An intermediate state of T7 RNA polymerase provides another pathway of nucleotide selection*

Zhan-Feng Wang(王展峰), Yu-Ru Liu(刘玉如), Peng-Ye Wang(王鹏业), and Ping Xie(谢平)†

Key Laboratory of Soft Matter Physics, Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Science, Beijing 100190, *China*

(Received 17 April 2017; revised manuscript received 8 July 2017; published online 10 September 2017)

Phage T7 RNA polymerase is a single-subunit transcription enzyme, transcribing template DNA to RNA. Nucleoside triphosphate (NTP) selection and translocation are two critical steps of the transcription elongation. Here, using all-atom molecular dynamics simulations, we found that between pre- and post-translocation states of T7 RNA polymerase an intermediate state exists, where the O helix C-terminal residue tyrosine 639, which plays important roles in translocation, locates between its pre- and post-translocation positions and the side chain of the next template DNA nucleotide has moved into the active site. NTP selection in this intermediate state was studied, revealing that the selection in the intermediate state can be achieved relying on the effect of Watson–Crick interaction between NTP and template DNA nucleotide, effect of stability of the components near the active site such as the nascent DNA–RNA hybrid and role of tyrosine 639. This indicates that another NTP-selection pathway can also exist besides the main pathway where NTP selection begins at the post-translocation state upon the entry of NTP.

Keywords: RNA polymerase, transcription elongation, fidelity control, molecular dynamics simulation

PACS: 02.70.Ns, 05.40.–a, 05.40.Jc **DOI:** 10.1088/1674-1056/26/10/100203

1. Introduction

Transcription of DNA into RNA, the first and highly regulated step of gene expression, is mainly carried out by RNA polymerases (RNAPs). Synthesis of RNA by RNAPs consists of three distinct stages: initiation, elongation, and termination.[1–3] Transcription elongation, the second and longest stage, is a processive process: one enzyme can catalyze thousands of nucleotide addition cycles (NACs) without dissociating from the DNA/RNA substrate.[4–6] The translocation of RNAP along DNA template is fueled by the chemical energy of nucleoside triphosphate (NTP) hydrolysis. Two main issues related to RNAPs are translocation mechanism and fidelity control or NTP selection.

Phage T7 RNAP is a prototype single-subunit molecular motor, carrying out all transcriptional functions without additional protein partners. The molecular architecture of C-terminal domain of T7 RNAP resembles the hand-like configuration of a DNA polymerase (DNAP) of DNAP I family, including fingers, palm, and thumb subdomains. The active site regions of DNAP I and T7 RNAP are particularly similar, including an α helix (termed "O-helix") that binds the incoming (d)NTP and abuts the nascent base pair.[7,8] The crystal structures of the elongation complex have formed the basis of our understanding of the NAC in T7 RNAP, showing four states in each NAC.[9–11] An NAC begins with an insertion state where the substrate NTP just inserts into the active site, accompanied by the closure of the fingers domain. After phosphoryl transfer reaction, the elongation complex (EC) transforms to pre-translocation state where the RNA transcript has been extended by one nucleotide, producing a pyrophosphate ion (PPi). After release of the product PPi, the EC transforms to post-translocation complex where the newly formed base pair has moved out of the active site along with a rotation of the fingers domain from closed to semi-open (simply called open) conformation and the C-terminus of O-helix resides in the active site. When a correct NTP comes, the EC transforms to a pre-insertion complex. When the fingers domain rotates to the closed conformation and the substrate NTP inserts into the active site, the EC transforms to the insertion state, completing the NAC.

