ARTICLES

Structural basis for transcription elongation by bacterial RNA polymerase

Dmitry G. Vassylyev[1], Marina N. Vassylyeva[1], Anna Perederina[1], Tahir H. Tahirov[2] & Irina Artsimovitch[3]

The RNA polymerase elongation complex (EC) is both highly stable and processive, rapidly extending RNA chains for thousands of nucleotides. Understanding the mechanisms of elongation and its regulation requires detailed information about the structural organization of the EC. Here we report the 2.5-Å resolution structure of the *Thermus thermophilus* EC; the structure reveals the post-translocated intermediate with the DNA template in the active site available for pairing with the substrate. DNA strand separation occurs one position downstream of the active site, implying that only one substrate at a time can specifically bind to the EC. The upstream edge of the RNA/DNA hybrid stacks on the β′-subunit 'lid' loop, whereas the first displaced RNA base is trapped within a protein pocket, suggesting a mechanism for RNA displacement. The RNA is threaded through the RNA exit channel, where it adopts a conformation mimicking that of a single strand within a double helix, providing insight into a mechanism for hairpin-dependent pausing and termination.

In all organisms, transcription performed by DNA-dependent RNA polymerases (RNAPs) can be divided into three mechanistically and structurally distinct stages: initiation, elongation and termination. During initiation, RNAP recognizes a promoter, unwinds DNA near the start site (register $i + 1$) and begins RNA synthesis, using NTPs as both a primer and the substrates. Initiation is characterized by multiple rounds of abortive synthesis during which RNAP synthesizes and releases short RNA products. Once RNAP has made an RNA molecule 13–15 nucleotide long, in which 7–9 nucleotides pair with the DNA template strand in an RNA/DNA hybrid, the transcription complex undergoes promoter clearance and makes a transition to the elongation phase.

Elongation is highly processive: the EC is capable of the uninterrupted synthesis of RNA chains thousands of nucleotides long, yet becomes abruptly destabilized at terminators that demarcate the RNA end, in many cases with single-nucleotide precision. The interplay between processive synthesis, transient halting at numerous 'roadblocks' and RNA release depends on the intricate network of interactions between RNAP, the nucleic acid signals and/or auxiliary transcription factors within the EC. High-resolution structural studies of the EC are therefore of central importance for understanding the general mechanisms and basic principles of transcription and the regulation of gene expression in the living cell.

In contrast with initiation complexes, which are vastly different in protein content and complexity, the core enzymes that perform elongation are highly homologous (in both sequence[1] and structure[2]) in prokaryotic and eukaryotic multisubunit RNAPs. Thus, the structural results and implications obtained on the ECs from either system would be of general interest. Until now there have been no structural data for the bacterial EC, whereas several medium-resolution structures were determined for various configurations of the EC formed by yeast RNAPII[3–7]. Although these studies yielded important insights into the overall arrangement of the nucleic acids within the EC, the available structures lack many details crucial for understanding the mechanistic aspects of transcript elongation. Moreover, the results and their interpretation provided by different groups were generally consistent but left several important issues unresolved, including, for example, the point of downstream DNA (dwDNA) strand separation. This further underscores the necessity of obtaining more definitive and highly resolved structural data for the multisubunit ECs.

We have determined the structure of the *T. thermophilus* RNAP EC (ttEC) at 2.5 Å resolution in which the core enzyme (subunit composition $\alpha_2\beta\beta'\omega$) was bound to a synthetic scaffold containing 14 base pairs (bp) of the dwDNA, 9 bp of the RNA/DNA hybrid and seven single-stranded nucleotides of the displaced RNA transcript (Fig. 1 and Supplementary Fig. 1; see also Methods). This structure—at the highest resolution so far for any multisubunit EC—unambiguously identifies the positions of all the nucleic acid chains at the atomic level and reveals the detailed network of their interactions with the protein side chains, thereby providing a framework for explaining the determinants of the high stability and processivity of the EC. Our data provide several testable implications on several crucial aspects that have remained obscure until now: first, the dwDNA configuration and the mechanism of the DNA strand separation; second, the optimal length of the RNA/DNA hybrid and the mechanism of RNA displacement from the template; third, the mechanism of DNA translocation; and last, the mechanism of the hairpin-dependent pausing and termination.

The core enzyme structure in the EC

Formation of the EC is not accompanied by large alterations of the core RNAP structure: indeed, in the holoenzyme[8] (which contains an additional protein subunit, σ, that enables promoter-specific transcription initiation) all functionally significant structural elements—the RNA exit channel as well as the binding cavities for the dwDNA and the RNA/DNA hybrid—are already preformed. The last two are, however, partly occupied by the domains of σ that hinder the growth of nascent RNA—thus, the release of most[9] of the σ-factor contacts with the core is required for the formation of the stable processive EC. This is consistent with the recently reported 'scrunching' of nucleic acids[10,11] during the transition from initiation to elongation in the

[1]Department of Biochemistry and Molecular Genetics, University of Alabama at Birmingham, Schools of Medicine and Dentistry, 402B Kaul Genetics Building, 720 20th Street South, Birmingham, Alabama 35294, USA. [2]Eppley Institute for Research in Cancer and Allied Diseases, University of Nebraska Medical Center, Lied Transplant Center, 10737A, 986805 Nebraska Medical Center, Omaha, Nebraska 68198-7696, USA. [3]Department of Microbiology, The Ohio State University, 484 West 12th Avenue, Columbus, Ohio 43210, USA.

bacterial multisubunit RNAP, and contrasts with the marked structural refolding that accompanies this transition in the single-subunit T7 RNAP[12,13]. In the bacterial system, the displacement and eventual release of the σ-factor takes the place of the refolding that is required to expand the active-site cavity and to accommodate the 8–9-bp-long RNA/DNA hybrid in the T7 system.

