DNA replication at high resolution
James L Keck and James M Berger

Several decades of research have delineated the roles of many proteins central to DNA replication. Here we present a structural perspective of this work spanning the past 15 years and highlight several recent advances in the field.

Address: Department of Molecular and Cell Biology, University of California, Berkeley, 229 Stanley Hall #3206, Berkeley, CA 94720, USA.

Correspondence: James M Berger
E-mail: jmberger@uclink4.berkeley.edu

Chemistry & Biology 2000, 7:R63–R71
1074-5521/00/$ – see front matter © 2000 Elsevier Science Ltd. All rights reserved.

Introduction
Replication, the duplication of an organism’s genetic material, is a fundamental process in biology. From viruses to multicellular organisms, life is inseparable from the genetic instructions that maintain and regulate its existence. Strong evolutionary pressures have, therefore, forced cells and viruses to ensure that the timing of replication is highly coordinated and that the process itself proceeds with as few errors as possible. Over the years, researchers have extensively examined these events, transforming replication into one of the most well-studied fields in biology.

Initially, replication research focused on prokaryotes and bacteriophage, using genetic methods to broadly delineate and define the roles of numerous proteins central to the process. These experiments demonstrated that some replication proteins act only in initiation, whereas others are required throughout the reaction cycle [1]. As methods in molecular biology were invented and improved, researchers sharpened this view by studying individual replication proteins and protein complexes in vivo. The field has further expanded to include the complex reactions of eukaryotic replication. Together, these studies have helped develop working models for the catalytic properties of several replication proteins and have demonstrated that many of the components act cooperatively as multiprotein complexes.

Over the past 15 years, structural studies have begun to resolve this picture further, delivering high-resolution, three-dimensional images of many of the individual proteins central to replication. These models have allowed researchers to interpret the abundant genetic and biochemical data generated from decades of replication research, as well as to formulate novel hypotheses about the control and execution of the process. Breakthroughs continue today in all areas of replication study and have been the subjects of many recent reviews [2–7]. Here we will focus primarily on several advances in our understanding of prokaryotic and bacteriophage DNA replication that have been made using high-resolution structural biology.

Most of the reactions of bacterial and phage replication have functional analogs in eukaryotic biology and are therefore meaningful across all kingdoms of life.

Bacterial DNA replication: a multiprotein event
DNA replication is a multi-staged reaction (Figure 1). First, a replication origin sequence within the genome is recognized by origin-binding protein components. Second, initiation of replication occurs through the recruitment of
replication proteins and the assembly of replication forks at the origin. Third, the general replication reaction duplicates both strands of DNA. Fourth, replication terminates and the two daughter chromosomes are separated. We will describe these processes as they occur in the bacterium *Escherichia coli* before delving into the structures of known replication proteins.

Replication initiation in *E. coli* begins with the recognition of a short origin DNA sequence ('ori') by an initiator protein called DnaA (Figure 1a). DnaA binding alters the structure of the 'ori' DNA and provides an associative platform for recruiting additional replication proteins [8,9]. The origin is then actively assembled into a bidirectional replication bubble with two oppositely directed replication forks. This process involves the recruitment of one helicase (DnaB) per replication fork by association with DnaA [10,11]. DnaB attracts a second important replication-fork enzyme called 'primase' (DnaG), which is a specialized, replication-priming RNA polymerase. Together, the helicase and primase form the core of an assembly referred to as the 'primosome' [12]. Because DNA polymerases are incapable of initiating DNA chains de novo, the primase portion of the primosome makes a short RNA chain, called a primer, from which DNA polymerase can extend its first few nucleotides. During these initial stages, and throughout the entire course of replication, the DnaB helicase catalyzes the unwinding of genomic DNA in an ATP-dependent reaction [13].

