the FEBS Journal

The N-terminal region of the bacterial DNA polymerase PolC features a pair of domains, both distantly related to domain V of the DNA polymerase III τ subunit

Kęstutis Timinskas and Česlovas Venclovas

Institute of Biotechnology, Vilnius University, Lithuania

Keywords
clamp loader; DNA polymerase; DNA replication; homology detection; template-based modeling

Correspondence
Č. Venclovas, Institute of Biotechnology, Vilnius University, Graičiūno 8, LT-02241 Vilnius, Lithuania
Fax: +370 5 260 2116
Tel: +370 5 269 1881
E-mail: venclovas@ibt.lt
Website: http://www.ibt.lt/bioinformatics

(Received 27 May 2011, revised 30 June 2011, accepted 6 July 2011)

doi:10.1111/j.1742-4658.2011.08236.x

PolC is one of two essential replicative DNA polymerases in *Bacillus subtilis* and other Gram-positive bacteria. The 3D structure of PolC has recently been solved, yet it lacks the N-terminal region. For this PolC region of ~ 230 residues, both the structure and function are unknown. In the present study, using sensitive homology detection and comparative protein structure modeling, we identified, in this enigmatic region, two consecutive globular domains, PolC-NI and PolC-NII, which are followed by an apparently unstructured linker. Unexpectedly, we found that both domains are related to domain V of the τ subunit, which is part of the bacterial DNA polymerase III holoenzyme. Despite their common homology to τ, PolC-NI and PolC-NII exhibit very little sequence similarity to each other. This observation argues against simple tandem duplication within PolC as the origin of the two-domain structure. Using the derived structural models, we analyzed residue conservation and the surface properties of both PolC N-terminal domains. We detected a surface patch of positive electrostatic potential in PolC-NI and a hydrophobic surface patch in PolC-NII, suggesting their possible involvement in nucleic acid and protein binding, respectively. PolC is known to interact with the τ subunit, however, the region responsible for this interaction is unknown. We propose that the PolC N-terminus is involved in mediating the PolC-τ interaction and possibly also in binding DNA.

Introduction

Genome replication in bacteria is carried out by the multicomponent protein machine, DNA polymerase III [1]. The actual DNA synthesis is performed by the catalytic α-subunit (PolIIIα), which belongs to the C-family of DNA polymerases [2]. Polymerases of the C-family fall into two major groups, DnaE and PolC, typified respectively by *Escherichia coli* PolIIIα and *Bacillus subtilis* PolC. DnaE and PolC can be readily distinguished by the different composition and arrangement of conserved modules. *E. coli*, similar to many other Gram-negative bacteria, possesses DnaE as its sole replicative polymerase. By contrast, Gram-positive bacteria such as *B. subtilis* have both PolC and DnaE. In *B. subtilis*, both polymerases have been shown to be essential for the elongation step in DNA replication [3]. Initially, it was proposed that PolC is responsible for leading strand synthesis, whereas DnaE replicates the lagging strand [3]. However, recent experiments with the reconstituted *B. subtilis* replisome [4] showed that the division of labor between PolC and DnaE is of a different nature. DnaE, much like eukaryotic DNA polymerase α, initially extends an

Abbreviations
OB, oligonucleotide/oligosaccharide-binding; PDB, Protein Data Bank; PHP, polymerase and histidinol phosphatase; RbfA, ribosome binding factor A.

RNA primer followed by more extensive rapid elongation by PolC [4]. These new results highlight the differences in *B. subtilis* and *E. coli* DNA replication at the elongation step, including the different interactions that coordinate leading and lagging strand synthesis.

Although bacterial DNA replication has been studied for decades, the first experimental structures of C-family polymerases were determined only a few years ago. DnaE representatives include full-length *Thermus aquaticus* [5,6] and C-terminally truncated *E. coli* [7] PolIIIα structures, whereas PolC is represented by the structure of *Geobacillus kaustophilus* replicative polymerase [8].

Gram-negative and Gram-positive bacteria separated over a billion years ago [9], providing ample time for divergent evolution of DnaE and PolC. However, despite the rearrangement of some domains and significant divergence at the sequence level, DnaE and PolC have many features in common. Both have a similar polymerase core consisting of 'palm', 'thumb' and 'fingers' domains. The polymerase core in both DnaE and PolC is flanked by a polymerase and histidinol phosphatase (PHP) domain on the N-terminal side, and by a tandem helix–hairpin–helix motif followed by the β-clamp binding motif on the C-terminal side. The PHP domain in some DnaEs of thermophylic bacteria exhibits Zn^{2+}-dependent 3′–5′ exonuclease activity [6,10], although this enzymatic activity is not universally conserved [8,11]. The tandem helix–hairpin–helix motif has been shown to be a major double-stranded DNA binding determinant in the *E. coli* DnaE [12]. Crystal structures revealed that this motif binds double-stranded DNA similarly in both PolC [8] and DnaE [6]. The β-clamp binding motif mediates interaction with the β-clamp [13], which confers processivity on the replicative polymerase by tethering it to DNA.