Nevertheless, the core enzyme conformations in the EC and the holoenzyme are not identical, demonstrating the most significant changes in those structural domains that switch their contacts from the σ-subunit in the holoenzyme to the nucleic acids in the EC. These alterations demarcate the transition from the initiation to the elongation phase and probably contribute to the high stability of the EC, a feature that distinguishes it markedly from the initiation complex. The most notable rearrangements involve the β and β′ domains forming the pincers of the crab-claw-like structure of the core enzyme[14] to constitute the main channel accommodating the dwDNA and the RNA/DNA hybrid (Fig. 1 and Supplementary Fig. 2). First, the β′-pincer (the major part of the 'clamp' in the eukaryotic system[15]) domain consisting of the amino-terminal α-helical coiled coil (β′CC1; residues 540–581), the major binding site for the σ-subunit, and two loops ('rudder', β′582–602, and 'lid', β′525–539) moves away from its position in the holoenzyme by ~5.5 Å, bringing the lid into a stack with the upstream edge of the RNA/DNA hybrid and positioning the rudder right between the dwDNA and the hybrid (see below) (Supplementary Fig. 2a, b). Second, the movement of the coiled coil is accompanied by the corresponding shift of the adjacent β′ N-terminal domain (β′51–499), which places it near the downstream edge of the dwDNA on one side (β′102–132; β′470–499) and contributes to the formation of the RNA exit channel on the other (β′51–87; Supplementary Fig. 2a, c, d). Third, in the β-pincer portion of the claws, the β-domain (β1–130; β335–396) that has lost its interactions with the σ-regions 2.4–3.1 (ND2 domain)[8] moves by ~3.0 Å towards the RNA/DNA hybrid, whereas the adjacent β-domain (β140–332) approaches dwDNA after the ~4.0 Å reorientation, thereby closing the dwDNA-binding cleft in the EC (Supplementary Fig. 2a–c). The latter movement is probably attributable to the presence of the dwDNA rather than to the absence of the σ-subunit. Last, the entire β-flap (β703–830) domain that forms one wall of the RNA exit channel moves slightly (~2.2 Å) towards the displaced transcript in the EC (Supplementary Fig. 2d). However, this rearrangement is compensated for by the corresponding repositioning of the β′ N-terminal domain (see above) forming another wall of the channel; thus, the size and shape of the RNA exit channel remain nearly unchanged. Overall, in comparison with the holoenzyme, the EC structure undergoes closure of the RNAP claws to reduce the size of the main channel substantially and probably to achieve surface complementarity between the protein and the nucleic acid chains, thereby favouring the high stability and processivity of the EC.

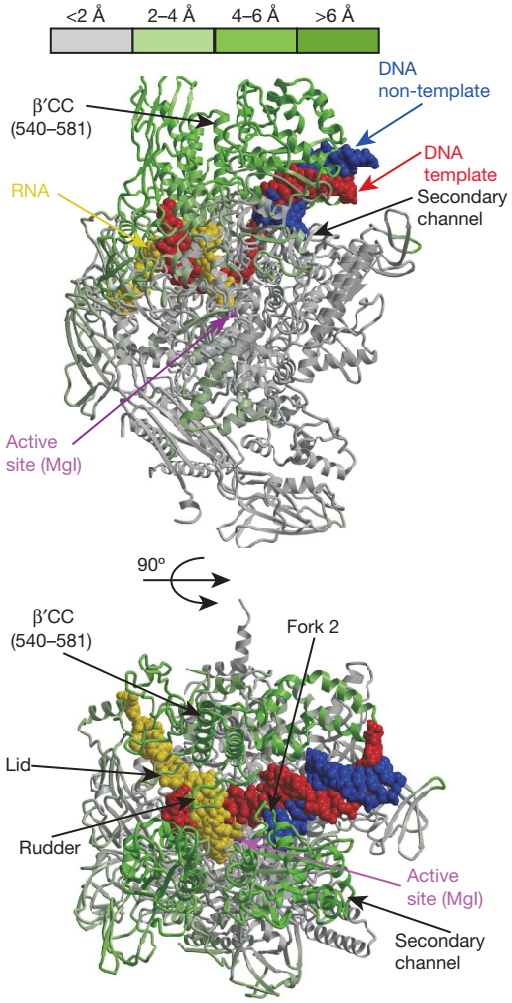

Figure 1 | Overall structure of the ttEC. In the two views shown (the lower panel shows the view through the RNAP main channel), the RNAP core enzyme structure in the ttEC is coloured from grey (less than 2 Å) to green (more than 6 Å) according to the deviation in the positions of the C_α atoms of the corresponding core and holoenzyme (PDB ID 2A6H) residues. Except for the variable colour palette used for RNAP in this figure, the same colour scheme (grey, RNAP; red, DNA template; blue, DNA non-template; yellow, RNA; magenta, the high-affinity Mg^{2+} ion, MgI, in the active site) is used in all figures.

Nucleic acid structure

As observed previously in the structures of T7 RNAP[13,16,17] and the yeast EC[4,5], the template strand in the ttEC forms a sharp kink (~90°) at the junction between the dwDNA and the RNA/DNA hybrid (Fig. 1 and Supplementary Fig. 1). The orientation of the nucleic acid duplexes differs substantially from that in the T7 EC (which possesses a distinct architecture) yet closely resembles the eukaryotic complexes, underscoring the structural homology between multisubunit RNAPs (Fig. 2a).

The nascent seven-nucleotide single-stranded RNA is threaded towards the surface through the previously proposed RNA exit channel[18]. In the channel, six out of seven RNA bases stack on each other, whereas the phosphate backbone conformation mimics that in the A-form nucleic acid duplex (Fig. 2b); only a few direct and several water-mediated hydrogen bonds are formed between the protein and the RNA nucleotides (Fig. 3), suggesting that the RNA conformation is maintained mostly through the self-stabilizing base-stacking contacts and/or solvent-mediated interactions. Several acidic side chains and carbonyl main-chain oxygens approach the RNA backbone, indicative of a possible involvement of the Mg^{2+} ions in mediating these interactions.

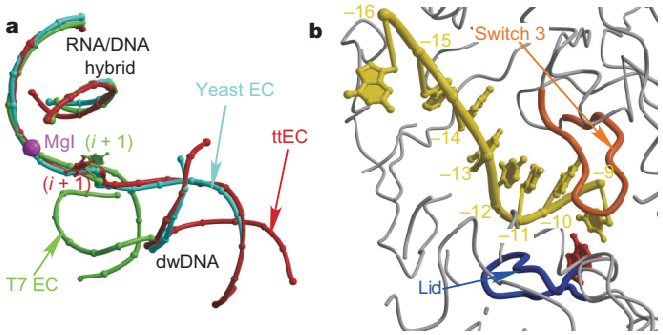

Figure 2 | Nucleic acid structure in the ttEC. a, Superposition of the nucleic acid phosphate backbones in the ttEC (red), yeast EC (cyan)[4] and T7 EC (green)[16]. **b**, Conformation of the displaced RNA nucleotide in the ttEC.

