J Mol Evol (1992) 34:351–357

Journal of
Molecular Evolution

Organization and Evolution of Bacterial and Bacteriophage Primase–Helicase Systems

Tatjana V. Ilyina,[1] Alexander E. Gorbalenya,[2] and Eugene V. Koonin[1]

[1] Institute of Microbiology, USSR Academy of Sciences, 7 Prospekt 60let Oktyabrya, 117811 Moscow, USSR
[2] Institute of Poliomyelitis and Viral Encephalitides, USSR Academy of Medical Sciences, 142792 Moscow Region, USSR

Summary. Amino acid sequences of primases and associated helicases involved in the DNA replication of eubacteria and bacteriophages T7, T3, T4, P4, and P22 were compared by computer-assisted methods. There are two types of such systems, the first one represented by distinct helicase and primase proteins (e.g., DnaB and DnaG proteins of *Escherichia coli*), and the second one by single polypeptides comprising both activities (gp4 of bacteriophages T7 and T3, and alpha protein of bacteriophage P4). Pronounced sequence similarity was revealed between approximately 250 amino acid residue N-terminal domains of stand-alone primases and the primase–helicase proteins of T7(T3) and P4. All these domains contain, close to their N-termini, a conserved Zn-finger pattern that may be implicated in template DNA recognition by the primases. In addition, they encompass five other conserved motifs some of which may be involved in substrate (NTP) binding. Significant similarity was also observed between the primase-associated helicases (DnaB, gp12 of P22 and gp41 of T4) and the C-terminal domain of T7(T3) gp4. On the other hand the C-terminal domain of P-alpha of P4 is related to another group of DNA and RNA helicases. Tentative phylogenetic trees generated for the primases and the associated helicases showed no grouping of the phage proteins, with the exception of the primase domains of bacteriophages T4 and P4. This may indicate a common origin for one-component primase–helicase systems. Two scenarios for the evolution of primase–helicase systems are discussed. The first one involves fusion of the primase and helicase components (T7 and T3) or fusion of the primase component with a different type of helicase domain (P4). The second possibility is the duplication of an ancestral gene encoding a gp4-like bifunctional protein followed by divergence of the copies, one of which retains the primase and the other the helicase domain.

Key words: Primase–helicase systems — Evolution — Bacteria — Bacteriophage

Introduction

Primase is an essential enzyme in all well–characterized systems of DNA replication. It promotes the synthesis of short oligoribonucleotides that serve as the primers for the Okazaki fragments on the lagging strand and in some systems also for the leading strand of the replicated DNA (Kornberg 1980, 1982; Van den Ende et al. 1985). As far as it is known, primases are always associated with DNA helicases. These helicases are thought not only to unwind the double-stranded DNA ahead of the replication fork but also to modify the conformation of the lagging strand in the initiation sites of Okazaki fragments and/or to interact with the primase protein itself so as to allow binding of the primase that functions in the distributive fashion (Alberts 1984; Nakai and Richardson 1988; Cha and Alberts 1990). There are

Present address and offprint requests to: E.V. Koonin at the National Center for Biotechnology Information, National Library of Medicine, Bldg. 38A, NIH 8600 Rockville Pike, Bethesda, MD 20894, USA

Table 1. Primase–helicase systems in bacteria and bacteriophages

	E. coli	Cryptic plasmid of *Chlamydia*	P22	T4	T7 T3	P4
Primase	DnaG (581)	Chlamydial DnaG homolog (?) (?)	DnaG (581)	gp61 (342)	gp4 (566)	gp alpha (777)
Helicase	DnaB (471)	gp1 (451)	gp12 (458)	gp41 (475)	gp4* (503)	gp alpha (777)
Type of primase–helicase	Two-component	Two-component	Two-component	Two-component	One-component	One-component
References	Van den Ende et al. 1985; Wong et al. 1988, and references therein	Hatt et al. 1988, and references therein	Wickner 1984a,b	Cha and Alberts 1990, and references therein	Bernstein and Richardson 1989; Beck et al. 1989, and references therein	Flensburg and Calendar 1987, and references therein

The numbers in parentheses specify the size of each protein. In the case of T7 gp4, the lengths of the complete 63-kd protein and the N-terminally truncated 56-kd protein are indicated for the primase and the helicase, respectively

two types of primase–helicase systems: two-component systems consisting of distinct primase and helicase polypeptides and one-component systems, in which both activities are exerted by a single gene product (Table 1). The sum of the sizes of the polypeptides constituting the two-component primase–helicase systems is approximately twice the size of the one-component primase–helicase (gp4) of T7 and T3 bacteriophages (Table 1). On the other hand, the latter gene is expressed in two forms of protein, the complete one (63 kd) and the short one (56 kd), which is produced by internal translation initiation and lacks the N-terminal 63 amino acid residues (Scherzinger et al. 1977; Dunn and Studier 1983; Bernstein and Richardson 1988b). The 63-kd form of gp4 is both primase and helicase, whereas the 56-kd form possesses only the helicase activity; the two forms function as a complex in phage DNA replication (Bernstein and Richardson 1988a, 1989).