There are three major differences between DnaE and PolC at the domain level. These include the proofreading 3′–5′ exonuclease domain, oligonucleotide/oligosaccharide-binding (OB) domain and the additional N- and C-terminal regions in PolC and DnaE, respectively. The PolC proofreading 3′–5′ exonuclease domain is inserted into the PHP domain and is an integral part of the polypeptide chain, whereas DnaE uses a separate proofreading subunit, ε [14]. Interestingly, the interaction between DnaE and ε is mediated by the PHP domain [15]. Thus, it may well be that the DnaE-bound ε and the intrinsic ε-like PolC domain represent structurally similar arrangements. The OB domain is present in both DnaE and PolC, but in opposite sequence regions. In DnaE, it is located next to the β-clamp binding site and close to the C-terminus. By contrast, the PolC OB domain is close to the N-terminus immediately preceding the PHP domain. However, it is interesting to note that, in 3D structures of DnaE and PolC, corresponding OB domains occupy positions that are much closer in space than might be expected from their distinct location in sequence. This suggests that the OB domain may play a similar role in binding the incoming template in both PolC and DnaE. The ability to bind single-stranded DNA has indeed been demonstrated for the *E. coli* DnaE OB domain [12,16]. The very N-terminal region of PolC and the C-terminal domain of DnaE appear to be specific for each type of polymerase. The small α/β C-terminal domain of DnaE has been shown to be responsible for binding the clamp loader τ subunit [13]. This interaction is critical for retaining DnaE within the replisome and for its recycling after the completion of each Okazaki fragment on the lagging strand. The experimental structure of the PolC N-terminal region (Pfam PF11490; ~ 230 residues) is not available because it has been removed in the crystallized PolC construct [8]. The function of this region is also unknown, except for the fact that its removal does not compromise core polymerase activity *in vitro* [8].

In the present study, we used sensitive homology detection methods in combination with comparative protein modeling to explore the structure of the PolC N-terminal region. We found that this region includes two consecutive structural domains. Both domains are distantly related to the structure of domain V of the clamp loader subunit τ. The identified relationship coupled with the results of functional analysis and structural considerations suggests an important role for the PolC N-terminal region in interacting with other components of the replisome and possibly DNA.

Results

Sequence searches identify two type II K homology (KH) fold-like domains within the PolC N-terminal region

For the PolC N-terminal region of ~ 230 residues, neither 3D structure nor function are known. It is also one of the least conserved regions in PolC sequences. For example, *B. subtilis* and *G. kaustophilus* full-length PolCs share 74% identical residues, whereas the corresponding N-terminal regions display only 44% sequence identity.

Standard sequence searches using BLAST and PSI-BLAST [17] failed to detect any homology between the N-terminal region of *B. subtilis* PolC (BsuPolC; National Center for Biotechnology Information GI

number: **143342**) and proteins with available 3D structures. Therefore, we turned to more sensitive homology detection methods based on sequence profiles. Thus, HHSEARCH [18] detected similarity between the second half of the BsuPolC N-terminal region (~ 100–200) and both domain V of the DNA polymerase III τ subunit [PolIIIτ-V; Protein Data Bank (PDB) code: 2aya] [19] and the N-terminal domain I of the replication initiator protein DnaA (DnaA-I; PDB: 2e0g) [20]. These structures were detected with high HHSEARCH probability (97% for both), strongly suggesting a common origin. Interestingly, the first half of the PolC N-terminal region (~ 1–100) also detected the PolIIIτ-V domain, albeit weakly (HHSEARCH probability of 16%). The structures of PolIIIτ-V and DnaA-I adopt a variant of the so-called type II KH fold [21]. One of their major differences from classical type II KH domains is the absence of the characteristic GXXG motif (where X denotes any amino acid) involved in nucleic acid binding. Two other profile-based methods, COMA [22] and COMPASS [23], also matched the second half of the PolC N-terminal region with PolIIIτ-V and DnaA-I, producing statistically significant scores (*E*-values $< 10^{-3}$). However, no significant matches were detected for the first half.

To further explore these tentative structural matches, we collected BsuPolC homologs using PSI-BLAST and constructed a multiple sequence alignment for the N-terminal region. The alignment was iteratively refined by removing sequences that were poorly aligned and had long gaps or insertions. Using this refined alignment as an input, the HHSEARCH results for the second half of the PolC N-terminal region were very similar, however, they improved dramatically for the first half. In this case, HHSEARCH detected PolIIIτ-V with a probability of 78%, up from 16%. Because additional sequence regions may sometimes interfere with homology detection, we decided to test whether the removal of the second half of the PolC N-terminus would help to improve the results further. Therefore, we took only the fragment of the multiple sequence alignment covering the first half of the PolC N-terminus (corresponding to residues 1–89 of BsuPolC; residue numbering is based on BsuPolC throughout the present study) and used it as an input into HHSEARCH for searching the PDB. PolIIIτ-V was again detected as the best match, with the probability increasing to 93%.

Taken together, the results of sequence-based searches suggested that the PolC N-terminal region has two adjacent structural domains, both related to PolIIIτ-V. We termed these two putative domains PolC-NI and PolC-NII (Fig. 1). The presence of the two similar domains is also supported by the predicted secondary structure, which consists of two repeating α-α-β-β-α-β topologies. Interestingly, we identified extensive intrinsic disorder within the linker between PolC-NII and the OB domain (approximately residues 170–224). The disorder in this linker region was predicted by three independent approaches (see Materials and methods), with the strongest consensus spanning residues 194–214. These data suggest that the linker connecting the N-terminal two-domain structure to the OB domain of PolC might be quite flexible.