Protein–nucleic-acid interactions

The dwDNA adopts a B-form conformation that is substantially distorted only at registers +2 and +3, immediately downstream of the unpaired acceptor DNA template (register $i+1$) where the kink occurs. A similar distortion was also observed in the eukaryotic EC structure[4], further underscoring the similarity between the two systems. Although all 13 bp of the dwDNA are buried within the closed channel formed by the β- and β′-pincers, there are very few direct polar interactions with the protein at the downstream dwDNA edge (Fig. 3), suggesting that the DNA conformation is maintained through self-stabilizing base pairing, whereas the protein stabilizes the duplex orientation through long-distance electrostatic and van der Waals interactions. The same pattern was observed in the T7 EC[13,16], implying that dwDNA mobility is a general feature that affords high processivity during elongation. The major contacts of RNAP to the dwDNA between registers +2 and +5 probably stabilize the kink in the nucleic acid and help to position the acceptor template base properly in the active site (Fig. 3). Particularly noteworthy is a β′-subunit three-helix bundle (β′101–131) that forms several polar and van der Waals interactions with the dwDNA phosphates (Fig. 4a and Supplementary Fig. 3a). Val β′108 located in the loop between the first and second helices inserts into the minor groove (Fig. 4a) and may prevent translational sliding of the nucleic acids while inducing the screw-like motion during DNA translocation. His 772 probably has a similar function in the T7 EC (Fig. 4b). In yeast RNAPII this region corresponds to the N terminus of the clamp head, which (according to the structural alignment) contains a uniform helix that superimposes well on the first and second helices of the bacterial bundle but lacks the Val β′108-loop insertion (Supplementary Fig. 4a). Inspection of the yeast EC structure[4] revealed another structural element at the very carboxy terminus of the cleft domain, which is represented by a uniform α-helix in the bacterial enzyme but is interrupted by insertions of two Rpb1 residues, Arg 1386 and His 1387, whose side chains are positioned in the dwDNA minor groove mimicking the Val β′108 loop (Fig. 4c and Supplementary Fig. 4b). The exact location of the loop insertion into the minor groove is, however, different in each complex: in yeast and T7 ECs it occurs closer to the upstream edge of the dwDNA than to the ttEC.

The RNA/DNA hybrid is tightly packed in the active-site cavity, where its phosphate backbone forms multiple polar and van der Waals interactions with the conserved protein residues (Fig. 3; see also Supplementary Table 2 in ref. 19). In this 'mixed' mode of binding (also observed in the T7 EC[13,16]) the van der Waals contacts may have two essential functions. First, repulsion between the hydrophobic side chains and charged phosphate groups may weaken hydrogen bonds that the latter form with the basic residues, thereby avoiding excessively strong binding of the nucleic acids that could reduce the EC processivity. Second, they may act as 'shape sensors' that monitor the hybrid configuration to provide feedback in response to mismatches and/or occasional dNMP incorporations, thereby contributing to the fidelity of transcription.

The rudder loop is positioned between the RNA/DNA hybrid and the dwDNA (Fig. 4d), suggesting that although it is unlikely to function in RNA displacement and/or separation of the upstream hybrid

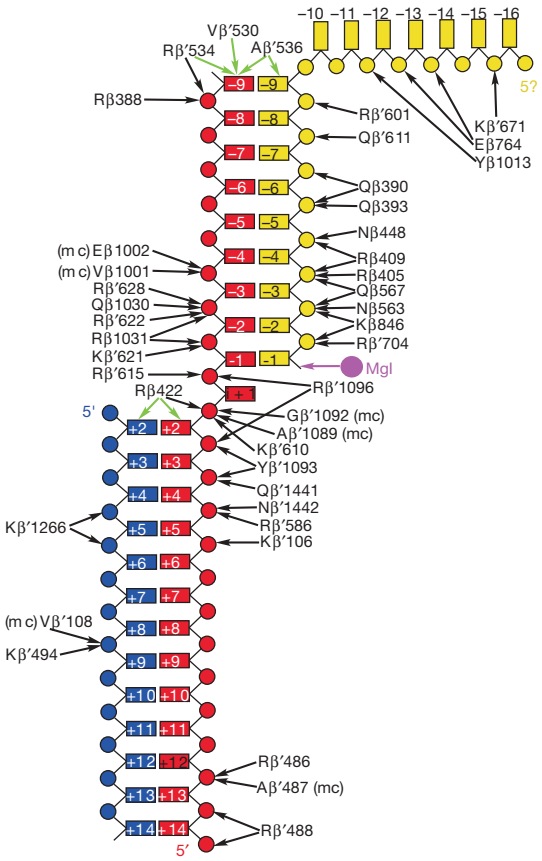

Figure 3 | Schematic drawing of the protein/nucleic acid contacts. Polar and stacking interactions are shown as black and green arrows, respectively. The ($i+1$) non-template nucleotide was not resolved in the electron density map and is therefore not shown.

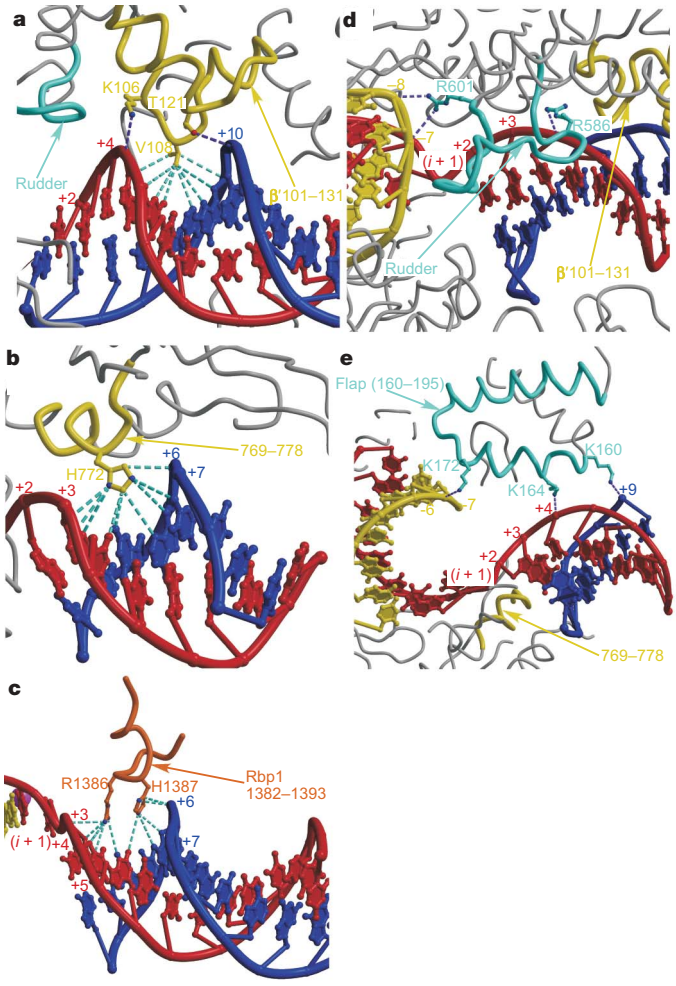

Figure 4 | Protein–nucleic acid interactions in the ttEC structure. **a**–**c**, Protein segments inserted in the minor groove of the dwDNA in the ttEC (**a**), T7 EC (**b**) and eukaryotic EC (**c**). **d**, **e**, Rudder loop (**d**) and flap domain (**e**) demarcate the RNA/DNA hybrid and dwDNA-binding sites in the ttEC and T7 EC, respectively. The van der Waals and polar interactions are shown by cyan and purple dashed lines, respectively.