Once the DNA is melted and primed, the general DNA replication reaction begins (Figure 1b). The central enzyme in this stage of replication is the DNA polymerase, which synthesizes a single DNA strand by covalently linking together appropriate deoxynucleotidetriphosphates (dNTPs) that are complementary to the template. Because all nucleic-acid polymerases add nucleoside triphosphates (NTPs) only to 3'-hydroxyl moieties on growing nucleic-acid chains, there is a strict directionality to DNA polymerase activity in the 5'→3' direction.
This restraint forces replication of the antiparallel DNA strands to be asymmetric: one newly synthesized strand is made continuously (leading strand), while the other is synthesized in segments (lagging strand) with new DNA polymerase initiating every ~1000 nucleotides [2]. In both cases, DNA polymerase III (pol III) is responsible for extending the nucleic-acid chains in E. coli [14,15]. Replication fork asymmetry also makes DNA polymerase initiation by primase obligatory throughout replication to ensure full duplication of the lagging strand. In order to maintain association between pol III and DNA, specialized proteins known as processivity factors act as ‘sliding clamps’ on duplex DNA. These proteins are loaded onto primed DNA by ‘clamp-loader’ proteins [16,17] and then associate with the polymerase, providing an additional tether between the polymerase and DNA. The activities of the leading- and lagging-strand polymerases in a replication fork are coordinated by a protein infrastructure forming a large complex referred to as the polymerase holoenzyme (pol-HE) [18]. As a replication fork progresses, single-stranded DNA (ssDNA) intermediates that are generated by the helicase are bound by ssDNA-binding proteins (SSBs) in a restricted conformation that ensures replication can proceed swiftly [2].

After copying the host chromosome, replication must terminate (Figure 1c). This process occurs in an organized manner, although certain details regarding this stage of replication are not as well understood as are initiation and elongation. In the E. coli chromosome, termination proteins bind specific DNA sequences and stop replication by inhibiting helicase translocation [19,20]. Following termination, the two completed copies of the genome are unlinked by topoisomerases to allow independent segregation of the two daughter chromosomes. At this end point, the cell has catalyzed the replication of 4.6 x 10^6 basepairs of genomic DNA in 40 minutes, at the astonishing rate of 1000 nucleotides per second per replication fork [3].

**Structures in DNA replication**

Over the past 15 years, X-ray crystallographic and nuclear magnetic resonance (NMR) studies have resulted in the high-resolution structure determination of at least one example of nearly every protein involved in DNA replication. In Figure 1, replication proteins and complexes are colored to highlight known high-resolution structures; all proteins with color have at least one representative structure that defines the three-dimensional shape of the molecule. The overwhelming number of proteins for which structures have been solved indicates that DNA replication is fast becoming one of the best understood processes in biology at the molecular level. To provide an overview of the molecular details of replication, we describe examples of known replication protein structures, with a particular emphasis on recent examples from bacteria and phage.

**Origin-binding proteins**

Although no structures have been solved for cellular origin-binding proteins, two such structures from viruses [21–23] and one from a plasmid replication system [24] have been described. The structures of these proteins are relatively distinct. All are seen to slightly bend DNA upon binding, but the functional significance of this feature is unknown. As these structures have recently been reviewed [25], we will not discuss them here.

**Helicases**

The double-stranded structure of genomic DNA is an impediment to replication because it buries the sequence information required to copy each strand. This feature makes unwinding double-stranded DNA (dsDNA) a requirement for replication, as well as for other types of reactions involving duplex nucleic acids, such as transcription. Because DNA unwinding is energetically unfavorable, cells have evolved enzymes known as helicases that couple the energy of NTP binding and hydrolysis to unwinding [26]. Families of helicases share several sequence and structural motifs, implying that there is a common unwinding mechanism used by these enzymes [27]. The differences that are observed between helicase groups are thought to be important for the specific roles that the enzymes serve in vivo. For example, one feature common to known replicative helicases is that they form hexameric rings that encircle DNA. Such a structure may be necessary for processive unwinding of DNA at replication forks.