Structural models strongly support the sequence-based homology inference

Sequence-based searches are a powerful tool for homology inference. However, the protein 3D structure provides a more rigorous means for the assess-

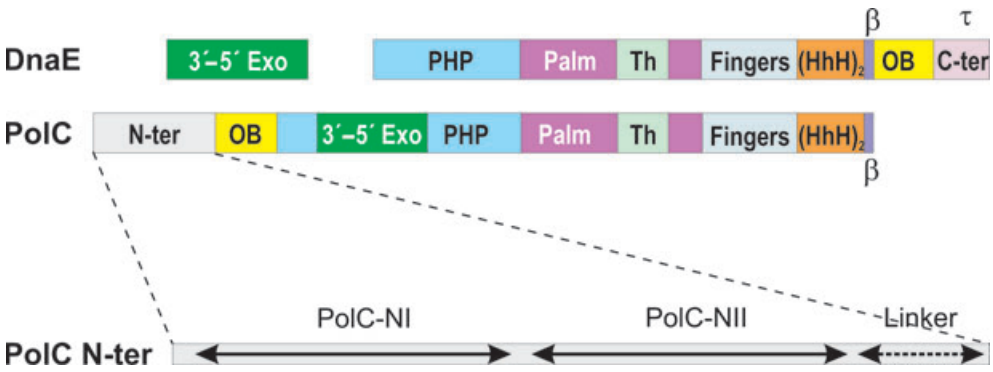

Fig. 1. DnaE and PolC domain architectures. Different domains are denoted by different colors and their common names. $(HhH)_2$, tandem helix–hairpin–helix motif; Th, thumb; C-ter, C-terminal domain; N-ter, N-terminal region. The 3′–5′ proofreading exonuclease activity in DnaE is provided by a separately encoded subunit. Greek letters β and τ indicate experimentally determined sites for binding corresponding subunits of the polymerase III holoenzyme. The expanded view shows the predicted domain composition for the PolC N-terminal region (PolC N-ter), which includes two globular domains (PolC-NI and PolC-NII) and a presumably flexible linker.

ment of any potential evolutionary relationship. In addition, protein structure is usually more informative in the search for a putative function. Therefore, we next constructed structural models for each of the two N-terminal domains.

Homology modeling of PolC-NII was fairly straightforward. Three structures identified in homology searches were used as modeling templates. One of them was the PolIIIτ-V domain (PDB: 2aya) [19] and two others represented DnaA domain I (PDB: 2e0g [20] and 2wp0 [24]). Models were constructed using iterative cycles of modeling and alignment refinement, as described in the Materials and methods. According to the structure assessment with PROSA2003 [25], the obtained models fare comparably to (or even better than) the corresponding experimental structures used as modeling templates (Table 1).

Because the sequence-based results for the PolC-NI domain were less convincing, we considered modeling to be especially useful for scrutinizing the inferred homology for this PolC domain. Initially, we used the structure of PolIIIτ-V (2aya) identified with HHSEARCH as the only modeling template. However, PolC-NI models based on this single template were considered to be inferior to the experimental structure of PolIIIτ-V. This suggested that the structure of PolIIIτ-V may not be the best approximation for the PolC-NI domain. Therefore, we also considered additional structural templates. The obvious choice was to include structures representing the related DnaA-I domain. In addition, we included structures of the ribosome binding factor A (RbfA) family identified by the structure-based search with DALILITE [26] using the structure of PolIIIτ-V as a query. We then used different combinations of structural templates to obtain a large number of PolC-NI models, all of which were assessed with PROSA2003. Somewhat unexpectedly, the assessment results showed that DnaA-I structures did not help to improve models, whereas RbfA structures (PDB: 2dyj [27] and 2e7g) did. After the iterative modeling procedure, the assessment results for the best *B. subtilis* PolC-NI model were slightly worse than for the PolC-NII domain, yet comparable to those for the template structures (Table 1). Additional PolC-NI models constructed for related sequences scored similarly or even better.

To obtain additional reference points for structure evaluation, we constructed homology models for PolIIIτ-V and DnaA-I, based on each other's experimental structure and the 'true' alignment derived from the structure comparison. This represents an idealized distant homology modeling case in which the optimal sequence alignment with the structural template is known beforehand. Notably, according to the PROSA2003 evaluation, PolC-NI models are clearly better than the homology models of either PolIIIτ-V or DnaA-I (Table 1). Thus, the evaluation results suggest that PolC-NI models are quite a reasonable approximation of their native structure.

Taken together, the modeling results reinforced the sequence-based homology finding that both N-terminal domains of PolC are related to domain V of the PolIII

Table 1. PROSA2003 evaluation results. PROSA2003 assessment includes both modeled and experimental structures. In addition to models of *B. subtilis* PolC N-terminal domains, five models of related sequences were evaluated. For experimental structures, the determination technique and the PDB code are indicated. For models, PDB codes in parentheses indicate the templates used in modeling. PROSA2003 *Z*-score represents the estimated energy of the structure (the range of *Z*-scores is for the five additional models). A more negative PROSA2003 energy *Z*-score suggests that the structure is more energetically favorable.