Multi-subunit RNAPs ensure faithful nucleic acid synthesis mainly through two major strategies, substrate selection and post-incorporation proofreading.[12–21] However, for single-subunit T7 RNAP, it seems that there is no proofreading activity[22] and its fidelity control relies entirely on the selection of correct NTP during a multi-step incorporation process.[23,24] To select the correct NTP, RNAPs need to discriminate ribo-NTPs (rNTPs) from deoxyribo-NTPs (dNTPs) and choose the rNTP that is complementary to the template DNA nucleotide. It appears apparent that for both types of

*Project supported by the National Natural Science Foundation of China (Grant Nos. 11374352 and 11674381) and the National Key Research and Development Program of China (Grant No. 2016YFA0301500).

†Corresponding author. E-mail: pxie@aphy.iphy.ac.cn

© 2017 Chinese Physical Society and IOP Publishing Ltd

RNAPs, NTP selection occurs in two steps and involves an isomerization of the enzyme from a catalytically inactive to active conformation.[16,20,24] The substrate first binds to an open active site in pre-insertion state with a catalytically inactive conformation to allow initial sampling of NTPs.[11] The correct rNTP could form incipient base pair with the template DNA nucleotide and is then delivered to the insertion state along with a conformational change of the enzyme to an active closed conformation. In multi-subunit RNAPs the position where NTP binds in pre-insertion state overlaps with that in the insertion state[18,24] (see Fig. A1 in Appendix A), whereas in the single-subunit T7 RNAP the two binding sites — pre-insertion and insertion sites — are in separate positions[9,11,24] (see Fig. A2 in Appendix A). An interesting but unclear issue is if in T7 RNAP there also exists an NTP binding site with the catalytically inactive open conformation which overlaps with that with the catalytically active closed conformation, as in multi-subunit RNAPs.

It is hard to catch crystal structures in all of the states during the elongation transcription process. Molecular dynamics (MD) simulation is a powerful method to provide dynamic information and to obtain transient intermediates,[25] thus making up this deficiency. Results from MD simulations of multi-subunit RNAPs have suggested that there may be as many as five check points to ensure the fidelity of transcription in multi-subunit RNAPs.[26] Our recent MD simulations[27] of single-subunit T7 RNAP showed that during the transition from pre- to post-translocation state there existed an intermediate conformation of the enzyme where the O helix C-terminal residue tyrosine 639 (Y639), which plays a critical role in translocation and NTP selection,[28,29] was located between its pre- and post-translocation positions. Here, starting from the available structure at post-translocation state, we confirmed the existence of this intermediate conformation by using MD simulations. Moreover, we found that in the intermediate conformation the side chain of the next template DNA nucleotide could sometimes stay steadily at the active site (for convenience, we refer to this intermediate conformation as the intermediate state). We investigated NTP selectivity in the intermediate state with MD simulations, revealing that the selectivity in the intermediate state can be achieved relying on the effect of Watson–Crick interaction between the substrate NTP and template DNA nucleotide, effect of stability of the components near the active site and role of Y639. The intermediate state thus could be the starting state of another NTP-selection pathway besides the main pathway where NTP selection takes place at post-translocation state upon the entry of NTP.

2. Results

2.1. Identification of the intermediate state

The MD simulations began with post-translocation complex, where the product pyrophosphate (PPi) has been released, the newly formed hybrid base pair $i+1$ has moved out the active site, the O helix C-terminal residue Y639 resides on the active site and the template DNA nucleotide is out of the active site (see Figs. 1(a), 1(b), and Fig. A2 in Appendix A), which is different from the case of multi-subunit RNAPs where the template DNA nucleotide has moved into the active site (see Fig. A1 in Appendix A). Here, three individual trajectories were conducted for more than 100 ns each with different initial velocities in this simulation system.