We were interested in comparing the amino acid sequences of the primases and the associated helicases to learn whether a direct functional and evolutionary relationship exists between the two types of primase–helicase systems. We show here that the primases and the associated helicases indeed constitute distinct protein families and discuss two alternative scenarios for the evolution of primase–helicase systems.

Methods

Computer-Assisted Analysis of Amino Acid Sequences. Amino acid sequences were from SWISSPROT data bank (Release 16). Initial sequence comparisons were done using the program DOTHELIX, generating complete local similarity plots for pairs of sequences (Leontovich et al. 1990). Additionally, putative conserved motifs were identified by program SITE, which converts alignments to position-dependent weight matrices and scans query sequences for segments scoring highest with these matrices (Koonin et al. 1990). Putative conserved segments found by these programs were used to delineate the boundaries of the regions of the proteins to be aligned by the multiple alignment program OPTAL using the amino acid residue comparison matrix MDM78 as previously described (Gorbalenya et al. 1989a). Briefly, this program implementing the Sankoff algorithm generates multiple sequence alignments in a stepwise manner and calculates adjusted alignment scores as the number of standard deviations (SD) over the mean of 25 random simulations. Protein secondary structure predictions were performed using the program PROTEIN2 implementing the Garnier algorithm. Programs DOTHELIX and PROTEIN2 are modules of the GENEBEE program package for biopolymer sequence analysis (Brodsky et al. 1991).

Phylogenetic Trees. Three methods for tentative phylogenetic tree generation were used: (1) The first is a simple minimal distance clustering algorithm, the unweighted pair-group method using arithmetic averages (UPGMA) (Sneath and Sokal 1973). (2) A protein parsimony algorithm, which is a version of the maximum parsimony method, is used to calculate the probabilities of amino acid changes based on genetic code assignments. The algorithm is implemented in the PROTPARS program of the PHYLIP package kindly supplied by Dr. J. Felsenstein (Felsenstein 1989). (3) Finally there is a maximum topological similarity algorithm, one of the "quartet" algorithms employing comparison of the sets of nearest neighbor quartets from a distance matrix and a tree and minimization of the number of different quartets (Chumakov and Yushmanov 1988; Yushmanov and Chumakov 1988). This algorithm was implemented in the TREE program of the GENEBEE package (Brodsky et al. 1991). The pairwise distances between the sequences were computed using a modification of the formula of Feng et al. (1985).

Results and Discussion

Bacterial and Bacteriophage primases and the Associated Helicases Constitute Two Compact Families of Related Proteins

Primases

Previously, Toh (1986) described a short region of similarity between *Escherichia coli* DnaG protein and gp4 of bacteriophage T7. On the other hand, it has been claimed that the sequence of bacteriophage P4 alpha protein is unrelated to that of gp4 (Flensburg and Calendar 1987). We sought to clarify the relationships between bacterial and phage primases by first detecting similar segments on local similarity plots, and then by generating multiple alignments. Statistically significant alignments with scores of at least 7.5 SD were obtained for the approximately 250 amino acid residue N-terminal domains of three bacterial primases and those of phages T4, T7, and P4.

A striking feature of the primase alignment (Fig. 1) is the conservation, close to the N-termini of the aligned proteins, of two pairs of Cys(His) residues in a configuration potentially allowing formation of a Zn-finger. That the finger-containing fragment is essential for primase activity is demonstrated by the fact that the N-terminally truncated 56-kd form of T7 gp4 is devoid of this activity (Bernstein and Richardson 1988a). It has been proposed by these authors that the finger may be involved in the recognition of the priming sites on the template DNA by the primase of T7. Our observations show that this may be a common property of bacterial and phage primases.

In addition to the putative finger motif, the primase alignment encompassed five other conserved sequence motifs (Fig. 1). Motifs 4 and 5 containing conserved negatively charged residues preceded by runs of hydrophobic residues predicted to form β-strands (data not shown) resemble the motifs involved in Mg^{2+}-mediated NTP binding by ATPases (Fry et al. 1986; Gorbalenya and Koonin 1989) and probably by various RNA and DNA polymerases (Poch et al. 1989; Bernad et al. 1990). It is tempting to speculate that one of these motifs might be directly involved in substrate binding by the primases. On the other hand, it is important to note that, despite extensive search, no statistically significant similarity between the primases and any other RNA or DNA polymerases (i.e., DNA-dependent RNA polymerases, DNA-dependent DNA polymerases, reverse transcriptases, and viral RNA-dependent RNA polymerases) could be detected.

