Supporting information for Pavlov *et al.* (2002) *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.202127199

Text

Construction, Expression, and Purification of Chimeric DNA Polymerases. All plasmids were constructed by common subcloning techniques and propagated in DH5α (Invitrogen) strain of *Escherichia coli*.

TaqTopopET21d. The polymerase chain reaction was used to amplify segments of the *M. kandleri top5* gene covering amino acids 685–984 from pAS6.5 plasmid (1) and the *Taq* polymerase gene covering amino acids 290–832 (Stoffel fragment) from pTTQ plasmid (gift of G. Belov). In the case of the *top5* gene fragment a linker 5'-GCCTACGACGTAGGCGCC-3' (translated into AYDVGA) was added at the 3' end of the fragment. The 5' end of the *top5* gene fragment contained the *Nde*I restriction site with the initiating AUG codon, while two stop codons were placed at the 3' end of the *Taq* fragment followed the *Hin*dIII restriction site. The 3' end of the top5 fragment was blunt ligated to the 5' end of the *Taq* fragment, digested with *Nde*I–*Hin*dIII, and the resulting DNA was cloned into the pET21d expression vector (Novagen).

TaqTopoC1-pET21d, TaqTopoC2-pET21d, and TaqTopoC3-pET21d. The *Taq* polymerase gene fragment covering amino acids 279–832 was amplified by PCR from pTTQ plasmid using primers with incorporated *Eco*RI and *Hin*dIII sites. The 1,684-bp fragment was then digested with *Eco*RI and *Hin*dIII and cloned into *Eco*RI–*Hin*dIII digested pBlueScript KSII vector (Stratagene) to yield the Stoffel-BS vector. Next, segments of the *top5* gene covering amino acids 384–984 (C1), 518–984 (C2), and 676–984 (C3) and including the *top5* terminating codon were PCR amplified from pAS6.5 plasmid using primers with incorporated *Hin*dIII and *Sal*I sites. These PCR products were digested with *Hin*dIII and *Sal*I and subcloned into the pBS KSII vector (Stratagene) to yield StoffelC1-C3 vectors. The inserts were cut out by *Hin*dIII and *Sal*I digestion and cloned into the *Hin*dIII–*Sal*I digested Stoffel-BS plasmid making StoffelC1-C3 fusions with AAGCTT (*Hin*dIII site) linker sequence. The resulting combined StoffelC1-C3 inserts were cut out by *Nco*I (the *Nco*I site was introduced by PCR primer used for

generating Stoffel-BS plasmid) and *Sal*I and cloned into pET21d vector to result in expression vectors TaqTopoC1-pET21d, TaqTopoC2-pET21d, and TaqTopoC3-pET21d. All subcloned DNA were sequenced to confirm proper position of initiation signals and the absence of adverse mutations.

PfuC2-pET21d. *Pfu* DNA polymerase cds (2,325 bp) was subdivided into two parts, 978- and 1,353-bp-long, and each one was individually PCR-amplified from *Pyrococcus furiosus* genomic DNA. The *Nco*I–*Eco*RI-digested upper PCR fragment (*Nco*I site was introduced in the PCR primer) was cloned into *Nco*I–*Eco*RI sites of modified pBlueScript II SK- vector (the modified vector carries *Nco*I–*Bgl*II recognition sites inserted between *Pst*I and *Eco*RI sites of the polylinker sequence). The *Eco*RI–*Hin*dIII-incompletely digested lower PCR fragment (*Hin*dIII site was introduced in the primer, an additional *Hin*dIII site is present in the *Pfu* cds) was cloned into *Eco*RI–*Hin*dIII sites of the modified pBlueScript II SK vector. Sequencing of several upper and lower inserts revealed clones carrying the correct sequences. The upper insert was cloned in the *Nco*I and *Eco*RI sites of the plasmid, which already carried the lower insert, thus joining both parts of the *Pfu* cds together. The *Pfu* cds was cut out by *Nco*I–*Hin*dIII digestion (*Hin*dIII-digestion was incomplete), the C-terminal TopoV-C2 domain was cut out from the Top5-C2-pBlueScript II SK plasmid by *Hin*dIII–*Sal*I double digestion, and both parts were ligated with *Nco*I–*Sal*I-digested pET21d vector. The resulting expression construct was verified by restriction digestion. The final protein starts with Met-Val instead of Met-Ile (as it is in the wild-type *Pfu* polymerase) at its N terminus and contain the Lys-Leu linker between *Pfu* polymerase and the Topo V's C2 domain.