strand[18,20], it contributes to the overall stability of the transcription bubble, a function consistent with the biochemical data[21]. Indeed, the rudder demarcates the hybrid and the dwDNA, while stabilizing their position and conformation through the direct contacts; two Arg residues (β′586 and β′601) located at its N and C termini interact directly with the dwDNA phosphates and the hybrid, respectively (Fig. 4d). The rudder interacts with the upstream edge of the RNA/DNA hybrid (registers −7/−8); thus, the contacts bridging the hybrid and the dwDNA may account for the higher stability of the EC than the early initiation complex. In contrast, the proper positioning of the rudder and the setting-up of these bridging interactions may help to achieve the successful transition to elongation. The flap subdomain in T7 RNAP (which is characteristic of the EC but is missing from the initiation complex[12,13,16,22]) also bridges the hybrid and the dwDNA and is probably a functional counterpart of the rudder[23] (Fig. 4e).

dwDNA and RNA/DNA hybrid strand separation

During elongation, RNAP performs thousands of nucleotide addition cycles. Each cycle must culminate in forward translocation by one nucleotide to allow for the incorporation of the next substrate; this step entails the separation of 1 bp of the dwDNA accompanied by the displacement of one nucleotide of the nascent RNA from the DNA template at the upstream edge of the RNA/DNA hybrid and subsequent annealing of the upstream DNA duplex. The mechanisms of the dwDNA and the RNA/DNA hybrid strand separation have been largely explained for the single-subunit enzymes[12,13], but remain poorly understood for the multisubunit RNAPs.

Two models based on the medium-resolution structures of the yeast ECs[3,6] differ with regard to the position of the melting of the dwDNA: in one model, strand separation occurs deep inside the dwDNA-binding channel (register +5)[6], whereas in another the dwDNA duplex is maintained at least up to the +3 register[4]. However, it is difficult to compare these models because the scaffold used in one work lacked the non-template strand up to the +4 register[6], whereas the mismatch was present in the dwDNA at register +2 in the other[4]. In the ttEC structure, in which we used the dwDNA that is fully complementary along its entire length (14 bp), the duplex is maintained up to and including the +2 register, with only one base pair open at register +1 (Supplementary Fig. 1), making the acceptor template in the active site available for base pairing with the incoming substrate. This observation is fully supported by a large body of work on the bacterial transcription complexes (see ref. 18 and references therein) and thus indicates that only one substrate at a time can bind specifically to the EC (see ref. 19 for additional evidence), while arguing against 'multisubstrate' models[24–26]. It is worth mentioning that our scaffold lacks the non-template strand upstream of the insertion site; the consequent lack of the upstream DNA duplex may indirectly affect the dwDNA strand separation. In the bacterial EC, the N terminus (β418–424) of loop 2 of the β fork (fork 2, β413–451; Fig. 1 and Supplementary Fig. 3b) acts as a steric barrier to the passage of the dwDNA towards the active site, probably having a central function in facilitating strand separation. Moreover, Arg β422 at the tip of this loop forms a hydrogen bond with the acceptor template phosphate and stacks on the +2 base pair (Fig. 5a); similarly, in T7 RNAP Phe 664 stacks on the +2 template base, and the main chain amino group of the neighbouring Gly 645 is hydrogen-bonded to the acceptor template phosphate[12] (Fig. 5b). It is noteworthy that the F644A and G645A mutants are among four T7 RNAP variants known to show a significant increase in the incorporation of mismatched bases[27], suggesting that these interactions may be coupled to substrate loading and selection[12]. This analogy suggests, first, that Arg β422 may be a fidelity determinant in the bacterial RNAP, and second, that base pairing at register +2 might be essential for the proper positioning of the acceptor DNA base in the active site, which in turn may affect recognition and binding of the cognate substrate. Although poor sequence alignment and conformational differences preclude unambiguous identification of the Arg β422 counterpart in eukaryotic fork 2, two basic residues that are spatially close to Arg β422 in the yeast structure (Arg 504 and Lys 507)[4] may have a similar function. In contrast, although the location of fork 2 in one of the eukaryotic ECs is similar, suggesting the same mechanism and position of separation of the dwDNA strand[4], it adopts different conformations in other complexes[5] and may disrupt the +5 base pair. Obtaining an unambiguous answer to this question would require a structural analysis of the eukaryotic ECs assembled on a nucleic acid scaffold containing a complete dwDNA duplex as in the present study.

The ttEC structure shows that RNAP can readily accommodate 9 bp of the RNA/DNA hybrid; its upstream edge (register −9) interacts with the lid loop that was previously proposed to facilitate the RNA displacement in eukaryotic RNAP[3] (Fig. 5c). In contrast to the yeast enzyme, however, in which this loop forms no direct contacts with the nucleic acids, in the ttEC several residues (conserved between bacteria and eukaryotes) from the lid stack onto the upstream base pair of the hybrid, mimicking the base-stacking interactions within the nucleic acid duplex (Fig. 5c and Supplementary Fig. 5). The lid therefore probably sterically blocks further growth of the hybrid, facilitating strand separation while simultaneously stabilizing the upstream base pair. Given the likely flexibility of this protruding structural segment, its potential reorientations could accommodate the 8-bp or 10-bp RNA/DNA hybrids. The lid may thus serve as an important regulatory element in stabilizing various non-standard hybrid configurations that form during pausing, backtracking or termination, for example.

Although deletion of the lid does indeed result in the formation of persistent RNA/DNA hybrids on single-stranded templates, it seems dispensable in the presence of the non-template strand, suggesting that reannealing of the upstream DNA is crucial for displacement of the RNA[28,29]. In this respect, the trapping of the first displaced RNA base in the hydrophobic pocket formed by the β1012–1022 loop (switch 3) is of particular interest (Fig. 5d). Indeed, if competition