Helicases

It has been shown previously that the C-terminal domains of bacterial DnaB helicases, gp12 of bacteriophage P22, and gp4 of T7 align well with each other (Backhous and Petri 1984; Wong et al. 1988). The functional equivalence of the helicases encoded by the *E. coli* dnaB gene and gene 12 of P22 has been demonstrated experimentally (Wickner 1984b). We show here that the gp41 helicase of T4 belongs to the same family of related proteins (Fig. 2). For

```
                                    1
                          *  *                   *  *
                     :    *  *             :   : *: *
DnaG  E.coli  36 F | HACCPFHNEKTPS-------FTVNGEKQFYHCFGCG | AHGNAIDFLMNYD
DnaG  S.typ.  36 Y | HACCPFHNEKTPS-------FTVNGEKQFYHCFGCG | AHGNAIDFLMNYD
DnaE  B.sub.  36 Y | FGLCPFHGESTPS-------FSVSPDKQIFHCFGCG | AGGNVFSFLRQME
g61   T4      34 F | R--CPVCGDSKTDQNKARGWYYGDNNEGNIHCYNCN | YHAPIGIYLKEFE
alpha P4      31 - | HQPCPVCGGSDRF------RFDDREGRGTWYCNQCG | A-GDGLKLVEKVF
gp4   T3      11 F | HAPCENCGSSDGN--------SVYSDGHEW-CFVCE | HRVPANE------
gp4   T7      13 Y | HIPCDNCGSSDGN--------SLFSDGHTF-CYVCE | KWTAGNE------

                 :    :                              :
DnaG  E.coli   KLEFVETVEELAAMHNLEVPFE-AGSGPSQIERHQRQTLYQLMDGLNTFYQQSL
DnaG  S.tyh.   KLEFVETVEELAAMHNLEIPYE-AGTGLSQIERHQRQNLYQLMNGLNDFYQQSL
DnaE  B.subt.  GYSFAESVSHLADKYQIDFPDDITVHSGARPESSGEQKMAEAHELLKKFYHHLL
g61   T4       PDLYREYIFEIRKEKGKSRPIE----KPKELPKQPEKKIIKSLPSCVRL--DKL
alpha P4       GVSPSDAAAKVAAVTGSLPPAD------PAVTTAAVDETDAARKNAAALAQTLM
gp4   T3       -----EREAKLSTRRR------------TGGSKPMSYDVWNFGDSNGRY-----
gp4   T7       -----DTKER-ASKRK------------PSGGKPMTYNVWNFGESNGRY-----
                                                  ↑
                              2
                         *  **   :          :    *    :
DnaG  E.coli   QQP-VATSARQ | YLEKRGLSHEVI | ARFAIGFAPPGWDNVLKRFGGNPENRQSL
DnaG  S.tyh.   THP-AAKPARD | YLQKRGLSAEII | QRFAIGFAPPGWDNALKRFGNNSDNKALL
DnaE  B.subt.  INTKEGQEALD | YLLSRGFTKELI | NEFQIGYALDSWDFITKFLVKRGFSEAQM
g61   T4       AE---DHPIIK | YVKARCIPKD-K | WKY-LWFT-TEWPKLVNSIAPGTYKKE--
alpha P4       AKT-RTGTGNA | YLTRKGFPGREC | RMLTGTH----RAGGVSWRA---------
gp4   T3       ----------S | DLTARGISKETC | QK--AGY----WLAKV------------
gp4   T7       ----------S | ALTARGISKETC | QK--AGY----WIAKV------------

                                           3
                                       ::  : *                  :      *
DnaG  E.coli   IDAGMLVTNDQGRSY-DRFRE | RVMFPIRDKRGRVIGF | GGRVLGNDTP-KYLN
DnaG  S.tyh.   LDAGMLVNNEQGSTY-DRFRN | RVMFPIRDKRGRVIGF | GGRVLGNDTP-KYLN
DnaE  B.subt.  EKAGLLIRREDGSGYFDRFRN | RVMFPIHDHHGAVVAF | SGRALGSQQP-KYMN
g61   T4       ------ISE-----------P | RLVIPIYNANGKAESF | QGRALKKDAPQKYI-
alpha P4       --------------------G | DLVVPLYDDSGELVNL | --QLISADGRKRTLK
gp4   T3       --------------------D | NRMYQVADYRDQNGSI | VSQKVRDKD--KNFK
gp4   T7       --------------------D | GVMYQVADYRDQNGNI | VSQKVRDKD--KNFK

                                                  4
                :  :          *:                  :: :  **  * : :          :
DnaG  E.coli   SPETDIFHKGRQLYGLYEAQQDNAEPN | RLLVVEGYMDVVAL | AQY-GINYAVA
DnaG  S.tyh.   SPETDIFHKGRQLYGLYEAQQYSAEPQ | RLLVVEGYMDVVAL | AQY-DINYAVA
DnaE  B.subt.  SPETPLFHKSKLLYNFYKARLHIRKQE | RAVLFEGFADVYTA | VSS-DVKESIA
g61   T4       TIEA--YPEATKIYGV----ERVKDGD | -VYVLEGPIDSL-- | --F--IENGIA
alha  P4       GGQVR--GTCHTLEG------QNQAGK | RLWIAEGYATALTV | HHL-TGETVMV
gp4   T3       TTGS---HKSDALFL----KHLWSGGK | KIVVTEGEIDMLTV | MELQDCKYPVV
gp4   T7       TTGS---HKSDALFG----KHLWNGGK | KIVVTEGEIDALTV | MELQDCKYPVV

                                                  5
                :*                        :    :::    *  *  *          :
DnaG  E.coli   SLGTSTTADHIQ-----LLFRATN- | NVICCYDGDRAGRD | AAWRALETALPYM
DnaG  S.tyh.   SLGTSTTADHMH-----MLFRATN- | NVICCYDGDRAGRD | AAWRALETAMPYM
DnaE  B.subt.  TMGTSLTDDHVK-----ILRRNVE- | EIILCYDSDKAGYE | A---TLKASELLQ
g61   P4       ITGGQLDLEVVP-------FK--D- | RRVWVLDNEPRHPD | ----TIKRMTKLV
alpha P4       ALSSVNLLSLAS----LARQKHPAC | QIVLAADRDLSG-D | -----GQKKAAAA
gp4   T3       SLGHGASAAKKTCAANYEYFDQFE- | QIILMFDMDEAGRK | ----AVEEAAQVL
gp4   T7       SLGHGASAAKKTCAANYEYFDQFE- | QIILMFDMDDAGRK | ----AVEEAAQVL

                                  6
                  *   :        *     * :      *
DnaG  E.coli   TDGRQLRF- | MFLP-DGEDPDTLVRKEG |   227
DnaG  S.tyh.   TDGRQVRF- | MFLP-DGEDPDTLVRKEG |   227
DnaE  B.subt.  KKGCKVRV- | AMIP-DGLDPDDYIKKFG |   249
gp61  T4       DAGERVMFW | DKSPWKSKDVNDMIRKEG |    30
alpha P4       ADACEGVV- | ALPP-CFGDWNDAFTQYG |   478
gp4   T3       PAG-KVRV- | AVLP-C-KDANEC-HLNG |   244
gp4   T7       PAG-KVRV- | AVLP-C-KDANEC-HIMG |   244
```