E. coli strain BL21 pLysS (Novagen) was transformed with expression plasmids. For each DNA polymerase, 2 liters of LB medium containing 100 μg/ml ampicillin and 34 μg/ml chloramphenicol was inoculated with transformed cells, and the protein expression was induced by adding 1 mM isopropylthio-β-galactoside (IPTG) and carried out at 37°C for 3 h. The cells were harvested and dissolved in 100 ml lysis buffer containing 50 mM Tris•HCl pH 8.0, 1mM EDTA, 5 mM β-mercaptoethanol, and protease inhibitors (Roche Applied Science). The lysate was centrifuged at 12,000 × *g* for 30 min, heated at 75ºC for 30 min, and centrifuged again at 15,000 × *g* for 1 h. The supernatant was filtered through a 0.22 μm Millipore filter and applied on a heparin high trap column (Amersham

Pharmacia Biotech) equilibrated with 0.5 M NaCl in 50 mM Tris•HCl buffer, pH 8.0, containing 2 mM β-mercaptoethanol. The column was washed with 100 ml of the same buffer, and the protein was eluted in 20 ml of 0.75 M NaCl in 50 mM Tris•HCl buffer, pH 8.0, with 2 mM β-mercaptoethanol.

Steady-state Kinetics of DNA Synthesis in Primer Extension Reactions. An enzymatic polymerization in primer extension reactions can be described by the following scheme:

$$\begin{array}{ccccccccccc} & & k_{b(0)} & & k_1 & & k_2 & & k_i & & k_{n\text{-}1} \\ E + S_0 & \longleftrightarrow & E^*S_0 & \longrightarrow & E^*S_1 & \longrightarrow & E^*S_2 & \ldots & \longrightarrow & E^*S_i \ldots & \longrightarrow & E + P_n. \\ & & k_{b(\text{-}0)} & & k_{b(1)} \Updownarrow k_{b(\text{-}1)} & & k_{b(2)} \Updownarrow k_{b(\text{-}2)} & & k_{b(\text{i})} \Updownarrow k_{b(\text{-}i)} & & & \end{array} \quad (1)$$

$$E + S_1 \qquad E + S_2 \qquad E + S_i$$

Using Scheme 1, one can derive a balance equation for the polymerase–substrate complexes in primer extension reactions:

$$[E_t] = [E] + [ES_0] + [ES_1] + \ldots + [ES_n], \quad (2)$$

where $[E_t]$ is total concentration of the active polymerase.

Also, at steady-state that occurs within several seconds for short PTJ substrates (2), for each consecutive step of elongation i, it is possible to write an equation:

$$d[ES_i]/dt = k_{b(i)} * [E]*[S_i] + k_i*[ES_{i\text{-}1}] - [ES_i]*(k_{b(\text{-}i)} + k_{i+1}) = 0, \quad (3)$$

where $k_{b(i)}$, and $k_{b(\text{-}i)}$ are the rate constants for binding PTJs (S_i) to the protein and dissociation of the corresponding complexes ES_i, and k_i and k_{i+1} are rate constants for the elongation that produces and extends these complexes, subsequently.