Several high-resolution structures of helicases have been described [28–32], and recently the first high-resolution glimpse of a replicative hexameric helicase (T7 bacteriophage gp4 protein) was published [33]. The full-length T7 gp4 protein contains both primase and helicase activities as individual domains in a single protein [34]. These domains share sequence homology to the bacterial DnaB and DnaG families, respectively [35]. Similar to DnaB, gp4 oligomerizes to an active hexameric form and can unwind DNA processively and rapidly in the context of its replication. Because all previously solved helicase structures are monomeric [28–32], the structural basis for this oligomerization has been unclear. To examine the structure of gp4, Ellenberger and coworkers [35] crystalized a truncated version of the protein (gp4-helix) that removed the primase domain and portions of the helicase domain. The effect of these truncations is the loss of both primase and helicase activities, but because an NTP hydrolysis activity was retained in the minimal construct, gp4-helix can be considered a good first representation of a hexameric helicase structure [35]. Solution of the gp4 structure represents a milestone in our understanding of replication at a structural level as it affords a visual explanation of many features unique to hexameric helicases.
The structure of gp4-hel reveals that the truncated molecule forms a helical filament with 6₁ symmetry, rather than a closed ring with true sixfold symmetry (Figure 2a). This arrangement is somewhat surprising because electron micrographs of full-length gp4 had previously revealed a closed-circular hexameric structure for the molecule [36]. However, the six adjacent protomers of one filament turn in the crystal can be reoriented with simple translations and other minor modifications to assemble a structure with true sixfold symmetry that is consistent with the lower-resolution electron micrographs of gp4 [33]. Furthermore, Wigley and coworkers have observed that a slightly longer gp4 helicase construct forms a closed-ring structure and is active as a helicase; the structure of this gp4 domain is currently in progress (D. Wigley, personal communication). The 6₁-symmetric arrangement of proteins within the crystal is strikingly similar to the symmetry observed for *E. coli* RecA, a ssDNA-binding recombination protein [37]. Structural similarity between RecA and other helicase structures has been noted before, but gp4-hel is the first helicase for which a RecA-like quaternary arrangement is also observed [28,38].

Oligomerization of the gp4-hel domains is accomplished primarily by domain swapping and, to a lesser extent, by shared active-site elements (Figure 2b). The amino-terminal 11 residues of each gp4 protomer are helical and form a subdomain with two helices that are present on the adjacent protomer. Domain swapping of this sort has been observed in a variety of oligomeric proteins and offers a means of coordinating the activities of one protomer with others in the complex [39]. Oligomeric helicases are well-known for their cooperativity in nucleotide binding and hydrolysis [26]; the observation of a swapped helix linking helicase protomers offers an exciting new structural platform for understanding this enzymological feature.

All known helicases share sequence homology in two motifs important for NTP binding and hydrolysis, the Walker A and B boxes [27,40]. This homology also extends to three-dimensional similarity in the positions of active-site residues in known helicase structures, including gp4-hel [41]. However, the active-site of gp4-hel offers another clue explaining the cooperativity of NTP binding and hydrolysis in oligomeric helicases. Although the positions of the relevant conserved residues are consistent with monomeric helicase crystal structures, the gp4-hel active site is a composite region made up of conserved residues from two individual protomers. Specifically, a conserved active-site arginine residue is contributed from one protomer to another (Figure 2b). The result is that the NTP-binding and -hydrolysis site is shared between two protomers [33]; a shared active site is also observed in the structure of another hexameric ATPase, the F₁ATP synthase [42]. Contrariwise, in monomeric helicase structures, the NTPase site is made up of two homologous domains contained within a single, contiguous polypeptide chain. It is therefore expected that cooperativity arises from the ability of residues in one protomer to ‘sense’ the presence or hydrolytic state of nucleotide in the active site of the adjacent protomer.

DNA polymerases

The first replication protein solved to high-resolution was the Klenow fragment of *E. coli* DNA polymerase I (pol I) [43]. Although this enzyme is not the replicative polymerase of *E. coli*, the pol I structure has been shown to define the general structure of all known replicative polymerases. Many excellent, up-to-date reviews of polymerase structure are available [44–48] so we will only briefly review their structural and functional highlights here. Over a dozen different DNA polymerase structures have been solved and each can be likened to a right hand with palm, thumb, and fingers domains (Figure 3a,d) [43]. The ‘palm’ portion of the polymerase acts as the catalytic center by binding two active-site divalent metals, as well as the 3’-hydroxyl nucleophile of the growing DNA chain.
and its \( \alpha \)-phosphate target on the incoming dNTP [45]. This region of the polymerase forms a canonical fold that is shared across nearly all known polymerase structures, consisting of an antiparallel \( \beta \) sheet buttressed on one side with two \( \alpha \) helices. The metal-liganding residues reside on the ends of two \( \beta \) strands of the sheet. The ‘thumb’ and ‘fingers’ regions form two canyon walls around the palm floor and contribute to template ssDNA, product dsDNA, and the \( 3' \rightarrow 5' \) exonuclease regions.