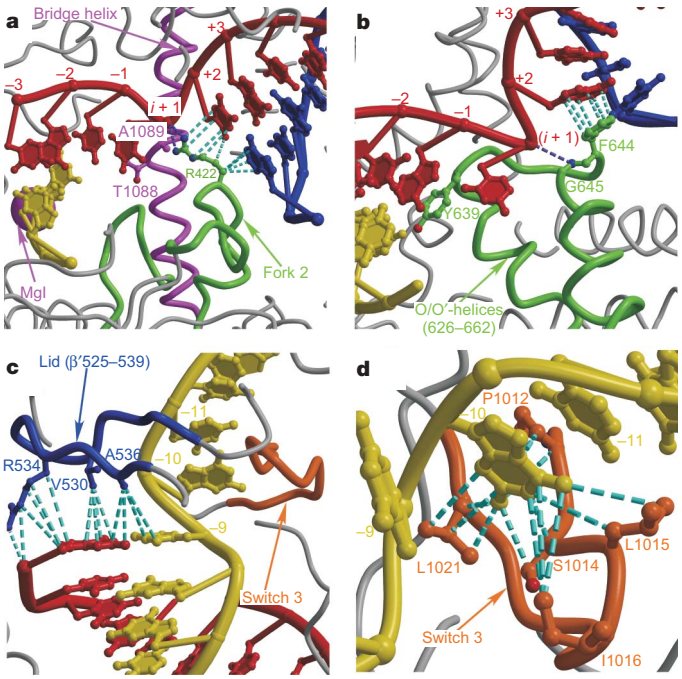

Figure 5 | dwDNA and RNA/DNA hybrid strand separation. a, **b**, Structural elements interacting with the +2 dwDNA base pair in the ttEC (**a**) and T7 EC (**b**). **c**, The lid loop stacks on the upstream RNA/DNA base pair in the ttEC. **d**, The first displaced RNA base is trapped in the protein pocket formed by the displacement loop in the ttEC. The van der Waals and polar interactions are shown by cyan and blue dashed lines, respectively.

is assumed between the displaced RNA and the non-template strand nucleotides for the template strand, the switch 3 loop harbouring the RNA base might have an essential function in the 'DNA-dependent' mechanism of the RNA separation.

Active site and implications for translocation

In the post-translocated ttEC, the acceptor template base $(i+1)$ resides in the active site, where it is sandwiched between the -1 template nucleotide and the bridge helix, mimicking the duplex configuration (Fig. 5a). As a result, its base pairing with the substrate may position the substrate phosphates near the active-site residues, in stark contrast to the post-translocated state of the T7 EC, in which the acceptor template base is displaced by the side chain of Tyr 639 (Fig. 5b), 'forcing' the initial substrate binding into the inactive, 'preinsertion' site located far from the active centre[12] and thus necessitating significant structural rearrangements to deliver the substrate to the active, 'insertion' position. Even though the transition to the 'insertion' state in the bacterial RNAP apparently does not call for such large changes, the structures of the substrate-bound ttECs[19] argue for the conceptually similar two-step mechanism: NTP first binds to the preinsertion configuration followed by isomerization to the closed, insertion state after the substrate-induced folding of the trigger loop, which is disordered in the ligand-free ttEC.

Although the protein/nucleic acid interactions formed at the upstream and downstream edges of the hybrid and the dwDNA, respectively, as well as those made with the first displaced RNA nucleotide and the acceptor template, potentially stabilize the post-translocated ttEC state, the same interactions are probably present in the pre-translocated complex, suggesting that these two states are equally stable in the absence of the substrate. Our analysis failed to reveal structural elements (for example, a 'ratchet') that would preclude the backward DNA translocation in the post-translocated ttEC. In particular, the bridge helix, whose local distortion[8] could potentially freeze the post-translocated state, is uniform in the ttEC. The ttEC structure therefore supports a 'brownian-ratchet' mechanism, in which translocation is driven by thermal motions giving rise to equilibrium between the post-translocated and pre-translocated states of the substrate-free ECs, whereas binding of the cognate substrate would favour the post-translocated state[30]. In contrast, structures of the T7 RNAP EC reveal markedly distinct orientations of the template in the register $(i+1)$ in the post-translocated and pre-translocated conformations[13], prompting a hypothesis that the Tyr 639 side chain replacing the DNA template in the active site acts as a 'ratchet' stabilizing the post-translocated state and blocking reverse translocation.

Implications for pausing and termination

Gene expression in bacteria is frequently controlled by nascent RNA hairpins, which, depending on the context, can slow RNAP down at a pause site to await a regulatory input or trigger RNA release at a terminator. The principal difference between the pause and the terminator hairpins is the register at which the RNA duplex begins, namely $(-11/-10)$ and (less than -8), respectively[31,32]. The 9 bp of the RNA/DNA hybrid and the duplex-like conformation of the displaced RNA observed in the ttEC suggest that the pause hairpin could be formed within the exit channel without major alterations of the enzyme structure. Indeed, modelling demonstrates that the RNA exit channel can accommodate the A-form 5-bp hairpin stem after subtle rotations of the experimental displaced RNA strand ($\sim 22^\circ$) used as a template for the modelling, and the flexible flap-tip helix ($\sim 30^\circ$; β761–785; Supplementary Fig. 6). The eukaryotic RNA exit channel that lacks the flap-tip helix and seems more open could accommodate nascent hairpins even more readily. In the model, the first displaced RNA base -10 is buried in the protein pocket, retaining its position observed in the experimental structure, and is not available for base pairing. In an alternative model the hairpin-induced local alterations of the lid and/or switch 3 loops may result in the formation of an additional RNA/DNA hybrid base pair that would increase the stability of the paused complex, while impeding the translocation and/or RNA displacement and prolonging the pause.

Modelling of the termination hairpin seemed impossible in view of the predicted significant rearrangements of the RNAP structure after hairpin formation at register -8 that triggers forward translocation and RNA release[33] or EC inactivation and dissociation[34]. However, our modelling attempts suggest that these rearrangements would probably involve those structural elements whose orientation is affected by the σ binding in the holoenzyme. Given that both the initiation and termination complexes are highly unstable, it is possible that the termination hairpin would destabilize the complex in a fashion similar to that of the σ-factor. For example, the hairpin might displace the lid/coiled-coil/rudder domain in a σ-like manner (see above), thereby disrupting the rudder–nucleic-acid interactions that affect the EC stability[21].

Concluding remarks

This study provides the first detailed view of the bacterial EC structure and identifies the determinants of the complex stability, processivity, and response to the regulatory nucleic acid signals, which in turn suggest the mechanisms for separation of the nucleic acid strand, DNA translocation, and transitions from the initiation to the elongation stage and finally to the termination stage of transcription. Most remarkably, it allows us to draw numerous parallels between single-unit and multisubunit enzymes and argue that these RNAPs use analogous structural mechanisms of nucleic acid binding and processing, because many fundamental features (such as duplex separation, RNA displacement and active-site architecture) seem strikingly similar at the molecular level. This conclusion is perhaps not surprising in view of the ancient evolutionary origin of the RNAPs and the similarities in response of the extant enzymes from phages, bacteria and eukaryotes to at least a subset of regulatory signals, yet it remains to be widely appreciated in the field. Indeed, scores of adaptive and accidental evolutionary innovations found in prokaryotes and eukaryotes obscure the profound congruity of the basic transcriptional mechanisms in all life kingdoms. However, the progress in understanding the differences in organization of transcriptional machinery between organisms and the impact they have on the regulation of gene expression depends on an understanding of the elemental similarity of the catalytic mechanisms and their structural framework. This work has fitted into place a few more pieces of the transcription puzzle, while highlighting the need for further advances on two fronts: first, continual improvement in the structural data to the near-ångström resolution required to explain the basic catalytic mechanism, and second, a progressive build-up of the crystallized complexes to include the non-template DNA strand, nascent RNA hairpins, and regulatory factors.