The steady-state concentration of a productive complex ES_i is then expressed using the recursive relationship:

$$[ES_i] = \frac{k_{b(i)} * [E] * [S_i]}{(k_{b(-i)} + k_{i+1})} + \frac{k_i * [E_{i-1}]}{(k_{b(-i)} + k_{i+1})} \ , \qquad (4)$$

which, as a result, produces, as a result subsequent replacements:

$$[ES_i] = \frac{[E] * [S_0]}{Km_0} * \prod_{j=1}^{i} \frac{k_j}{(k_{b(-j)} + k_{j+1})} \quad (5),$$

where

$$Km_0 = \frac{(k_{b(-0)} + k_1)}{k_{b(0)}}.$$

Since the rate of addition of the first nucleotide to PTJ, $\mathbf{v}_1 = k_1 * [ES_0]$, substituting $[ES_i]$ in (2) by (5) and rearranging produces a Michaelis–Menten equation:

$$\mathbf{v}_1 = \frac{k_{app} \cdot [E_t] \cdot [S_0]}{(Km_{app} + [S_0])} \ , \qquad (6)$$

where

$$k_{app} = \frac{k_1}{\sum_{i=1}^{n} \prod_{j=1}^{i} \frac{k_j}{(k_{b(-j)} + k_{j+1})}} \text{ and } Km_{app} = \frac{Km_0}{\sum_{i=1}^{n} \prod_{j=1}^{i} \frac{k_j}{(k_{b(-j)} + k_{j+1})}}.$$

At sufficiently low concentrations of S_0, ($Km_{app} >> [S_0]$), then (6) becomes:

$$\mathbf{v}_1 \approx \frac{k_{app}}{Km_{app}} * [E_t] * [S_0] \ , \qquad (7)$$

and the initial rates of reactions are proportional to [PTJ].

Affinity of *Taq* DNA polymerase or its fragments to various DNA substrates was not clearly defined in many cases. Using results obtained in pre-steady state experiments (2), it is possible to conclude that, in the reaction with addition of first deoxynucleotide to PTJ, K_m (DNA) would be higher than 0.1 μM for *Taq* polymerase and higher than 2.2 μM for the fragments. Also, K_m could be sensitive to the structure of PTJ duplex substrate. We measured initial rates of primer extension reactions using a wide range of PTJ concentrations to assess the affinity of the hybrid enzyme TopoTaq to DNA (Fig. 5).

The initial rates of primer extension by *Taq* DNA polymerase and TopoTaq were almost proportional to the substrate concentration up to1μM PTJ (Fig. 6*A*). Therefore, the apparent K_ms for these reactions were higher than 2 μM. The rates were also proportional to total protein concentration, as shown in Fig. 6*B*. Since, according to Scheme 1, $\frac{k_{app}}{Km_{app}} = \frac{k_{b(0)} \cdot k_1}{(k_{b(-0)} + k_1)}$, the slope of a line in Fig 2*A* at low [PTJ] is equal to $\frac{k_{b(0)} \cdot k_1}{(k_{b(-0)} + k_1)} * [E_t]$, where $k_{b(0)}$, and $k_{b(-0)}$ are rate constants for binding S_0 to the protein and dissociation of the corresponding complex ES_0 in Scheme1, and k_1 is the rate constant for the incorporation of the first nucleotide, subsequently.

Furthermore, as $\frac{k_1}{(k_{b(-0)} + k_1)}$ represents a *microscopic processivity parameter* in the reaction of addition of the first nucleotide (3), it is evident that $\mathbf{v}_1 = k_{b(0)} * \boldsymbol{p}_0 * [E_t][S_0]$. Hence, this finding demonstrates that the dependencies of initial rates in Fig. 6 are functions of the change in both the rates of binding of the DNA substrate to the polymerases and the processivity of synthesis. At low salt concentrations, the values of processivity are quite close to unity. Under these conditions, it was possible to estimate rate constants for bimolecular association of the enzymes with PTJ ($k_{b(0)}$) that were in the range $0.4 – 1.6 \cdot 10^6$ M^{-1} s^{-1}.