Structure	Type	Length	PROSA2003 *Z*-score
PolC N-terminal domain I			
PolC-NI, *B. subtilis*	Model (based on 2aya, 2dyj, 2e7g)	79	−6.6
PolC-NI, other (5)	Models (based on 2aya, 2dyj, 2e7g)	79	(−6.6; −7.9)
PolC N-terminal domain II			
PolC-NII, *B. subtilis*	Model (based on 2aya, 2e0g, 2wp0)	74	−8.4
PolC-NII, other (5)	Models (based on 2aya, 2e0g, 2wp0)	74	(−7.8; −8.2)
Reference structures			
PolIIIτ-V, *E. coli*	NMR, 2aya	72	−8.0
DnaA-I, *E. coli*	NMR, 2e0g	77	−5.6
DnaA-I, *H. pylori*	X-ray, 2wp0	86	−6.8
RbfA, *T. thermophilus*	X-ray, 2dyj	82	−7.1
RbfA, *Homo sapiens*	NMR, 2e7g	89	−7.1
PolIIIτ-V, *E. coli*	Model (based on 2e0g)	72	−5.3
DnaA-I, *E. coli*	Model (based on 2aya)	77	−5.2

τ subunit. In addition, these results suggested that the PolC-NI structure may be more similar to that of RbfA, whereas PolC-NII may be more similar to DnaA. Interestingly, PolC N-terminal domains are only remotely related to each other. Although the corresponding structural models are fairly similar, their structure-based sequence alignment shows < 10% sequence identity. Moreover, we were unable to detect the similarity between the two PolC N-terminal domains with either HHSEARCH or other sensitive profile-based homology detection methods. Collectively, these observations suggest that the tandem structure is not the result of domain duplication within the PolC but rather has been acquired by PolC, either as an already diverged two-domain structure or, sequentially, one domain at a time, from different parental sources.

Structure and surface properties of PolC N-terminal domains

Although the type II KH fold-like structure and the relationship to domain V of the PolIII τ subunit are convincing for both PolC N-terminal domains, their function is not immediately obvious. At the same time, the established structural similarity with additional functionally characterized domains (e.g. DnaA-I and RbfA) suggests that either of the two domains might be involved in protein–protein interactions and/or nucleic acid binding. To obtain more specific clues regarding the possible function of PolC N-terminal domains, we used their structural models to analyze surface properties, including residue conservation, electrostatic potential and hydrophobicity.

Conserved surface residues in the PolC-NI domain tend to cluster on its N-terminal side, including the N-terminal part of α1-helix, β1-strand and the loops connecting β1 with β2 and α3 with β3 (Fig. 2A,C). Interestingly, this surface region shows an increased positive electrostatic potential. The most conserved positively charged position in BsuPolC corresponds to Lys44. Other moderately conserved positively-charged residues include Lys36 and Lys41. In addition, species of the class *Bacilli* often have one to four Lys or Arg residues in variable positions of the N-terminal part of the α1 helix. These residues also contribute to an elevated positive electrostatic potential. Our PolC-NI structural models revealed several other conserved residues on the surface, including Gln17, Phe11, Leu15 and Ile75. The reason for their conservation is not clear; however, at least for the hydrophobic residues, the possibility that their localization on the surface is a result of inaccuracies in the modeled structures cannot be disregarded. On the other hand, even some positional errors within the cluster of positively-charged residues in PolC-NI would not alter its surface electrostatic properties significantly. Therefore, the patch of an increased positive electrostatic potential appears to be the most distinct feature of the PolC-NI domain surface. In turn, this suggests that the very N-terminal domain of PolC may at least weakly bind DNA or RNA. If so, the putative interaction is likely to be nonspecific because the modeled structure of PolC-NI lacks any prominent clefts that might contribute to the structure or sequence specificity.

The PolC-NII domain does not have a positively-charged surface patch, as was predicted for PolC-NI. Nevertheless, some of the conserved positions are no less intriguing. For example, Trp98 and its neighbor, Tyr97, are highly conserved in the α1 helix (Fig. 2B,D). Notably, Trp98 corresponds to the conserved Trp residue in both *E. coli* PolIIIτ-V (Trp523) and DnaA-I (Trp6). The hydrophobic patch including Trp6 has been implicated in *E. coli* DnaA dimerization [20]. In addition, the same hydrophobic patch in DnaA-I features the conserved Leu10 that corresponds to the similarly conserved Ile102 in PolC-NII. Another highly conserved site includes dipeptide Gly157-Phe158, located in the loop between α3 and β4. The strong conservation of Gly157 suggests severe conformational constraints imposed at this position, making the burial status of Phe158 uncertain. Interestingly, no position is as highly conserved in corresponding loops in either PolIIIτ-V or DnaA-I. One additional moderately conserved surface site corresponds to Thr134 at the N-terminus of the α3 helix. It might be that this residue has been conserved for structural reasons (e.g. specifically as the N-cap for the α3 helix). Alternatively, it might be an interaction site because the corresponding region in *Helicobacter pylori* DnaA-I mediates the interaction with HobA [24]. However, unlike PolC-NII, the DnaA-I surface area for the HobA interaction includes multiple (rather than a single) conserved residues. Overall, the surface analysis suggests that PolC-NII is more likely to participate in mediating protein–protein interactions than in nucleic acid binding.