Processivity factors/clamp loaders
To ensure that DNA polymerases stay bound to their substrates throughout replication, cells and viruses have evolved proteins that enhance polymerase processivity [53,54]. These processivity factors fall into two structural categories: ring-shaped ‘sliding clamps’ that encircle dsDNA and are therefore topologically linked to the polymerase product, and globular accessory proteins that stabilize the polymerase–DNA complex (Figure 3a,b,d). Cells appear to use only the first mode of processivity enhancement. To date, a number of structures of sliding clamps have been determined: the dimeric \( \beta \) subunit of \( E. coli \) pol-HE [55], the trimeric gp45 protein from RB69 bacteriophage [51] and the trimeric proliferating cell nuclear antigen (PCNA) of \( S. cerevisiae \) [56] and \( H. sapiens \) [57].

Although the oligomerization state of clamps differs among species, their overall structures are strikingly similar (Figure 3b,d). The sliding clamps all form circular, pseudo-six-fold ring structures with a continuous \( \beta \) sheet framing the outside of the toroid and \( \alpha \) display of helices on the inside. Individual protomers of the
dimeric *E. coli* β protein have three nearly identical domains, whereas the trimeric clamps have two homologous domains per protomer. The result is that each clamp consists of six structurally similar domains, regardless of whether the functional form is dimeric or trimeric. The dimensions of the ring holes are similar and consistent with dsDNA fitting through the center (Figure 3d). It is interesting to note that the site for clamp–polymerase interactions has been identified in cocystalization studies [51,57]. This identification has allowed one of the first reasonable attempts at building a high-resolution model of interacting replication proteins (see the section on future structural studies).

Besides the clamp mode of processivity enhancement, at least one phage DNA replication system uses the second strategy, namely that of binding a monomeric, globular protein that stabilizes the polymerase–DNA interaction. In the T7 phage replication system, the *E. coli* thioredoxin protein is coopted by the polymerase to act as an enhancer of replication processivity [58]. In this complex, thioredoxin binds to the DNA polymerase at the thumb domain and is thought to partially wrap around the dsDNA replication product [49] (Figure 3a). This mechanism therefore appears to be a simplified version of the sliding-clamp method.

Because sliding clamps form ring structures spontaneously, cell-replication systems require specialized clamp-loading proteins to open and position the clamp on DNA. Clamp loaders utilize cycles of ATP binding and hydrolysis to open sliding-clamp rings and load them onto primed DNA in a reaction that is coordinated with primase activity [18,59]. In *E. coli*, the clamp loader (γ complex) is composed of five subunits (γ, δ, δ', χ, and ψ). δ' is the only component of known structure from this complex. δ' forms a ‘C-shaped’ molecule with an NTP-binding site located in one of its domains [60] (Figure 3c). Although inactive as an ATPase, δ' significantly resembles γ (the active clamp-loader ATPase subunit of the γ complex) in sequence and therefore helps in understanding the catalytic process of clamp loading [60]. It is thought that ATP binding and/or hydrolysis facilitates a conformational change in the γ subunit to cause the two domains of the ‘C’ to separate, thus opening the processivity ring and loading it onto primed DNA. The mechanistic details of this reaction are still unclear and await biochemical and structural insights into the workings of the γ–β supercomplex.

### Single-stranded DNA-binding proteins

Unwinding of dsDNA by helicases produces ssDNA, the intermediate DNA structure used by polymerases in replication. This form of DNA is particularly susceptible to damage by nucleases and can also form intranastand secondary structures that inhibit both DNA replication and other cellular DNA metabolic processes. To protect ssDNA and restrict its conformation, cells use SSBs to sequester the unpaired nucleic acid strands [61,62].