The stability of the protein was checked firstly. The temporal evolutions of root mean square deviation (RMSD) values of CA atoms of the protein are shown in Fig. 1(c). It is seen that the RMSD values are always smaller than 0.7 nm and fluctuate stably around a value of about 0.45 nm for two of the three trajectories and of about 0.35 nm for the other trajectory. The root mean square fluctuation (RMSF) values of CA atoms of the protein are shown in Fig. 1(d). The RMSF values in different trajectories show good convergence. The most notable feature is that there are two regions where the fluctuation is pretty large. One is around residue Glu600 which resides on a loop of fingers domain and the other is around Ala380 which resides on the largest helix of the thumb domain. The large flexibility of these two regions is consistent with the relatively high B-factor of these parts in crystal structures. It is also consistent with the fact that these two regions are even lost in many crystal structures. The flexibility of the thumb domain has been noted previously in structural studies.[30] The flexible region of the thumb domain agrees with a previous MD simulation where the helix could break at Ala380–Val384.[31] Additionally, the observed location of the break in the thumb helix is also consistent with the observation that the thumb helix in the crystal structures of nucleic acid-free RNAP becomes disordered at this location.[32]

The stability of the newly formed base pair (3′-end of DNA–RNA hybrid or base pair $i+1$) in the post-translocation state was then checked. The distribution of RMSD values of the newly formed base pair shows small deviations from its initial position (see Fig. A3 in Appendix A). To quantitatively describe the movement of the newly formed base pair, its displacement along the direction pointing from base pair $i+1$ to base pair i was calculated after alignment with CA atoms of the palm sub-domains (residues 412–553 and 785–883). Statistical analyses of the distribution of displacements (see Fig. 2(a)) were performed with all of three trajectories. The distribution showed an asymmetrical pattern, with a slight forward movement.

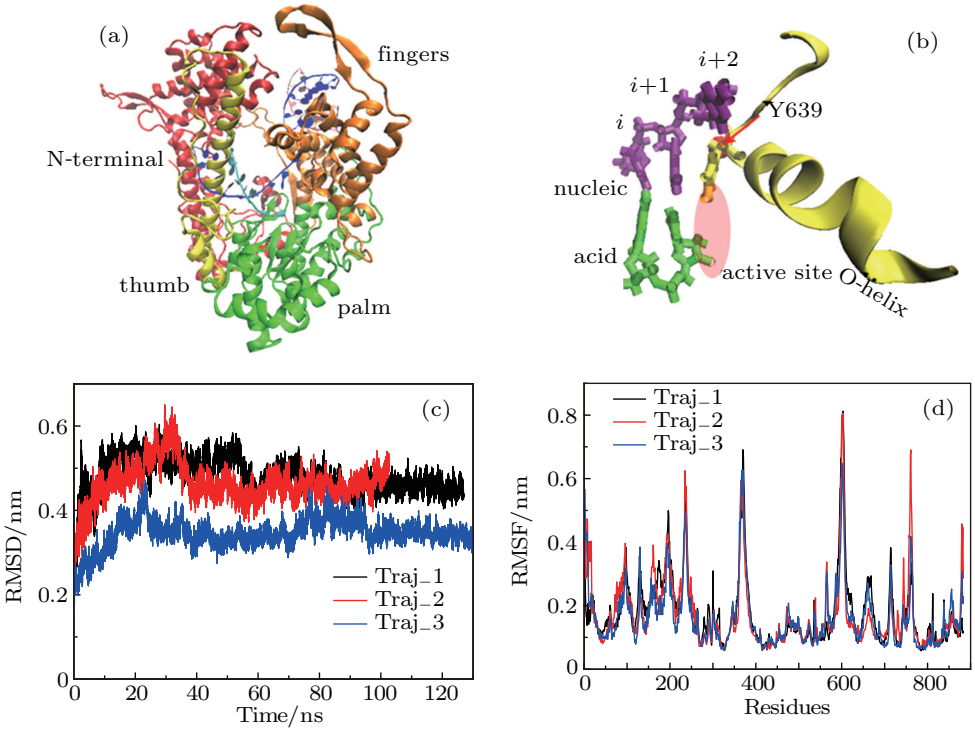


Fig. 1. (color online) MD simulations of T7 RNAP at post-translocation state. (a) The crystal structure of T7 RNAP at post-translocation state. Different domains are depicted in different colors: fingers domain in orange, palm domain in green, thumb domain in yellow, and N-terminal domain in red. (b) Components near the active site. Amino acids are drawn with cartoon method in yellow, DNA and RNA strands are drawn with bonds method in purple and green, respectively. Residue Y639 is also drawn with bonds method. (c) Time series of root mean square deviation (RMSD) values of CA atoms of the whole protein in three trajectories with different initial velocities. (d) Root mean square fluctuation (RMSF) values of different CA atoms of the protein residues.