Determination of Processivity Equivalence Parameter. On heterogeneous templates, each position on the template has an individual value of microscopic processivity that has its own sensitivity to salts. An example is shown in Fig. 7.

The premature termination of extension renders the geometric mean microscopic processivity parameter,

$$P = \sqrt[n]{\prod_{i=0}^{n-1} \boldsymbol{p}_i} \tag{8}$$

(4) calculated for addition of defined number of nucleotides zero.

It was shown theoretically (4) that the probability of producing a primer extended by exactly n residues could be written using the microscopic processivity parameter

$$\boldsymbol{p_i} \text{ as: } \mathbf{P}_n = (1 - \boldsymbol{p}_n) * \prod_{i=0}^{n-1} \boldsymbol{p}_i \ . \tag{9}$$

Therefore, by definition of the average, the average length of extension (***Lav***) would be equal to

$$\boldsymbol{Lav} = \sum_{n=1}^{\infty} n \cdot \mathbf{P}_n \ . \tag{10}$$

At increased salt concentrations only short extended products are present in the reaction mixture. It means that $\boldsymbol{p_i}$ for longer extension products decreases to zero within the experimental error (Fig. 7). Then the average length of extension per polymerase binding event can be calculated directly from experimental values of $\boldsymbol{p_i}$:

$$\boldsymbol{Lav} = \sum_{n=1}^{n_{\max}} n \cdot \mathbf{P}_n \ , \tag{11}$$

where $n$max is the maximum number of nucleotide attachments allowed by the template. However, for highly processive synthesis, the average extension per binding event could be greater than the physical length of the template. In this case:

$\boldsymbol{Lav} = \sum_{n=1}^{n_{\max}} n \cdot \mathbf{P}_n + \sum_{n=n_{\max}+1}^{\infty} n \cdot \overline{\mathbf{P}}_n$ (11), where $\overline{\mathbf{P_n}} = (1-P) * P^n$ [P is the geometric mean microscopic processivity parameter (4)] was calculated. ***Lav*** here can be considered to be

the average extension per binding event on an infinite hypothetical substrate that has nmax nucleotides of the original sequence continued by a homopolymer tail, on which DNA polymerase synthesize with processivity equal to the geometric mean microscopic processivity parameter of the original sequence. Then, since for a homopolymer, $\boldsymbol{Lav} = 1/(1 - P)$ (5), we define the *processivity equivalence parameter*, $\boldsymbol{P_e} = 1 - 1/\boldsymbol{Lav}$. For practical purpose of calculation, in equation (11) the sum of infinite number of terms is replaced by a sum of finite number of terms. We usually stopped the summation as soon as the next added term would become less than $10^{-5} \cdot \boldsymbol{Lav}$.

Modeling Domains of TopoV and Design of Chimeras. Although the crystal structure of Topo V is not known, current biochemical information suggests that the HhH motifs of the protein are folded into distinct units, which are further organized into bigger structures as it was revealed by limited proteolysis (1, 6). We attempted to use computer modeling for 3D structures of the individual TopoV HhH domains based on structural information obtained for other proteins with HhH domains. Use of protein 3D modeling servers, such as SwissModel (7–10) or Geno3D (11) with the automatic mode of sequence recognition allowed only for modeling of TopoV domain G because of its high similarity to RuvA DNA binding domain. In all other cases, low sequence similarity of TopoV domains to the proteins with known structures prevented finding out a proper template for modeling. Therefore, the structural data bank was screened for non-redundant proteins with double HhH repeats (the majority of structures already existed in EBI and NCBI databases); also, structures found by Shao and Grishin (12) were added. The found proteins were checked against Fold classification based on Structure–Structure alignment of Proteins [FSSP (13)] database for closely related proteins. If structures of the corresponding protein-DNA complexes with resolution <3 Å existed, then these were used instead of the structures of individual proteins. Seven proteins were found and used as templates: 1bpy (human DNA polymerase β), 1c7y (holiday junction DNA helicase RUVA; *E. coli*), 1coo (RNA polymerase alpha subunit; *E. coli*), 1dgs (NAD^+-dependent DNA ligase; *Thermus filiformis*), 1ebm (human 8-oxoguanine DNA glycosylase), 2abk (endonuclease III; *E. coli*), and 2pjr (helicase PCRA; *Bacillus stearothermophilus*). The structures of HhH domains excised from the Protein Data Bank files served as templates for TopoV domain modeling with SwissModel server. All TopoV HhH domains, except domain J, were successfully modeled with at least one of the templates tried. If the server