Discussion

Sensitive sequence profile–profile comparison methods combined with comparative modeling revealed that the N-terminal region of the bacterial replicative polymerase PolC includes two structural domains: PolC-NI and PolC-NII. Both domains are distantly related to domain V of the DNA polymerase III τ-subunit, adopting

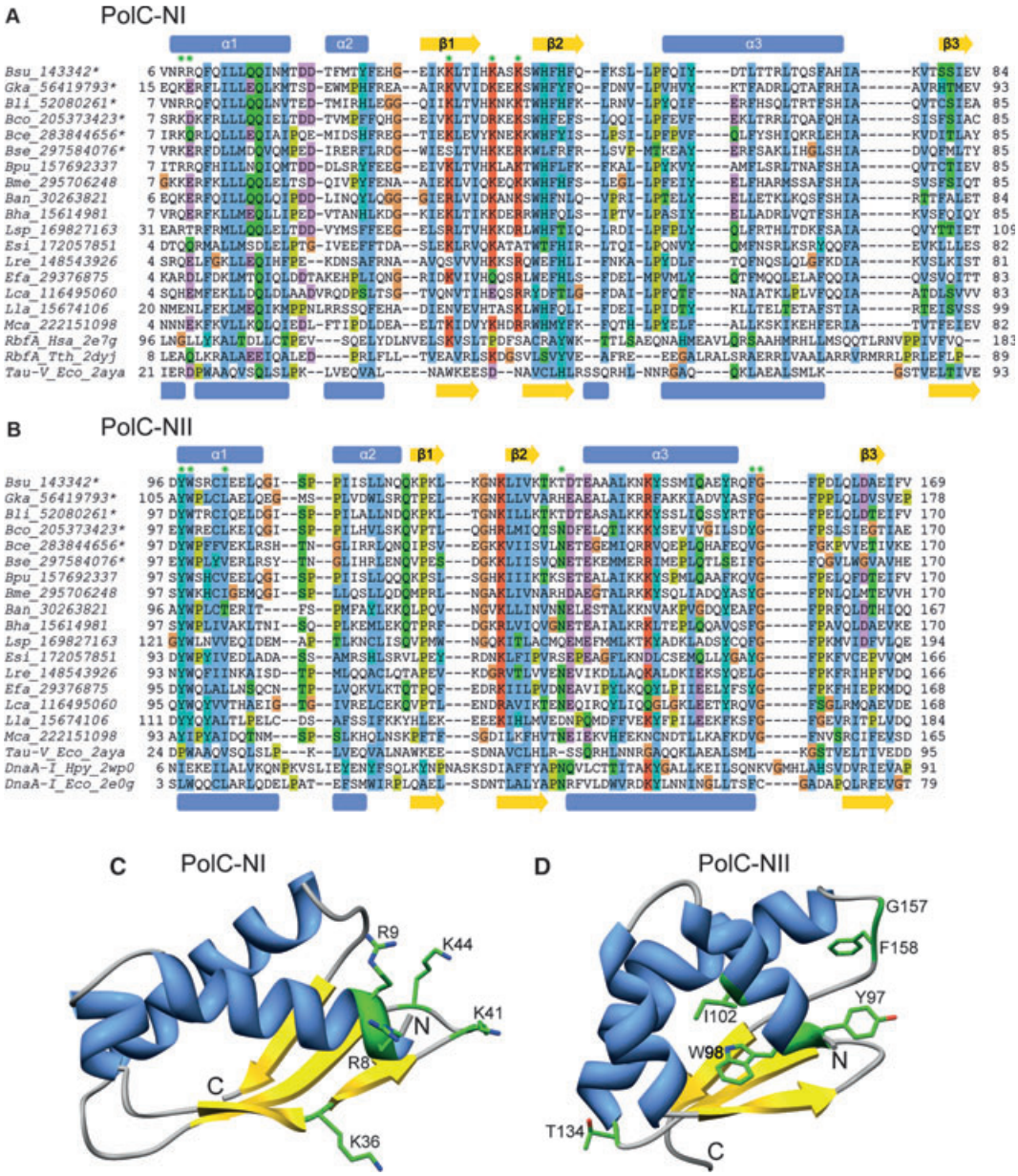

Fig. 2. Sequence alignments and corresponding structural models for the two domains of the PolC N-terminal region. Sequences of the PolC-NI (A) and PolC-NII (B) domains aligned with the structures used for the construction of corresponding structural models (C, D). Labels for PolC sequences include species abbreviation and the GI number. Labels for sequences of experimental structures include the name of the protein, species abbreviation and the PDB code. PolC sequences for which models were constructed are indicated with an asterisk next to the sequence label. Predicted secondary structures for the two domains of the *B. subtilis* PolC sequence (Bsu_143342) are shown above the corresponding alignments, whereas the secondary structures shown below the alignments were derived from the experimental structures of domain V of the *E. coli* τ-subunit (Tau-V-Eco_2aya) (A) and the *E. coli* DnaA-I domain (DnaA-Eco_2e0g) (B). Green stars above the alignments indicate conserved surface residues shown with their side chains in the corresponding structural models of *B. subtilis* PolC-NI (C) and PolC-NII (D) domains. The coordinates of PolC-NI and PolC-NII structural models are available at: http://www.ibt.lt/bioinformatics/models/polc_nterm/.

type II KH fold-like structure. In addition, PolC-NII shows an even higher similarity to domain I of the initiator of chromosomal replication DnaA (DnaA-I).