Fig. 1. Alignment of the amino acid sequences of bacterial and bacteriophage primase domains. Asterisks: identical amino acid residues conserved in at least six out of seven sequences; colons: structurally similar residues conserved in at least six out of seven sequences; double asterisks: Cys and His residues probably involved in Zn-finger formation; arrow: the start Met of the N-terminally truncated (56-kd) form of gp4 of T7. The six distinct conserved motifs are shown. The distances from the aligned regions to the protein termini are indicated by numbers. Abbreviations: B.subt., *Bacillus subtilis*; S.typh., *Salmonella typhimurium*.

```
                                                                  1 ("A")
                    *:                 :  *   ::  :  *     :::::     :**:
dnaB E.coli   191  ARIEQ---LFQQPHDG-VTGVNTGYDDLNKKT-AGLQPS DLIIVAARPSMGKT
dnaB S.typh.  191  ARIEQ---LFQQPHDG-VTGVDTGYQDLNKKT-AGLQRS DLIIVAARPSMGKT
gp1 C.trac.   180  ARIKKRRATFSENQNSFFDAFPTGYKDIDDKG-VILAKG NFVIIAARPSIGKT
gp12 P22      155  DTLDR---RLRNGEES--DTLKTGIEELDAIT-GGMNAE DLVIIAARPGMGKT
gp41 T4       159  AR------WLSYMNKA--RKVPFKLRILNKITKGGAETG TLNVLMAGVNVGKS
gp4  T3       272  DRVKE---AMTSEDAV--GLLFDGCTGLNDRT-LGARGG EVVMVTSGSGMGKS
gp4  T7       272  ERIRE---HLSSEESV--GLLFSGCTGINDKT-LGARGG EVIMVTSGSGMGKS


                                           1a
                                    :   *    :*           ::    :
dnaB E.coli    TFAM-NLVENAA--MLQD---KP- VLIFSLEM PSEQIMMRSLASLSRVDQTKIR-T
dnaB S.typh.   TFAM-NLCENAA--MLQD---KP- VLIFSLEM PGEQIMMRMLASLSRVDQTRIR-T
gp1  C.trac.   ALAI-DMAINLA--VTQQ---RR- VGFLSLEM SAGQIVERIIANLTGISGEKLQ-R
gp12 P22       ELAL-KIAEGVASRVIPGSGVRRG VLIFSMEM SAIQVVERGIAGAGMMSVSVLRNP
gp41 T4        -LGLCSLAADY---LQLG---HN- VLYISMEM AEEVCAKRIDANMLDVSLDDID-D
gp4  T3        TFVR-QQALAWG--KRMG---KR- VGLAMLEE SVEDTIQDMMGLNNKVRLRQSDEV
gp4  T7        TFVR-QQALQWG--TAMG---KK- VGLAMLEE $VEETAEDLIGLHNRVRLRQSDSL


                                 *   ::       :         : *
dnaB    E.coli     GQLDDEDW-ARISG-TMGILLEKRNIYIDDSSGLTPTEVRSRARRIAREHGGI-
dnaB S.typh.       GQLDDEDW-ARISG-TMGILLEKRNMYIDDSSGLTPTEVRSRARRIFREHGGL-
gp1  C.trac.       GDLSKEEL-FRVEE-A-GETVRESHFYICSDSQYK-LNLIANQIRLLRKEDRV-
gp12 P22           SRMDDEGW-ARVAS-GM-KLLAELDVWVVDASRLSVEEIRSISERHKQEHPNL-
gp41 T4            GHISYAEYKGKMEKWREKSTLGRLIVKQYPTGGADANTFRSLLNELKLKKNFVP
gp4  T4            KKAIAED--GRFDE-WYDELFGDDTFHLYDSFAEAEAD-RLLAKLAYMRTGLGC
gp4  T7            KREIIEN--GKFDQ-WFDELFGNDTFHLYDSFAEAETD-RLLAKLAYMRSGLGC


                      2("B")                               3
                     :*::* :  :   :      ::     *  :   *:: *     :: :
dnaB E.coli        G LIMIDYL QL---MRVPALSDNRTLE IAEISRSLKALAKELVPVVALSQL -
dnaB S.typh.       S LIMIDYL QL---MRVPSLSDNRTLE IAEISRSLKALAKELVPVVALSQL -
gp1  C.trac.       D VIFIDYL QL---IN-SSVGENRQNE IADISRTLRGLASELIPIVCLSQL -
gp12 P22           S LIMADYL GL---IEKPK-AERNDLA IAHISGSLKAMAKDLTPVISLSQL -
gp41 T4            T IIIVDYL GICKSCRIRVYSENSYTT VKAIAEELRALAVETTVLWTAAQV -
gp4  T3            D VIVLDHI SI--VVSASEESDERKM- IDRLMTKLKGFAKSTGVVLVVICH L
gp4  T7            D VIILDHI SI--VVSASGESDERKM- IDNLMTKLKGFAKSTGVVLVVICH L


                                           4
                             :       :*** ** : *  * ::   :  :     : :
dnaB E.coli        --NRSLEQRADKR- PVNSDLRESGSIEQDADLIM -FIYRDEVYHENSDLKGIA
dnaB S.typh.       --NRSLEQRADKR- PVNSDLRESGSIEQDADLIM -FIYRDEVYHENSDLKGIA
gp1  C.trac.       --SRKVEDRANKV- PMLSDLRDSGQIEQDADVIL -FINRKE---SSSN----C
gp12 P22           --SRDVEKRPNKR- PTNADLRDSGSIEQDADSII -MLYREAVYDENSSAAPFA
gp41 T4            --GK--QAWDSSD- VNMSDIAESAGLPATADFML AVIETEELAAAEQQLIKQI
gp4  T3            KNPEKGKAHEEGRP VSITDLRGSGALRQLSDTII ALERNQ-----QGDMPNLV
gp4  T7            KNPDKGKAHEEGRA VSITDLRGSGALRQLSDTII ALERNQ-----QGDMPNLV


                    : : * *
dnaB E.coli        EIIIGKQRN  28
dnaB S.typh.       EIIIGKQRN  28
gp1  C.trac.       EITVGKNRH  24
gp12 P22           EIIVTKNRF  46
gp41 T4            KSRYGDKNK  62
gp4  T3            LVRLLKCRF  44
gp4  T7            LVRILKCRF  43
```