suggested several structures, the one with the lowest calculated free energy was chosen. Domain J that had a too short similarity range to 1dgs, according to parameter of the server, was folded by Swiss-PDBViewer using the 1dgs structure as a template, followed by an energy minimization procedure in vacuum with GROMOS (14). Domain L was found to have two overlapping parts; one (amino acids 910-940 in TopoV) was similar to the DNA ligase HhH domain (1dgs), while the other (amino acids 959-984) was similar to the N-terminal HhH domain of human DNA polymerase β. The intermediate loop (amino acids 941-958) could be folded using both templates, and it had shown almost identical conformation in both cases. Therefore, the two folded parts of the domain were joined in an orientation that provided the best overlay of the residues in the intermediate loop, and the resulting structure was subjected to the energy minimization procedure.

Fig. 10 summarizes the results of TopoV domain modeling, along with structural alignment of TopoV's HhH motifs with the template HhH domains. TopoV repeats B, D, E, J, K, and L could be folded using the ligase (1dgs) HhH domain as a template. Sequences in the domains C, G, H, and I had similarity to the helicase RuvA (1c7y) HhH fold. Domains A and L had similarity to polymerase β (1bpy); domain F was found to be similar to one in helicase PCRA (2pjr). No similarity was detected with HhH domains from RNA polymerase alpha subunit (1coo), glycosylase (1ebm), or endonuclease (2abk).

As the structure of the helicase RuvA complex with DNA is known (1c7y), we located the conserved amino acid of the TopoV domains that correspond to the amino acids of RuvA in contact with DNA, according to the structural alignment. Those were found in domains C, G, and H, but not in I. The sequence in domain F was similar to a region in helicase PCRA (2pjr) that did not include any amino acid residue contacting with DNA. Likewise, the domain A had similarity to a part of polymerase β that did not contact DNA in 1bpy. Six out of twelve TopoV domains have similarity to the ligase HhH domain, however the structure of the ligase-DNA complex is unknown. Moreover, the commonly used structural alignments produced by programs DALI (15) or VAST (16) did not show any similarity of the specific part of 1dgs, which was chosen by both SwissModel and Geno3D servers as a template for TopoV, to any other protein with known contacts with DNA. However, we successfully used a combinatorial extension (CE) approach (17)

provided by a server at San Diego Supercomputer Center (http://cl.sdsc.edu/ce.html) and obtained the structural alignment of the ligase with helicase RuvA and polymerase β (structures 1c7y and 1bpy). Fig. 11 displays the structural alignment found by CE. It is important that the ligase HhH domain, which is the one with the highest similarity to TopoV domains, also contains a well conserved amino acid sequence that is responsible for DNA binding in polymerase β. It seems very likely that this sequence (colored blue in Fig. 10) binds DNA in the ligase and in similarly folded domains of TopoV. Consequently, we located regions of TopoV domains B, D, E, J, K, and L, which have ligase-like folds, and marked the similar conservative residues along with adjacent basic amino acid residues, as the expected sites for DNA binding.

We designed four chimeric proteins consisting of the catalytic (Stoffel) fragment of *Taq* DNA polymerase and three C-terminal amino acid sequences of TopoV, which include repeats B-L, E-L, and H-L, respectively. These sequences sequentially encompass the complete structures produced by the HhH domains in TopoV, as revealed by limited proteolysis (6), starting from the COOH-terminal H-L formation. As in TopoV, we attached the three sequences to the COOH-termini of the polymerase domain.