The stability of residue Y639 which abuts the active site and the newly formed base pair $i+1$ was also checked. Interestingly, the distribution of its RMSD values shows two peaks (see Fig. A4 in Appendix A), indicating that two stable positions of Y639 may exist. To further study the two possible positions, the displacement of CA atom of Y639 relative to its initial position along the direction pointing from its position at pre-translocation conformation (PDB: 1S77) to that at post-translocation conformation (PDB: 1H38) (see Fig. A5 in Appendix A) was calculated after alignment of CA atoms of the palm domains (residues 412–553 and 785–883). The distribution of the displacements of CA atom of Y639 also shows two peaks (Fig. 2(b)), one locating at about its initial position (0 nm) which means no movement relative to the palm domain and the other locating at about -0.1 nm which means a movement relative to the palm domain in a direction pointing away the palm domain or the newly formed base pair $i+1$. When the displacement equals to 0 nm Y639 still occupies the active site, while when it equals to -0.1 nm Y639 has moved out of the active site, as shown in Fig. A6 in Appendix A. The movement of the O helix N-terminal residue arginine 627 (R627) was also analyzed. The displacement of CA atom of residue R627 relative to its initial position in the direction pointing from its position at pre-translocation conformation (PDB: 1S77) to that at post-translocation conformation (PDB: 1H38) (see Fig. A5 in Appendix A) was also calculated after alignment of CA atoms of the palm domains (residues 412–553 and 785–883). The displacements of CA atom of R627 have a rather broad distribution (see Fig. A7 in Appendix A). A two-dimensional (2D) sampling map of the displacements of CA atoms of Y639 and R627 was made based on above results (Fig. 2(c)). The map shows clearly two minimum positions. Besides its initial position at $x=0$ and $y=0$, there is another stable position with $x=-0.1$ nm and $y=-0.2$ nm, which is consistent with our previous results showing an intermediate conformation between pre- and post-translocation complexes.[27] A large flexible area around its original position indicates a loose couple between the two terminuses of O-helix. However, for the intermediate state, the movement of C-terminus of O-helix in one direction often couples with the movement of the other terminus of O-helix in the opposite direction.

Careful observations in Visual Molecular Dynamics (VMD) show that in two of the three trajectories there exists an interesting situation in the intermediate conformation where the side chain of the next template DNA nucleotide $i+2$ (denoted by TN $i+2$) has moved into the active site, as shown in Fig. 2(d). This is understandable because in the post-translocation state the side chain of Y639 occupies the active site (see, e.g., Fig. 1(b)), occluding sterically the side chain of TN $i+2$ to move into the active site, and the movement of Y639 away from the active site in the intermediate conformation makes room for the side chain of TN $i+2$ to move into.