Fig. 2. Alignment of the amino acid sequences of bacterial and bacteriophage primase-associated helicase domains. The designations are as in Fig. 1. The five conserved motifs are shown. The designation 1a is used to follow the nomenclature of the conserved motifs in other helicases (Gorbalenya et al. 1989b). A and B are the two motifs constituting the purine NTP-binding pattern. Gp1 C.trac. is gene 1 product of the cryptic plasmid of *Chlamydia trachomatis* (Hatt et al. 1988).

each step of the alignment, highly significant scores were obtained, but it should be noted that the P22 protein is much more closely related to the bacterial helicases than the proteins of T7 and T4. Specifically, the score of 29.5 SD was observed for the alignment of gp12 with the bacterial helicases as opposed to 20.8 SD and 7.7 SD for gp4 of T7 and gp41, respectively. The alignment of these DnaB-related helicases encompassed five distinct conserved sequence segments (Fig. 2). Segments 1 and 2 correspond to the A and B motifs of the purine NTP-binding site, respectively; these motifs have been directly implicated in NTP binding and hydrolysis (Walker et al. 1982; Gorbalenya and Koonin 1989). The function(s) of the other conserved segments remain unknown, but it is tempting to speculate that they may relate to the helicase activity. Three superfamilies of DNA and RNA helicases characterized previously also have highly conserved sequence segments, in addition to those constituting the NTP-binding site (Gorbalenya et al. 1989b, 1990). However, the DnaB-related helicases showed no significant sequence similarity to any of the other helicases, and the patterns of amino acid residue conservation in the motifs typical of each of the three superfamilies are quite distinct from those described here for the DnaB family.

