & Honig, B. (1988) *J. Comp. Chem.* **9,** 327–335.
35. Gilson, M. K. & Honig, B. (1988) *Proteins* **4,** 7–18.
36. Sharp, K., Fine, R. & Honig, B. (1987) *Science* **236,** 1460–1463.
37. Burley, S. K. & Petsko, G. A. (1986) *FEBS Lett.* **203,** 139–143.
38. Matsumura, M., Becktel, W. J. & Matthews, B. W. (1988) *Nature (London)* **334,** 406–410.
39. Matsumura, M., Wozniak, J. A., Dao-Pin, S. & Matthews, B. W. (1989) *J. Biol. Chem.* **264,** 16059–16066.
40. Lim, W. A. & Sauer, R. (1989) *Nature (London)* **339,** 31–36.
41. Lim, W. A., Hodel, A., Sauer, R. T. & Richards, F. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 423–427.
42. Alber, T. (1989) *Annu. Rev. Biochem.* **58,** 765–798.
43. Matthews, B. W., Nicholson, H. & Becktel, W. J. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 6663–6667.
44. Herning, T., Yutani, K., Inaka, K., Kuroki, R., Matsushima, M. & Kikuchi, M. (1992) *Biochemistry* **31,** 7077–7085.