What might the function of these PolC N-terminal domains be? The involvement of related structures in protein–protein interactions [20,24] and nucleic acid

binding [27] suggests similar functions for these domains. Taking into account the biological context, an obvious hypothesis is that either one or both domains mediate the interaction of PolC with the τ-subunit. It is known that PolC interacts with the clamp loader subunit τ [28–30], however, the region mediating the interaction has not yet been identified. This interaction is relatively weak compared to the corresponding DnaE-τ interaction in *E. coli* [30]. The τ-binding determinants in *E. coli* DnaE have been mapped to the very C-terminus after the OB domain. A single point mutation in this region decreased τ-binding by more than 700-fold [13], whereas the deletion of 48 residues from the C-terminus completely abolished binding [31]. Because PolC does not have the corresponding C-terminal region, its interaction with τ must be mediated by other domains. The N-terminal region, specific to PolC, appears to be the most likely candidate for this role. Both the PolC N-terminal region and the DnaE C-terminal domain are attached to the OB domain, which likely binds the DNA template in both polymerases. Although the exact positions of the corresponding OB domains in PolC [8] and DnaE [5,6] structures differ, the PolC N-terminal region and the DnaE C-terminus may potentially occupy very similar spatial positions with respect to other domains. First, our analysis suggests that the PolC N-terminal region is connected to the OB domain through a flexible linker. Second, the analysis of full-length DnaE crystal structure suggests that both C-terminal and OB domains may be mobile with respect to one another and the other polymerase domains [5]. Collectively, these general structural arguments strongly support a τ-binding role for the PolC N-terminal region.

Our analysis of surface properties suggests that PolC-NII is more likely to be involved in protein–protein interactions, whereas PolC-NI might have a role in nucleic acid binding. Therefore, of the two domains, PolC-NII appears to be more suitable for the putative τ-binding role. Interestingly, the τ subunit in *B. subtilis* and many other Gram-positive bacteria is shorter than that in Gram-negative bacteria such as *E. coli*. The difference in length appears primarily the result of a shorter domain IV, which has been shown to be largely unstructured in *E. coli* and to participate in binding both the replicative helicase [32] and the DNA [33]. One of the possibilities is that PolC-NI contributes to DNA binding to compensate for the shorter domain IV of τ. It also cannot be excluded that one of the PolC N-terminal domains might bind the replicative helicase in addition to binding τ.

In summary, the results obtained in the present study suggest several possible interactions for PolC N-terminal domains. We consider that the corresponding structural models coupled with the analysis of their surface properties provides a useful framework for testing the proposed interactions not only at the domain, but also at the residue level.

Materials and methods

Sequence search and alignment

Standard sequence similarity searches were performed using BLAST and PSI-BLAST [17] with default parameters in locally installed and weekly updated databases of all nonredundant protein sequences ('nr') and sequences corresponding to known protein structures ('pdb'). The 'nr' database was obtained from the National Center for Biotechnology Information (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) and the 'pdb' database was obtained from the PDB (http://www.pdb.org). Sequence searches aimed at the increased sensitivity and accuracy were performed using web server implementations of HHSEARCH [18], COMA [22] and COMPASS [23], which comprise methods based on sequence profile–profile comparison. For all methods except HHSEARCH, an *E*-value of 0.001 or less was considered to represent statistically significant matches. For HHSEARCH, the probability of 95% and higher was considered statistically significant.

Multiple sequence alignments for homologous sequences identified during sequence searches were constructed with MAFFT [34] using the accuracy-oriented L-INS-i algorithm. Visualization and analysis of multiple sequence alignments was carried out using JALVIEW [35].

Structure search and alignment

Structure similarity searches were performed in the PDB database using the DALILITE server [26]. Dali *Z*-scores > 2 were considered to indicate a nonrandom structural similarity. Structure-based alignments were generated from the consensus of three methods: DALILITE [26], TM-ALIGN [36] and FATCAT [37].

Prediction of secondary structure and disordered regions

Predicted secondary structures and natively disordered regions were derived from the consensus of results obtained using several methods. PSIPRED [38], JNET [39] and two variants of PROF [40,41] were used for secondary structure prediction. Disorder prediction was performed using DISOPRED2 [42], IUPRED [43] and POODLE-I [44].