In a previous study (Gorbalenya et al. 1990), we showed that the C-terminal half of bacteriophage P4 alpha protein encompasses a domain related to the (putative) helicase domains encoded by small DNA viruses (papova-, parvo-, and geminiviruses) and positive-strand RNA viruses (picorna-, como-, and nepoviruses). Thus term, T antigen-related helicases, may be coined for this family, after the large T antigen of SV40, a well-characterized DNA and RNA helicase.

Evolutionary Relationships among Bacterial and Bacteriophage Primase–Helicase Systems

The present study demonstrates that direct evolutionary relationships apparently exist between bacterial and bacteriophage primases and associated helicases comprising both one-component and two-component primase–helicase systems (Fig. 3). To gain further insight into the possible pathways of evolution of these systems, we generated tentative phylogenetic trees separately for the primase and helicase domains. Figure 4 shows such trees produced by the clustering algorithm. Although cluster dendrograms are subject to artifacts caused by unequal evolutionary rates in different lineages, the (relatively) rate-independent parsimony and maximal topological similarity algorithms produced rootless trees with topologies fully compatible with those presented in Fig. 4. Three aspects of the resulting trees seem remarkable: (1) Both the primase and the helicase domains of different bacteriophages generally do not tend to group together, suggesting

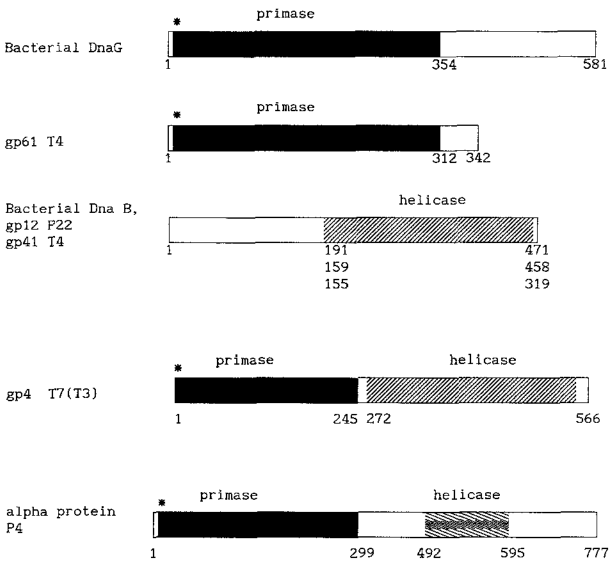

Fig. 3. A scheme showing the organization of primase and helicase domains in bacterial and bacteriophage proteins. Hatching highlights the conserved primase, DnaB-related helicase (DnaB, gp12, gp41, gp4), and T antigen-related helicase. These domains are designated by numbers. The data on the conserved helicase domain in the alpha protein were from Gorbalenya et al. (1990). Asterisks show the location of the putative Zn-fingers.

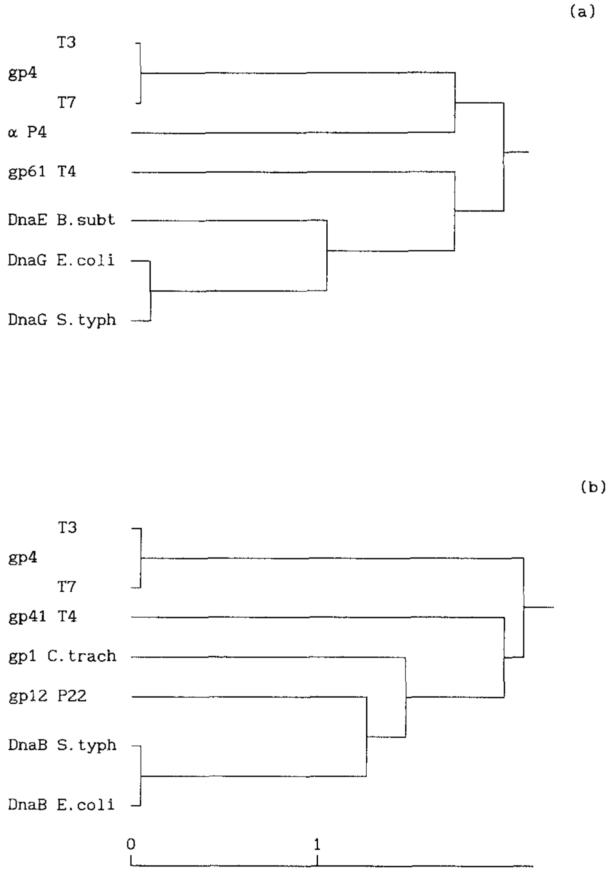

Fig. 4. Cluster dendrograms for bacterial and bacteriophage primase and helicase domains. The dendrograms were derived from the alignments shown in Figs. 1 and 2 using the UPGMA algorithm as indicated in the Methods. The scale of distances (in arbitrary units) is shown at the bottom. **a** Primase domains. **b** Helicase domains.