METHODS SUMMARY

The structure was determined by the molecular replacement technique[35] by using the core enzyme portion of the *T. thermophilus* holoenzyme structure[8] (PDB entry 1IW7) as a search model (Supplementary Table 1). Application of the zonal scaling to the experimental diffraction data greatly improved the R_{free} and omit electron density (Supplementary Figs 7 and 8 and Supplementary Table 2) and permitted model building of all nucleic acid components of the EC. The structure was refined to a final R-factor/R_{free} of 0.238/0.267 at 2.5 Å resolution (Supplementary Table 3).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 28 January; accepted 11 May 2007.
Published online 20 June 2007.

1. Archambault, J. & Friesen, J. D. Genetics of eukaryotic RNA polymerases I, II, and III. *Microbiol. Rev.* **57**, 703–724 (1993).
2. Cramer, P. Multisubunit RNA polymerases. *Curr. Opin. Struct. Biol.* **12**, 89–97 (2002).

3. Gnatt, A. L., Cramer, P., Fu, J., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 Å resolution. *Science* **292,** 1876–1882 (2001).
4. Kettenberger, H., Armache, K. J. & Cramer, P. Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS. *Mol. Cell* **16,** 955–965 (2004).
5. Wang, D., Bushnell, D. A., Westover, K. D., Kaplan, C. D. & Kornberg, R. D. Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. *Cell* **127,** 941–954 (2006).
6. Westover, K. D., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: nucleotide selection by rotation in the RNA polymerase II active center. *Cell* **119,** 481–489 (2004).
7. Westover, K. D., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: separation of RNA from DNA by RNA polymerase II. *Science* **303,** 1014–1016 (2004).
8. Vassylyev, D. G. *et al.* Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 Å resolution. *Nature* **417,** 712–719 (2002).
9. Mooney, R. A., Darst, S. A. & Landick, R. Sigma and RNA polymerase: an on-again, off-again relationship? *Mol. Cell* **20,** 335–345 (2005).
10. Kapanidis, A. N. *et al.* Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. *Science* **314,** 1144–1147 (2006).
11. Revyakin, A., Liu, C., Ebright, R. H. & Strick, T. R. Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching. *Science* **314,** 1139–1143 (2006).
12. Temiakov, D. *et al.* Structural basis for substrate selection by t7 RNA polymerase. *Cell* **116,** 381–391 (2004).
13. Yin, Y. W. & Steitz, T. A. The structural mechanism of translocation and helicase activity in T7 RNA polymerase. *Cell* **116,** 393–404 (2004).
14. Zhang, G. *et al.* Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution. *Cell* **98,** 811–824 (1999).
15. Cramer, P., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: RNA polymerase II at 2.8 ångstrom resolution. *Science* **292,** 1863–1876 (2001).
16. Tahirov, T. H. *et al.* Structure of a T7 RNA polymerase elongation complex at 2.9 Å resolution. *Nature* **420,** 43–50 (2002).
17. Yin, Y. W. & Steitz, T. A. Structural basis for the transition from initiation to elongation transcription in T7 RNA polymerase. *Science* **298,** 1387–1395 (2002).
18. Korzheva, N. *et al.* A structural model of transcription elongation. *Science* **289,** 619–625 (2000).
19. Vassylyev, D. G. *et al.* Structural basis for substrate loading in bacterial RNA polymerase. *Nature* doi:10.1038/nature05931 (this issue).
20. Naryshkin, N., Revyakin, A., Kim, Y., Mekler, V. & Ebright, R. H. Structural organization of the RNA polymerase-promoter open complex. *Cell* **101,** 601–611 (2000).
21. Kuznedelov, K., Korzheva, N., Mustaev, A. & Severinov, K. Structure-based analysis of RNA polymerase function: the largest subunit's rudder contributes critically to elongation complex stability and is not involved in the maintenance of RNA–DNA hybrid length. *EMBO J.* **21,** 1369–1378 (2002).
22. Cheetham, G. M. & Steitz, T. A. Structure of a transcribing T7 RNA polymerase initiation complex. *Science* **286,** 2305–2309 (1999).
23. Jiang, M., Ma, N., Vassylyev, D. G. & McAllister, W. T. RNA displacement and resolution of the transcription bubble during transcription by T7 RNA polymerase. *Mol. Cell* **15,** 777–788 (2004).
24. Foster, J. E., Holmes, S. F. & Erie, D. A. Allosteric binding of nucleoside triphosphates to RNA polymerase regulates transcription elongation. *Cell* **106,** 243–252 (2001).
25. Gong, X. Q., Zhang, C., Feig, M. & Burton, Z. F. Dynamic error correction and regulation of downstream bubble opening by human RNA polymerase II. *Mol. Cell* **18,** 461–470 (2005).
26. Landick, R. NTP-entry routes in multi-subunit RNA polymerases. *Trends Biochem. Sci.* **30,** 651–654 (2005).
27. Huang, J., Brieba, L. G. & Sousa, R. Misincorporation by wild-type and mutant T7 RNA polymerases: identification of interactions that reduce misincorporation rates by stabilizing the catalytically incompetent open conformation. *Biochemistry* **39,** 11571–11580 (2000).
28. Naryshkina, T., Kuznedelov, K. & Severinov, K. The role of the largest RNA polymerase subunit lid element in preventing the formation of extended RNA-DNA hybrid. *J. Mol. Biol.* **361,** 634–643 (2006).
29. Toulokhonov, I. & Landick, R. The role of the lid element in transcription by *E. coli* RNA polymerase. *J. Mol. Biol.* **361,** 644–658 (2006).
30. Abbondanzieri, E. A., Greenleaf, W. J., Shaevitz, J. W., Landick, R. & Block, S. M. Direct observation of base-pair stepping by RNA polymerase. *Nature* **438,** 460–465 (2005).
31. Artsimovitch, I. & Landick, R. Interaction of a nascent RNA structure with RNA polymerase is required for hairpin-dependent transcriptional pausing but not for transcript release. *Genes Dev.* **12,** 3110–3122 (1998).
32. Yarnell, W. S. & Roberts, J. W. Mechanism of intrinsic transcription termination and antitermination. *Science* **284,** 611–615 (1999).
33. Santangelo, T. J. & Roberts, J. W. Forward translocation is the natural pathway of RNA release at an intrinsic terminator. *Mol. Cell* **14,** 117–126 (2004).
34. Gusarov, I. & Nudler, E. The mechanism of intrinsic transcription termination. *Mol. Cell* **3,** 495–504 (1999).
35. Navaza, J. Implementation of molecular replacement in AMoRe. *Acta Crystallogr. D Biol. Crystallogr.* **57,** 1367–1372 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Temiakov and M. Anikin for assistance in crystallization at the initial stage of the project. We are grateful to R. Landick for helpful discussions and critical reading of the manuscript. Use of the Advanced Photon Source was supported by the US Department of Energy, Office of Energy Research. This work was supported by NIH grants to D.G.V. and I.A.