Modeling and assessment of protein 3D structure

Protein structure models were constructed using a slightly modified template-based modeling methodology developed previously [45]. The main feature of this methodology is the iterative improvement of models by optimizing the set of structures used as modeling templates and by refining the query sequence alignment with those templates. The improvement is monitored by the assessment of structural and energy properties of the constructed 3D model. Here, modeling templates were identified by sequence profile-profile searches with HHSEARCH [18], COMA [22] and COMPASS [23]. Additional templates were identified using structure searches with DALILITE [26]. To obtain a set of starting sequence-to-structure alignments, three different profile–profile methods (HHSEARCH, COMA and COMPASS) were used. Four alignment variants were produced with HHSEARCH by changing two parameters: inclusion of secondary structure information (yes/no) and the MAC (maximum accuracy algorithm) parameter set to 0.3 or disabled. Two additional alignments were generated by COMA and COMPASS, respectively. To ensure that alignments would be produced with all the templates, the *E*-value threshold was set to 1000 for COMA and COMPASS, and the probability threshold set to 2% for HHSEARCH. One additional sequence-to-structure alignment was produced in the context of multiple sequence alignment using PROMALS3D [46], a method that is capable of including structural data. Alignment regions showing agreement between all of the methods were considered to be reliable. For the remaining regions, a number of different alignment variants were explored by constructing corresponding models followed by their assessment. Structural models were generated automatically with MODELLER [47] from sequence alignment with the specified structural templates. Models were assessed by estimating their energies with PROSA2003 [25], as well as by using visual inspection for major flaws, such as steric clashes, buried uncompensated charges, etc. Optimization of the template set and the alignment was applied iteratively until energy scores could no longer be improved and no significant defects could be revealed by the visual assessment.

Analysis of surface features and conservation

Residue conservation analysis was performed with the CONSURF server [48] using locally constructed multiple sequence alignments. Sequences for alignment construction were collected by running up to five iterations of PSI-BLAST and then retaining only sequences that are no more than 50% identical to each other in the analyzed region. Sequence filtering was carried out with CD-HIT [49]. Alignments were constructed with MAFFT using the L-INS-i algorithm. Visual analysis of protein surface conservation, electrostatic and hydrophobic properties was performed using UCSF CHIMERA [50].

Acknowledgements

The authors wish to thank Penny Beuning, Digby Warner and Valerie Mizrahi for their useful comments and suggestions. This work was supported by Howard Hughes Medical Institute and Ministry of Education and Science of Lithuania.

References

1 Kornberg A & Baker TA (1992) *DNA Replication*, 2nd edn. WH Freeman, New York.
2 Ito J & Braithwaite DK (1991) Compilation and alignment of DNA polymerase sequences. *Nucleic Acids Res* **19**, 4045–4057.
3 Dervyn E, Suski C, Daniel R, Bruand C, Chapuis J, Errington J, Janniere L & Ehrlich SD (2001) Two essential DNA polymerases at the bacterial replication fork. *Science* **294**, 1716–1719.
4 Sanders GM, Dallmann HG & McHenry CS (2010) Reconstitution of the *B. subtilis* replisome with 13 proteins including two distinct replicases. *Mol Cell* **37**, 273–281.
5 Bailey S, Wing RA & Steitz TA (2006) The structure of *T. aquaticus* DNA polymerase III is distinct from eukaryotic replicative DNA polymerases. *Cell* **126**, 893–904.
6 Wing RA, Bailey S & Steitz TA (2008) Insights into the replisome from the structure of a ternary complex of the DNA polymerase III alpha-subunit. *J Mol Biol* **382**, 859–869.
7 Lamers MH, Georgescu RE, Lee SG, O'Donnell M & Kuriyan J (2006) Crystal structure of the catalytic alpha subunit of *E. coli* replicative DNA polymerase III. *Cell* **126**, 881–892.
8 Evans RJ, Davies DR, Bullard JM, Christensen J, Green LS, Guiles JW, Pata JD, Ribble WK, Janjic N & Jarvis TC (2008) Structure of PolC reveals unique DNA binding and fidelity determinants. *Proc Natl Acad Sci USA* **105**, 20695–20700.
9 Hori H & Osawa S (1979) Evolutionary change in 5S RNA secondary structure and a phylogenic tree of 54 5S RNA species. *Proc Natl Acad Sci USA* **76**, 381–385.
10 Stano NM, Chen J & McHenry CS (2006) A coproofreading Zn(2+)-dependent exonuclease within a bacterial replicase. *Nat Struct Mol Biol* **13**, 458–459.
11 Aravind L & Koonin EV (1998) Phosphoesterase domains associated with DNA polymerases of diverse origins. *Nucleic Acids Res* **26**, 3746–3752.
12 McCauley MJ, Shokri L, Sefcikova J, Venclovas Č, Beuning PJ & Williams MC (2008) Distinct double- and single-stranded DNA binding of *E. coli* replicative DNA polymerase III alpha subunit. *ACS Chem Biol* **3**, 577–587.