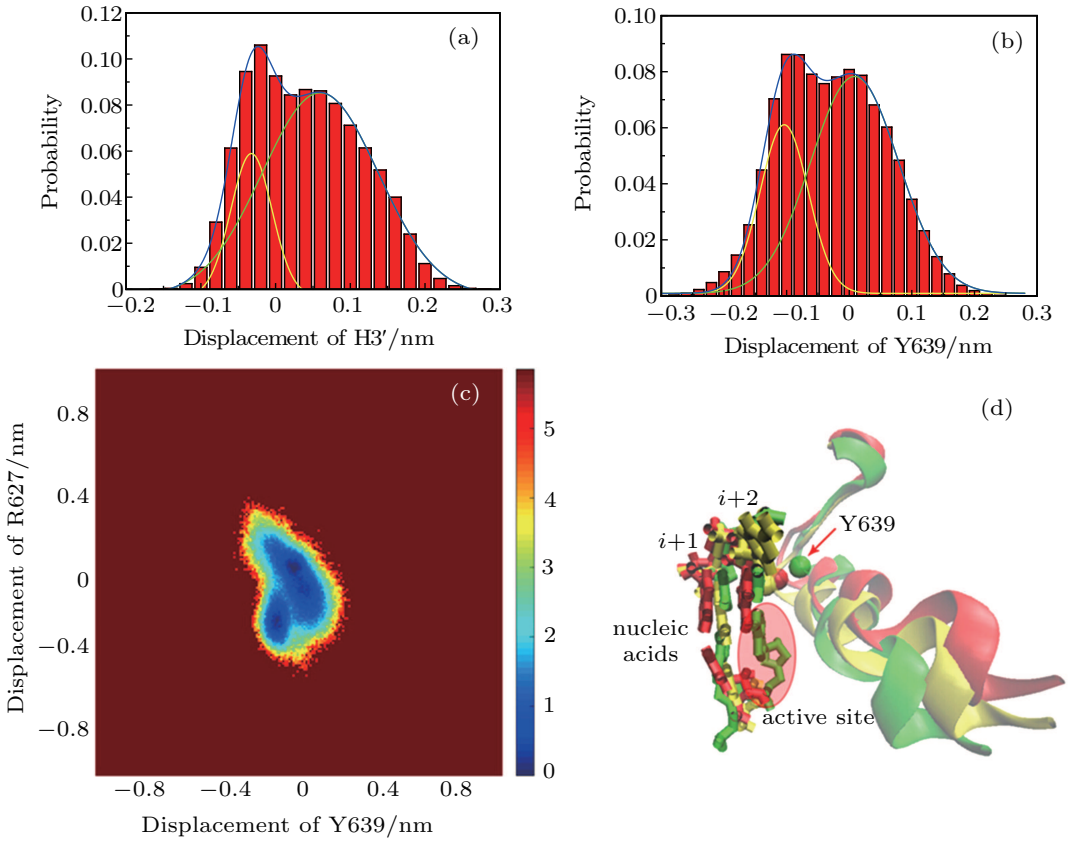


Fig. 2. (color online) The intermediate conformation derived from the post-translocation state. (a) Distribution of the displacement of 3′-end of DNA–RNA hybrid (H3′). Yellow and green lines represent Gaussian fits; the blue line represents the sum of two Gaussians. (b) Distribution of displacements of CA atom of residue Y639. Yellow and green lines represent Gaussian fits; the blue line represents the sum of two Gaussians. (c) A 2D sampling map along the displacement of CA atom of Y639 and the displacement of CA atom of R627. Color-coding represents relative free energies ($-\ln P$) as indicated with color bar. (d) A snapshot of the intermediate state where the side chain of the template DNA nucleotide has moved into the active site. Components near the active site of pre- and post-translocation conformations are shown in green and yellow, respectively. The red one is the snapshot derived from our MD trajectories.