Author Contributions D.G.V. determined and analysed the structure and guided the project. A.P. purified RNAP. M.N.V. and A.P performed crystallization and data collection. T.H.T assisted with data collection and analysis. I.A. contributed scaffold design and analysis. D.G.V. and I.A. jointly wrote the manuscript.

Author Information The atomic coordinates are deposited in the Protein Data Bank under accession number 2O5I. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.G.V. (e-mail: dmitry@uab.edu).

METHODS

Crystallization and data collection. Initially, the two nucleic acid scaffolds were designed for co-crystallization with the *T. thermophilus* core enzyme that differed only in the length (8 and 9 bp) of the RNA/DNA hybrid[36,37]. Surprisingly, the preliminary biochemical characterization of the ECs assembled on these scaffolds (that might be designated as the EC8 and EC9) revealed that whereas the EC9 was resistant, the EC8 seemed highly sensitive to pyrphosphorolysis, suggesting that the EC9 and EC8 are preferentially in the post-translocated and pre-translocated states, respectively. Consistent with this interpretation, crystallization of these complexes resulted in the crystals obtained under essentially distinct crystallization conditions, belonged to the different space groups (tetragonal and hexagonal for the EC9 and EC8, respectively) and were not cross-reproducible[36,37]. Whereas the EC8 hexagonal (presumably pre-translocated) crystals diffracted to only 7 Å resolution, the very first relatively small EC9 tetragonal crystals diffracted beyond 4 Å and we therefore concentrated on the improvement of these promising crystals.

The optimized crystallization conditions for the EC9 (the well solution contained 30% 2-methyl-2,4-pentandiol, 0.1 M sodium cacodylate pH 6.5 and 0.2 M magnesium acetate) resulted in the large bipyramidal crystals that belonged to the space group $P4_12_12$ with unit cell dimensions $a = b = 156.21$ Å, $c = 499.23$ Å. In spite of their large size, most of the crystals exhibited poor diffraction quality and high mosaicity. After numerous trials, we were able to select only one well-diffracting crystal, from which the preliminary high-resolution data were collected at synchrotron beam line BL6, Photon Factory, Japan[36,37]. Although the data allowed us to solve the structure by molecular replacement and indicated the presence of the nucleic acids in the complex, the electron density (ED) for the nucleotides was of the overall modest quality and was especially poor for the upstream edge of the RNA/DNA hybrid, the downstream edge of the dwDNA and for the displaced RNA transcript. In particular, the ED corresponding to the two template nucleotides at the upstream edge of the hybrid was nearly absent. We presumed that although the ECs showed high stability during the relatively short biochemical experiments, they might be still affected by the oscillation between the post-translocated and pre-translocated states that would permit the excision of the 3′-RNA nucleotides during crystallization lasting several weeks, thereby resulting in a mixture of the different complexes coexisting in the crystal. The analysis of the nucleic acids from the crystals that were growing for about three months revealed that this hypothesis was correct (Supplementary Fig. 9). To resolve this problem, a new scaffold was designed, in which the two 3′-RNA nucleotides were linked by phosphorothioate bonds to inhibit nucleolytic cleavage, and the G/C content in the hybrid and the dwDNA was increased to achieve better stability of the complexes and presumably prevent backtracking. Indeed, the new scaffold seemed substantially more stable in the lengthy crystallization experiments (Supplementary Fig. 9), which resulted in faster-growing and better-diffracting crystals. Though the percentage of crystals diffracting to high resolution was substantially increased, all of them were still characterized by a high, anisotropic mosaicity that, in combination with a quite weak diffraction power, seriously affected the quality of the processed data. To improve the mosaicity, we applied an annealing procedure to the crystals. After multiple trials we found that freezing the crystals in the cryo-stream followed by annealing and soaking them in the cryoprotectant for ~1–2 min greatly improved the mosaicity for some crystals on the second cycle of freezing. This allowed us finally to select a few well-diffracting crystals with a reasonably low mosaicity (0.5–0.7°), from which the best sample was used for high-resolution data collection at the SERCAT beam line (APS, Argonne, USA); the structure was refined by using the twinning option of the CNS program[38] alternating with manual model building[39] to a final *R*-factor/R_{free} of 0.238/0.267 at 2.5 Å resolution (Supplementary Table 3).

All structural figures were prepared with the programs Molscript, Bobscript and Raster3D[40–42].

Zonal scaling: the key for structure solution. Molecular replacement[35] applied to the EC data provided a clear single solution for both rotation and translation functions using the core enzyme extracted from the previously determined structure of the *T. thermophilus* holoenzyme[8] (PDB entry 1IW7) as a search model (Supplementary Table 1). Surprisingly, however, the rigid body and positional refinements performed against the original diffraction data were highly unstable and did not yield a decisive decrease in *R*-factor from the initial value of ~49.8% at 2.5 Å resolution (Supplementary Table 2) providing, in particular, the abnormally high *R*-factors in the lowest resolution shell. As a result, the $(2F_{obs} - F_{calc})$ ED map was characterized by a largely smeared ED in the protein region that was lacking the atomic details, demonstrating ~4.0 Å rather than 2.5 Å resolution quality, and by a fairly poor omit ED for the nucleic acids (Supplementary Figs 7a and 8a). Careful analysis revealed that the data were affected by strong 'mis-scaling' between the low-resolution (~20–7 Å) and high-resolution data. Because this problem could not be resolved through a procedure of the bulk solvent correction[38,43] that uses the uniform solvent approach, we presumed that in the EC crystals the solvent might be non-uniform, thereby providing systematic, but distinct, contributions to the intensities in different resolution ranges. To test this hypothesis, we devised an original procedure in which 'zonal scaling' was applied to the diffraction data before the refinement and modelling of the nucleic acids in the EC using the unbiased (right after the molecular replacement) RNAP coordinates to eliminate or minimize the potential resolution-dependent systematic errors in the experimental data that might arise, in particular, from the non-uniformity of the solvent in the crystals. The EC data were divided into ~180 resolution shells (to allow for ~2,000 reflections per shell), and the individual scale factors between the observed and model structure factor amplitudes were determined and applied to the experimental data for each shell. The zonal scaling protocol is not entirely new and was previously implemented in the program PRISM[44] that used skeletonization of the ED as a technique of phase improvement to provide a better match between the observed and calculated structure factor amplitudes and was used successfully by us in the determination of the T4 endonuclease V/DNA complex structure[45]. Moreover, another commonly accepted scaling procedure, 'local' scaling, although different in details, uses the same idea of the determination of the individual scale factor for a particular reflection through scaling applied to only a limited subset of the reflections adjacent to this reflection in a reciprocal space[46]. This protocol would therefore apparently introduce a similar, resolution-dependent correction to the structure factor amplitudes as zonal scaling. Consistently, application of local scaling to scale the native and heavy-atom derivative data was reported to make a significant improvement in the signal-to-noise ratio in the difference Patterson maps[47], while this option is incorporated in the major programs using direct methods to localize the heavy-atom sites in the derivatives of protein crystals, where it is used to obtain a better estimate for the normalized structure factor amplitudes[48,49]. The novelty of our approach is that we used the zonal scaling procedure towards a refinement of the atomic model.