13 Dohrmann PR & McHenry CS (2005) A bipartite polymerase-processivity factor interaction: only the internal beta binding site of the alpha subunit is required for processive replication by the DNA polymerase III holoenzyme. *J Mol Biol* **350**, 228–239.
14 Barnes MH, Hammond RA, Kennedy CC, Mack SL & Brown NC (1992) Localization of the exonuclease and polymerase domains of *Bacillus subtilis* DNA polymerase III. *Gene* **111**, 43–49.
15 Wieczorek A & McHenry CS (2006) The NH2-terminal php domain of the alpha subunit of the *Escherichia coli* replicase binds the epsilon proofreading subunit. *J Biol Chem* **281**, 12561–12567.
16 Georgescu RE, Kurth I, Yao NY, Stewart J, Yurieva O & O'Donnell M (2009) Mechanism of polymerase collision release from sliding clamps on the lagging strand. *EMBO J* **28**, 2981–2991.
17 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.
18 Söding J (2005) Protein homology detection by HMM–HMM comparison. *Bioinformatics* **21**, 951–960.
19 Su XC, Jergic S, Keniry MA, Dixon NE & Otting G (2007) Solution structure of domains IVa and V of the tau subunit of *Escherichia coli* DNA polymerase III and interaction with the alpha subunit. *Nucleic Acids Res* **35**, 2825–2832.
20 Abe Y, Jo T, Matsuda Y, Matsunaga C, Katayama T & Ueda T (2007) Structure and function of DnaA N-terminal domains: specific sites and mechanisms in inter-DnaA interaction and in DnaB helicase loading on oriC. *J Biol Chem* **282**, 17816–17827.
21 Grishin NV (2001) KH domain: one motif, two folds. *Nucleic Acids Res* **29**, 638–643.
22 Margelevičius M & Venclovas Č (2010) Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison. *BMC Bioinformatics* **11**, 89.
23 Sadreyev R & Grishin N (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol* **326**, 317–336.
24 Natrajan G, Noirot-Gros MF, Zawilak-Pawlik A, Kapp U & Terradot L (2009) The structure of a DnaA/HobA complex from *Helicobacter pylori* provides insight into regulation of DNA replication in bacteria. *Proc Natl Acad Sci USA* **106**, 21115–21120.
25 Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355–362.
26 Holm L & Rosenstrom P (2010) Dali server: conservation mapping in 3D. *Nucleic Acids Res* **38**, W545–W549.
27 Datta PP, Wilson DN, Kawazoe M, Swami NK, Kaminishi T, Sharma MR, Booth TM, Takemoto C, Fucini P, Yokoyama S *et al.* (2007) Structural aspects of RbfA action during small ribosomal subunit assembly. *Mol Cell* **28**, 434–445.
28 Noirot-Gros MF, Dervyn E, Wu LJ, Mervelet P, Errington J, Ehrlich SD & Noirot P (2002) An expanded view of bacterial DNA replication. *Proc Natl Acad Sci USA* **99**, 8342–8347.
29 Bruck I & O'Donnell M (2000) The DNA replication machine of a gram-positive organism. *J Biol Chem* **275**, 28971–28983.
30 Bruck I, Georgescu RE & O'Donnell M (2005) Conserved interactions in the *Staphylococcus aureus* DNA PolC chromosome replication machine. *J Biol Chem* **280**, 18152–18162.
31 Kim DR & McHenry CS (1996) Biotin tagging deletion analysis of domain limits involved in protein-macromolecular interactions. Mapping the tau binding domain of the DNA polymerase III alpha subunit. *J Biol Chem* **271**, 20690–20698.
32 Gao D & McHenry CS (2001) tau binds and organizes *Escherichia coli* replication proteins through distinct domains. Domain IV, located within the unique C terminus of tau, binds the replication fork, helicase, DnaB. *J Biol Chem* **276**, 4441–4446.
33 Jergic S, Ozawa K, Williams NK, Su XC, Scott DD, Hamdan SM, Crowther JA, Otting G & Dixon NE (2007) The unstructured C-terminus of the tau subunit of *Escherichia coli* DNA polymerase III holoenzyme is the site of interaction with the alpha subunit. *Nucleic Acids Res* **35**, 2813–2824.
34 Katoh K, Misawa K, Kuma K & Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–3066.
35 Waterhouse AM, Procter JB, Martin DM, Clamp M & Barton GJ (2009) Jalview version 2 – a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191.
36 Zhang Y & Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302–2309.
37 Ye Y & Godzik A (2004) FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res* **32**, W582–W585.
38 Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 195–202.
39 Cuff JA & Barton GJ (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* **40**, 502–511.
40 Rost B (2001) Review: protein secondary structure prediction continues to rise. *J Struct Biol* **134**, 204–218.
41 Ouali M & King RD (2000) Cascaded multiple classifiers for secondary structure prediction. *Protein Sci* **9**, 1162–1176.

42 Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF & Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **337**, 635–645.
43 Dosztanyi Z, Csizmok V, Tompa P & Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* **347**, 827–839.
44 Hirose S, Shimizu K & Noguchi T (2010) POODLE-I: disordered region prediction by integrating POODLE series and structural information predictors based on a workflow approach. *In Silico Biol* **10**, 0015.
45 Venclovas Č & Margelevičius M (2009) The use of automatic tools and human expertise in template-based modeling of CASP8 target proteins. *Proteins* **77**(Suppl 9), 81–88.
46 Pei J, Tang M & Grishin NV (2008) PROMALS3D web server for accurate multiple protein sequence and structure alignments. *Nucleic Acids Res* **36**, W30–W34.
47 Šali A & Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779–815.
48 Ashkenazy H, Erez E, Martz E, Pupko T & Ben-Tal N (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* **38**, W529–W533.
49 Li W & Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659.
50 Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC & Ferrin TE (2004) UCSF Chimera – a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612.