We calculated the distance between the side chain of TN $i+2$ and 3′-end of DNA–RNA hybrid (H3′) and that between the residue Y639 and 3′-end of DNA–RNA hybrid (H3′) in all of the three trajectories, with the results being shown in Figs. 3(a)–3(c). When the distance between the side chain of TN $i+2$ and the H3′ is larger than 0.7 nm, we consider that the TN is out of the active site and is in the pre-insertion site, whereas when the distance is close to 0.5 nm, we consider that the TN has moved into the active site and is in the insertion site. When the distance between the residue Y639 and 3′-end of DNA–RNA hybrid is larger than 0.6 nm, there is enough room for the accommodation of TN $i+2$ at the active site. In the first trajectory, the distance between the residue Y639 and 3′-end of DNA–RNA hybrid (red line in Fig. 3(a)) is kept small stably, and thus there is not enough space for TN $i+2$ to move into the active site. Consequently, the side chain of TN $i+2$ always stays out of the active site, and the distance between the 3′-end of DNA–RNA hybrid and side chain of TN $i+2$ (black line in Fig. 3(a)) even becomes larger at about 5 ns and then maintains at about 1.0 nm. In the second trajectory, the distance between the residue Y639 and 3′-end of DNA–RNA hybrid (red line in Fig. 3(b)) becomes large at the start, making room for the accommodation of TN $i+2$, and the side chain of TN $i+2$ inserts into the active site at about 2 ns, but moves out at about 10 ns and does not come back (black line in Fig. 3(b)). For the third trajectory, the distance between the residue Y639 and 3′-end of DNA–RNA hybrid (red line in Fig. 3(c)) becomes large at about 45 ns. The side chain of TN $i+2$ squeezes into the active site at about 50 ns and then stays there stably (black line in Fig. 3(c)). We call this intermediate conformation with the side chain of TN $i+2$ squeezed into the active site as the intermediate state.

As there is the stacking interaction between residue Y639 and H3′, squeezing of the side chain of TN $i+2$ into the active site would destroy this interaction, thus increasing the free energy of the intermediate state. On the other hand, we also noted that after squeezing into the active site the side chain of TN $i+2$ may form stacking interaction with H3′ to compensate for the free energy increment. The calculated results of the time series of the non-bond interaction between residue Y639 and H3′ (red line) and that between H3′ and the side chain of TN $i+2$ (black line) in trajectory 3 are shown in Fig. 3(d). At about 45 ns when the distances between residue Y639 and H3′ get large, the interaction between residue Y639 and H3′ decreases from about 7.5 kcal/mol to about 1 kcal/mol, while at about 50 ns when TN $i+2$ inserts into the active site, the interaction between H3′ and

the side chain of TN $i+2$ increases from about 1 kcal/mol to about 7.5 kcal/mol. As a result, the increment in the later interaction compensates for the decrement in the former interaction.