Zonal scaling, in combination with the bulk solvent correction, led to a marked decrease in the initial *R*-factors (particularly at low resolution) and a smooth rigid-body refinement using the CNS program[38] that converged to an *R*-factor/R_{free} of ~39% after *B*-factor refinement (Supplementary Table 2). Identification of merohedral twinning coupled with the non-crystallographic symmetry similar to that observed previously[37,38,50] (Supplementary Table 3) additionally improved the *R*-factors to ~33% at 2.5 Å resolution just through the rigid-body refinement and provided a high-quality ED for both the existing protein model and, most importantly, for the nucleic acid chains missing from the model (Supplementary Figs 7b and 8b). This ED allowed us to replace the β′-subunit dispensable region (β′165–449) mistraced in the RNAP holoenzyme with the proper model[51], to model all the nucleic acid in the EC unambiguously and to obtain a final high-quality model on refinement[38] (alternating with manual model building[39]) of the whole complex against the 2.5-Å resolution data (*R*-factor/R_{free} = 0.238/0.267; Supplementary Table 2). The final model includes the RNAP core enzyme (α_2, β, β′ and ω subunits, molecular mass ~400 kDa), the 13 bp of the dwDNA, 9 bp of the RNA/DNA hybrid, and seven nucleotides of the displaced RNA transcript (Fig. 1 and Supplementary Fig. 1).

All available corrections to the experimental data and the atomic parameters used during refinement of the macromolecular structures are aimed at decreasing the crystallographic *R*-factor and subsequently increasing the quality of the ED map. Among these, those variable parameters against which the major minimization is performed (coordinates, *B*-factors) might be overfitted during the course of the refinement, resulting in a so-called 'model bias'. There are at least two solid criteria that allow us to detect and avoid a model bias: R_{free}, which would begin increasing, and the omit ED maps, which would become worse on overfitting of the model. Zonal scaling demonstrated an excellent match to both of these criteria. Indeed, this correction applied only once to a partial unrefined model resulted in a decisive decrease in the initial R_{free} value by ~6%, while the refinement against the corrected experimental data yielded an additional overall decrease in R_{free} by ~18% (note that the refinement against the original data set did not converge at all; Supplementary Table 2). In contrast, application of zonal scaling to the EC diffraction data permitted a marked improvement in the omit nucleic acid ED (Supplementary Fig. 8b) thereby providing additional evidence that the zonal correction does not introduce model bias but significantly improves signal-to-noise ratio. It is also worth noting that, although zonal scaling and bulk solvent corrections are both designed to provide a better match between the low-resolution and high-resolution data, their contributions to the structure factors are largely distinct, suggesting that these corrections may account for somewhat different properties of the crystals and are therefore complementary, rather than alternative, techniques. Indeed, application of the bulk solvent correction after zonal scaling resulted in an additional improvement in R_{free} (Supplementary Table 2).

Our preliminary data indicate that zonal scaling would probably aid in the refinement of many crystal structures and in some cases (like that of the EC structure) may emerge as a key technique for structure determination. The detailed analysis of this technique is now underway and the results on the application of zonal scaling to the refinement of the various crystal structures will be published elsewhere.

36. Kashkina, E. *et al.* Elongation complexes of *Thermus thermophilus* RNA polymerase that possess distinct translocation conformations. *Nucleic Acids Res.* **34,** 4036–4045 (2006).
37. Vassylyeva, M. N. *et al.* Purification, crystallization and initial crystallographic analysis of RNA polymerase holoenzyme from *Thermus thermophilus*. *Acta Crystallogr. D Biol. Crystallogr.* **58,** 1497–1500 (2002).
38. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54,** 905–921 (1998).
39. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
40. Merrit, E. A. & Bacon, D. J. Raster3D: photorealistic molecular graphics. *Methods Enzymol.* **277,** 505–524 (1997).
41. Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24,** 946–950 (1991).
42. Esnouf, R. M. Further additions to MolScript version 1.4, including reading and contouring of electron-density maps. *Acta Crystallogr D* **55,** 938–940 (1999).
43. Afonine, P. V., Grosse-Kunstleve, R. W. & Adams, P. D. A robust bulk-solvent correction and anisotropic scaling procedure. *Acta Crystallogr. D Biol. Crystallogr.* **61,** 850–855 (2005).
44. Baker, D., Bystroff, C., Fletterick, R. J. & Agard, D. A. PRISM: topologically constrained phased refinement for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **49,** 429–439 (1993).
45. Vassylyev, D. G. *et al.* Atomic model of a pyrimidine dimer excision repair enzyme complexed with a DNA substrate: structural basis for damaged DNA recognition. *Cell* **83,** 773–782 (1995).
46. Matthews, B. W. & Czerwinski, E. W. Local scaling: a method to reduce systematic errors in isomorphous replacement and anomalous scattering measurements. *Acta Crystallogr. A* **31,** 480–497 (1975).
47. Wang, J. H. *et al.* Structure of a functional fragment of VCAM-1 refined at 1.9 Å resolution. *Acta Crystallogr. D Biol. Crystallogr.* **52,** 369–379 (1996).
48. Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta Crystallogr. D Biol. Crystallogr.* **58,** 1772–1779 (2002).
49. Weeks, C. M. & Miller, R. Optimizing Shake-and-Bake for proteins. *Acta Crystallogr. D Biol. Crystallogr.* **55,** 492–500 (1999).
50. Yeates, T. D. Detecting and overcoming crystal twinning. *Methods Enzymol.* **276,** 344–358 (1997).
51. Chlenov, M. *et al.* Structure and function of lineage-specific sequence insertions in the bacterial RNA polymerase β′ subunit. *J. Mol. Biol.* **353,** 138–154 (2005).