It is known that *Taq* polymerase contains an HhH fold in the 5' → 3' exonuclease domain; however no direct contacts of this structure with DNA have been demonstrated. The X-ray structure of *Taq* polymerase with DNA shows the conformation of the protein with the HhH domain at distant position with respect to the DNA substrate (1tau, "open" conformation). In contrast, the x-ray structure of the protein with the HhH domain in proximity to the polymerase active site is solved without the DNA (1cmw, "closed" conformation). We attempted to overlay the catalytic domains of these structures and position the DNA substrate from 1tau into 1cmw. It was possible further to bring the *Taq* HhH domain in contact with the DNA, after a relatively small turn of the entire exonuclease domain. The resulting structure is shown in Fig. 13*A*. Similarly, a sequence containing HhH domains H-L could be attached through a suitable linker to the NH_2-termini of the catalytic domain of *Taq* polymerase to bring the TopoV HhH domain L in proximity to the DNA substrate (Fig. 13*B*). Therefore, a TopoTaq chimera was designed, such that the entire TopoV structure containing HhH motifs H-L has been fused with NH_2-termini of the Stoffel fragment through a linker. In addition, a significant homology

of the entire 5' → 3' exonuclease domain of *Taq* polymerase to the sequence containing the TopoV HhH repeats has been found, which might provide better interactions of the TopoV polypeptide with the catalytic domain of *Taq* polymerase (Fig. 14).

1. Belova, G. I., Prasad, R., Kozyavkin, S. A., Lake, J. A., Wilson, S. H. & Slesarev, A. I. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 6015–6020.

2. Brandis, J. W., Edwards, S. G. & Johnson, K. A. (1996) *Biochemistry* **35,** 2189–2200.

3. von Hippel, P. H. & Yager, T. D. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 2307–2311.

4. Fairfield, F. R., Newport, J. W., Dolejsi, M. K. & von Hippel, P. H. (1983) *J. Biomol. Struct. Dyn.* **1,** 715–727.

5. von Hippel, P. H., Fairfield, F. R. & Dolejsi, M. K. (1994) *Ann. NY Acad. Sci.* **726,** 118–131.

6. Belova, G. I., Prasad, R., Nazimov, I. V., Wilson, S. H. & Slesarev, A. I. (2002) *J. Biol. Chem.* **277,** 4959–4965.

7. Peitsch, M. C. (1996) *Biochem. Soc. Trans.* **24,** 274–279.

8. Guex, N. & Peitsch, M. C. (1997) *Electrophoresis* **18,** 2714–2723.

9. Guex, N., Diemand, A. & Peitsch, M. C. (1999) *Trends Biochem. Sci.* **24,** 364–367.

10. Schwede, T., Diemand, A., Guex, N. & Peitsch, M. C. (2000) *Res. Microbiol.* **151,** 107–112.

11. Combet, C., Jambon, M., Deleage, G. & Geourjon, C. (2002) *Bioinformatics* **18,** 213–214.

12. Shao, X. & Grishin, N. V. (2000) *Nucleic Acids Res.* **28,** 2643–2650.

13. Holm, L. & Sander, C. (1996) *Nucleic Acids Res.* **24,** 206–209.

14. Brunne, R. M., van Gunsteren, W. F., R., B. & Ernst, R. R. (1993) *J. Am. Chem. Soc.* **115,** 4764–4768.

15. Holm, L. & Sander, C. (1997) *Nucleic Acids Res.* **25,** 231–234.

16. Gibrat, J. F., Madej, T. & Bryant, S. H. (1996) *Curr. Opin. Struct. Biol.* **6,** 377–385.

17. Shindyalov, I. N. & Bourne, P. E. (1998) *Protein Eng.* **11,** 739–747.