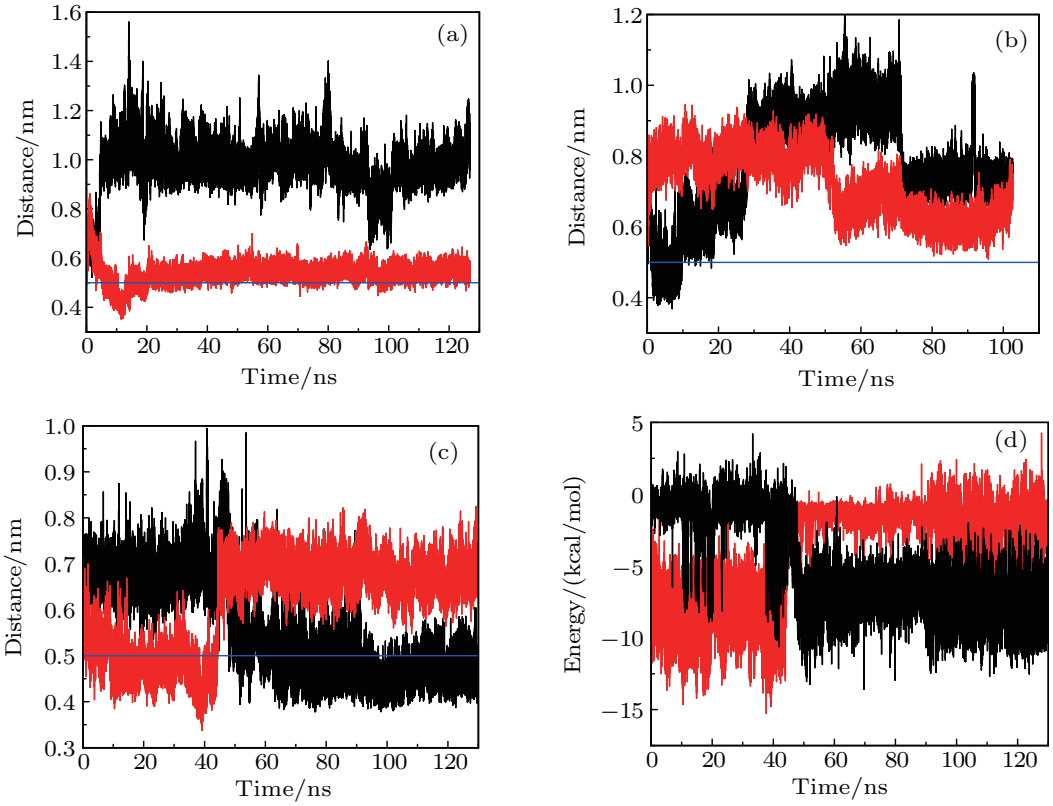


Fig. 3. (color online) Stability of the intermediate state. (a)–(c) Time series of the distance between the side chain of TN $i+2$ and the 3′-end of DNA–RNA hybrid (black lines) and that between the residue Y639 and the 3′-end of DNA–RNA hybrid (red lines) in the three trajectories. (d) Time series of the non-bond interaction energy between the base of the template DNA nucleotide and the 3′-end of DNA–RNA hybrid (black line) and the non-bond interaction energy between the 3′-end of DNA–RNA hybrid and residue Y639 (red line).

2.2. NTP selection in the intermediate state

As in the intermediate state (Fig. 2(d)) the side chain of TN $i+2$ is located at the active site, it is expected that NTP selection can also take place in the intermediate state, similar to that in DNAPs[33] and multi-subunit RNAPs.[25,26] To check this, in this section we studied the selection of different NTPs in the intermediate state by inserting different rNTPs and complimentary dNTP into the active site.

The template DNA base to pair with the incoming NTPs is a cytosine (dC10), and thus the ribo-guanine (GTP or rGTP or rG) is the complimentary or correct NTP (cNTP). By contrast, a ribo-adenine (ATP or rATP or rA), a ribo-cytosine (CTP or rCTP or rC) and a ribo-uracil (UTP or rUTP or rU) are considered as non-complimentary or incorrect NTP (ncNTP) due to their mismatched bases. A complimentary dNTP (cdNTP) deoxyribo-guanine (dGTP or dG) is also considered as an incorrect one due to its sugar deficiency. The coordinates of different NTPs or dNTP after binding were determined by superposing the derived structure with insertion complex which includes $\alpha\beta$ methylene ATP (PDB ID: 1S76). The components near the active site bound with cGTP after two times of energy minimization are shown in Fig. 4(a) (views showing the side chain of Y639 are shown in Fig. A8(a) in Appendix A). This state bound with substrate NTP can be viewed as the starting state of another NTP-selection pathway besides the main pathway where NTP selection takes place at post-translocation state upon the entry of NTP.

2.2.1. Stability of substrate NTP at the active site

For systems bound with cGTP and ncATP, three individual realizations were performed with different initial velocities. For systems bound with ncCTP, ncUTP and cdGTP, two individual trajectories were performed with different initial velocities. Each individual trajectory was performed for about 100 ns. For the system bound with cGTP, the base of substrate GTP is rather stable in all of three trajectories (see Fig. 4(b), and Fig. A8(b) in Appendix A). For the system bound with cdGTP, the base of substrate dGTP is as stable as that in the system bound with cGTP (see Fig. 4(c), and Fig. A(8c) in Appendix A). However, as shown in Fig. 4(c), a deformation of the $(i+1)$-th base pair is observed in the system bound with cdGTP. For other three systems bound with ncNTP, there is always more than one trajectory that the base of substrate NTP left the template DNA nucleotide and moved out of the active site (see Figs. 4(d)–4(f), (Figs. A8(d)–A8(f) in Appendix A).