an independent origin for the bacteriophage primase–helicase systems. (2) The only exception is presented by the primases of bacteriophages T7(T3) and P4 (Fig. 4a), indicating that the primase moieties of the one-component primase–helicase systems may have a common origin. (3) The primase-associated helicase (gp12) of bacteriophage P22 falls within the bacterial domain of the tree (Fig. 4b). This is very compatible with the fact that this phage does not encode a primase of its own but utilizes the bacterial primase DnaG for its DNA replication (Wickner 1984a,b). Notably, the related bacteriophage λ encodes neither primase nor a helicase and completely relies on the respective host proteins to replicate its DNA (McMacken et al. 1987). Thus, acquisition of the helicase gene by P22 may be a relatively recent evolutionary event.

Basically, two types of evolutionary scenarios for the primase–helicase systems are imaginable. The first one assumes that the primase and helicase genes are of independent origin and the ancestral bacterial primase–helicase system consisted of two distinct proteins, which is similar to the extant systems. Under this assumption, the strategies adopted by the four bacteriophages for the evolution of their primase–helicase systems may be described as follows: (1) Phage P22 has captured only the helicase gene that functions in conjunction with the bacterial primase. (2) Phage T4 has captured the separate primase (gp61) and helicase (gp41) genes from the bacterium. (3) The capture of the primase and helicase genes by the ancestor of T7 and T3 phages involved fusion of their parts or fusion of complete genes followed by deletions of their large segments. (4) Finally, evolution of the P4 primase–helicase system included fusion of the primase gene with a gene encoding a helicase unrelated to DnaB. The bacterial homolog of this helicase is not (yet) known.

The alternative scenario relies on the assumption that the ancestral primase–helicase system consisted of only one component and resembled gp4 of phages T7 and T3. Clearly, the gene for this ancestral primase–helicase could exist only at an early stage of bacterial evolution, as such distant bacteria as representatives of gram-positive (*Bacillus subtilis*) and gram-negative (*E. coli* and *Salmonella typhimurium*) groups already have distinct primase and helicase genes. The evolution of this gene could have involved a duplication followed by divergence of the copies, with the N-terminal primase domain retained by one of the copies and the C-terminal helicase domain by the other. The additional domains emerging as the result of this process could provide some accessory functions. e.g., interactions with the other protein components of DNA replication machines. Under this scenario, T7(T3) and probably P4 (or their common progenitor) derived the primase–helicase gene from the bacterial genome before the duplication. In P4, the DnaB-like domain could be subsequently substituted by a different type of helicase. T4 could acquire the primase and helicase genes later in evolution, after the proposed duplication.

The gene fusions that could have occurred in the evolution of the primase–helicase systems of T7(T3) and P4 or conservation of the ancestral one-component organization of primase–helicase suggested by the second scenario are compatible with the general trend of the economy of the genome coding capacity typical of (relatively) small viruses. However, it has been shown that T7 still utilizes two different forms of gp4 as the helicase (56-kd protein) and the primase (63-kd protein). Apparently, the 56-kd form provides the processive DNA-unwinding activity at the leading edge of the replication fork, whereas the 63-kd protein mediates the distributive synthesis of the primers on the lagging strand (Bernstein and Richardson 1989). It will be most interesting to find out if P4 utilizes a similar strategy to generate two functionally distinct forms of primase–helicase.

References

Alberts BM (1984) The DNA enzymology of protein machines. Cold Spring Harbor Symp Quant Biol 48:1–12

Backhaus H, Petri JB (1984) Sequence analysis of a region from the early right operon in phage P22 including the replication genes 18 and 12. Gene 32:289–303

Beck PJ, Gonzalez S, Ward CL, Molineaux IJ (1989) Sequence of bacteriophage T3 DNA from gene 2.5 through gene 9. J Mol Biol 210:687–701

Bernad A, Lazaro JM, Salas M, Blanco L (1990) The highly conserved amino acid sequence motif Tyr–Gly–Asp–Thr–Asp–Ser in alpha-like DNA polymerases is required by phage ϕ29 DNA polymerase for protein-primed initiation and polymerization. Proc Natl Acad Sci USA 87:4610–4614

Bernstein JA, Richardson CC (1988a) A 7-kDa region of the bacteriophage T7 gene 4 protein is required for primase but not for helicase activity. Proc Natl Acad Sci USA 85:396–400

Bernstein JA, Richardson CC (1988b) Purification of the 56-kDa component of the bacteriophage T7 primase/helicase and characterization of its nucleoside 5′-triphosphate activity. J Biol Chem 263:14891–14899

Bernstein JA, Richardson CC (1989) Characterization of the helicase and primase activities of the 63-kDA component of the bacteriophage T7 gene 4 protein. J Biol Chem 264:13066–13073