The structures of several SSBs have been determined, including phage T4 gp32 protein [63], *E. coli* SSB [64], the subunits of human replication protein A (RPA) [65,66], human mitochondrial SSB [67], several filamentous phage SSBs [68–70] and adenovirus SSB [71] (Figure 4a). Most of these structures share a common oligonucleotide/oligosaccharide-binding (OB) domain fold, which consists of a five-stranded, antiparallel β barrel that is puckered on one side. In known ssDNA–SSB complexes [63,65],
ssDNA is bound in the puckered surface of the β barrel, similar to binding strategies seen in other types of OB folds [72]. As with the sliding clamps, differences in oligomeric state among different SSB species can be compensated for by repetition of homologous domains within the proteins. For example, E. coli SSB forms a homotetramer containing one OB fold per protomer [64], whereas human RPA is a heterotrimer composed of one large subunit with two OB folds and two smaller subunits, each bearing a single OB fold [65,66]. Both proteins, therefore, have four OB folds in their final active form, but arrive at that number in different ways. The importance of this fourfold arrangement for replicative SSBs is unclear.

Terminators
At the end of replication, ~180° away from the origin in a circular bacterial chromosome, replication forks are halted by replication termination proteins (called RTP in Bacillus subtilis or ‘tus’ in E. coli) [73]. These proteins bind to specific DNA sequences (‘ter’ sites) and inhibit replication-fork helicase activity in a polar fashion. For a given tus-ter or RTP-ter complex, therefore, only a single fork progressing from one direction will be remobilized while a replication fork approaching from the other side is allowed to pass [74]. This polarity ensures that a clockwise-moving replication fork starting at on 0° passes through one tus-ter site at 179.9° but terminates at another, oppositely facing tus-ter at 180.1°. Meanwhile, the counterclockwise fork passes through the tus-ter at 180.1° but terminates at the 179.9° complex. In this way, the circular bacterial chromosome can be fully copied by bidirectional replication.

Two terminator proteins have been solved to high resolution, B. subtilis RTP [75] and a ter-bound form of E. coli tus [76] (Figure 4b). The structures are significantly different: RTP has a helix–turn–helix (HTH) domain that has been shown biochemically to be its DNA recognition surface, whereas tus binds ter primarily through β sheet–major groove interactions. The tus–ter complex is asymmetric, as would be expected given the polarity of its biochemical activity. Mechanisms to explain terminator activity invoke an asymmetric steric block that inhibits DNA helicases in one direction but allows translocation from the other direction. Because terminators also inhibit nonreplicative processes such as transcription in a polar fashion, it appears unlikely that a specific complex is formed between the replicative helicase and the termination apparatus [73].

Future structural studies in replication systems
Structural studies have helped reveal the molecular details of replication in several ways. First, determination of the structures of several examples of functionally analogous proteins reveals striking structural similarities between phage and bacteria, as well as with eukaryotes, often despite a lack of clear-cut sequence homology. This suggests that the basic replication reaction is ancient and conserved across the biosphere. Second, structures have helped answer several mechanistic questions about the reactions of replication, including the nature of processivity enhancement and the structural basis of dNTP addition to growing DNA chains. Finally, structural data have offered a visual basis for understanding replication that might not have arisen solely from biochemical or genetic studies.

What lies ahead? Figure 1 indicates that a majority of individual replication proteins have been determined to high-resolution. What is not known, however, is how these components fit together into the biologically relevant complexes that coordinate their activities. Structure determination of complexes is difficult, and is made even more difficult for replication proteins by the dynamic nature of these complexes. One approach to solving this problem is to crystallize minimal interactive regions from multiple proteins of known structure and to extrapolate these interactions to a model of the larger protein–protein complex. A second approach is to find stabilized homogenous samples of replication complexes, either by covalently linking together proteins or perhaps by exploring replication complexes from thermophilic organisms. An example of the first approach is exemplified in a recent experiment that has led to a convincing polymerase–sliding clamp complex model from the bacteriophage RB69 [51]. In this work, Shamo and Steitz [52] solved the structures of the sliding clamp and DNA-bound form of the replicative DNA polymerase from RB69. They then determined the structure of a complex between the clamp and the region of the DNA polymerase that links the two proteins, thus allowing them to build a model of the clamp–polymerase–DNA complex (Figure 4d). Studies employing both approaches in determining how multiprotein complexes assemble and coordinate their functions will dictate the future of structural research on replication.

Acknowledgements
We thank Dale Wigley for communicating results prior to publication and Peter Garber for careful reading of the manuscript. This review was supported by the G. Harold & Leila Y. Mathers Charitable Foundation (J.M.B.) and The Jane Coffin Childs Memorial Fund for Medical Research (J.L.K.).

References