Brodsky LI, Drachev AL, Tatuzov RL, Chumakov KM (1991) GENEBEE: a package of computer programs for biopolymer sequence analysis. Biopolim Kletka 7 (1):10–14

Cha T-A, Alberts BM (1990) Effects of the bacteriophage T4 gene 41 and gene 32 proteins on RNA primer synthesis: coupling of leading- and lagging-strand synthesis at a replication fork. Biochemistry 29:1791–1798

Chumakov KM, Yushmanov SY (1988) Maximum topological similarity principle in molecular taxonomy. Mol Genet Mikrobiol Virusol 3:3–9 [in Russian]

Dunn JJ, Studier FW (1983) Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. J Mol Biol 166:477–535

Felsenstein J (1989) PHYLIP 3.2 manual. University of California Herbarium, Berkeley CA

Feng DF, Johnson MS, Doolittle RF (1985) Aligning amino acid sequences: comparison of commonly used methods. J Mol Evol 21:112–125

Flensburg J, Calendar R (1987) Bacteriophage P4 DNA replication. Nucleotide sequence of the P4 replication gene and the cis replication region. J Mol Biol 196:439–445

Fry DC, Kuby SA, Mildvan AS (1986) ATP-binding site of adenylate kinase: mechanistic implications of its homology with ras-encoded p21, F-ATPase, and other nucleotide-binding proteins. Proc Natl Acad Sci USA 83:907–911

Gorbalenya AE, Koonin EV (1989) Virus proteins containing the purine NTP-binding pattern. Nucleic Acids Res 17:8413–8440

Gorbalenya AE, Blinov VM, Donchenko AP, Koonin EV (1989a) An NTP-binding motif is the most conserved sequence in a highly diverged group of proteins involved in positive strand RNA viral replication. J Mol Evol 28:256–268

Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM (1989b) Two related superfamilies of putative helicases involved in replication, recombination, repair and expression of DNA and RNA genomes. Nucleic Acids Res 17:4713–4730

Gorbalenya AE, Koonin EV, Wolf YI (1990) A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. FEBS Lett 00:145–148

Hatt C, Ward ME, Clarke IN (1988) Analysis of the entire nucleotide sequence of the cryptic plasmid of *Chlamydia trachomatis* serovar L1. Evidence for involvement in DNA replication. Nucleic Acids Res 16:4053–4067

Koonin EV, Chumakov KM, Gorbalenya AE (1990) A method for localization of motifs in amino acid sequences. Biopolym Kletka 6(6):43–48 [in Russian]

Kornberg A (1980) DNA replication. Freeman, San Francisco

Kornberg A (1982) Supplement to DNA replication. Freeman, San Francisco

Leontovich AM, Brodsky LI, Gorbalenya AE (1990) Compilation of a complete map of local similarity for two biopolymers (DOTHELIX program of the GENBEE package. Biopolim Kletka 6(6):14–21

McMacken R, Alfano C, Gomes B, LeBowitz JH, Mensa-Wilmot K, Roberts JD, Wold M (1987) Biochemical mechanisms in the initiation of bacteriophage λ DNA replication. In: Kelly TJ, McMacken R (eds) DNA replication and recombination. Alan R. Liss, New York, pp 227–245

Nakai H, Richardson CC (1988) Leading and lagging strand synthesis at the replication fork of bacteriophage T7. J Biol Chem 263:9818–9830

Poch O, Sauvaget I, Delarue M, Tordo N (1989) Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. EMBO J 8:3867–3874

Scherzinger E, Lanka E, Morelli G, Seiffert D, Yuki A (1977) Bacteriophage T7 induced DNA-priming protein. Eur J Biochem 72:543–558

Sneath P, Sokal R (1973) Principles of numerical taxonomy. San Francisco

Toh H (1986) T7 and *E. coli* share homology for replication-related gene products. FEBS Lett 194:245–248

Van den Ende A, Baker TA, Ogawa T, Kornberg A (1985) Initiation of enzymatic replication at the origin of *Escherichia* *coli* chromosome: primase as the sole priming enzyme. Proc Natl Acad Sci USA 82:3954–3958

Walker JE, Saraste M, Runswick MJ, Gay NJ (1982) Distantly related sequences in the a- and b-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J 2:945–951

Wickner S (1984a) DNA-dependent ATPase activity associated with phage P22 gene 12 protein. J Biol Chem 259:14038–14043

Wickner S (1984b) Oligonucleotide synthesis by *Escherichia coli* primase in conjunction with phage P22 gene 12 protein. J Biol Chem 259:14044–14047

Wong A, Kean L, Maurer R (1988) Sequence of the dnaB gene of *Salmonella typhimurium.* J Bacteriol 170:2668–2675

Yushmanov SY, Chumakov KM (1988) Algorithms for construction of maximum topological similarity phylogenetic trees. Molek Genet Mikrobiol Virusol 3:3–9 [in Russian]

Received July 24, 1991/Revised and accepted October 15, 1991