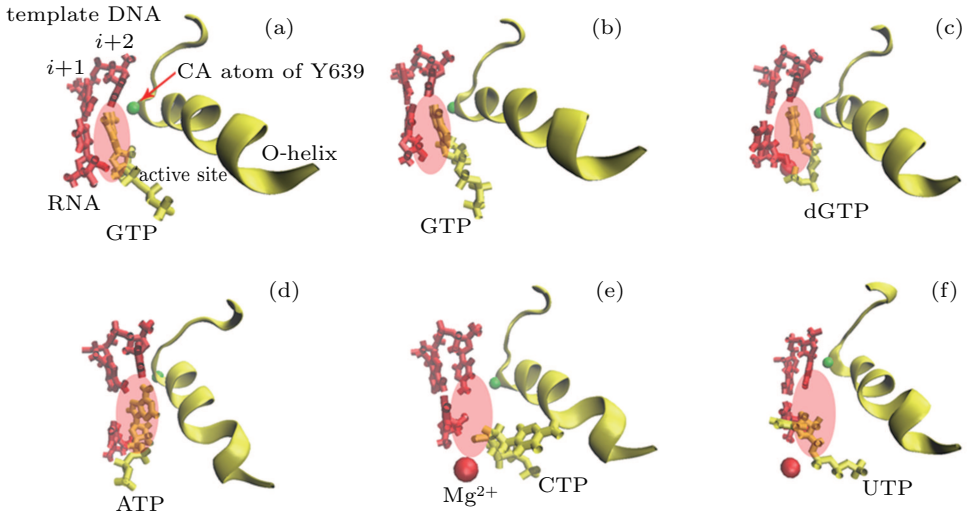


Fig. 4. (color online) Molecular views of the active site with different (d)NTPs bound. Nucleic acids and amino acids are drawn with bonds and cartoon method, respectively. DNA and RNA are shown in red; protein and substrate NTP are shown in yellow. Magnesium ions are shown in red balls and the CA atom of Y639 in green ball. (a) The conformation in GTP-binding system after energy minimization. Representative conformations in systems bound with GTP, dGTP, ATP, CTP and UTP are shown in panels (b)–(f), respectively.

To quantitatively describe the movement of the base of the incoming NTP, two reaction coordinates were used: one is the distance between the center of mass of the base of substrate NTP and that of the template DNA nucleotide dC10 (simplified as NTP/base–dC10/base distance) and the other one is the distance between the center of mass of the base of NTP and that of the 3′-end of RNA strand (simplified as NTP/base–rU8/base distance). The distributions of these two distances in five different NTP-binding systems are shown in Fig. 5(a) and Fig. 5(b), respectively. For systems bound with cNTP and cdNTP, the NTP/base–dC10/base distances are narrowly distributed around 5.5 Å, whereas for systems bound with ncNTPs, significantly larger NTP/base–dC10/base distances as well as shorter NTP/base–dC10/base distances for ncUTP are observed. Larger distances mean that the NTPs are not optimally positioned in the active site while shorter distances also show deformations near the active site (see Fig. A9 in Appendix A). Largest distances in the system bound with ncUTP reach as large as 2.2 nm. The NTP/base–rU8/base distances are narrowly distributed around 4.5 Å in the systems bound with cGTP and cdGTP. For both NTP/base–dC10/base distances and NTP/base–rU8/base distances, their distributions show that the most stably bound NTPs are GTP and dGTP. However, it is difficult to discriminate the differences between GTP and dGTP. To show the difference of the stability of NTPs intuitively, two-dimensional (2D) sampling maps of the base of NTP depicted along above two distances in different NTP-binding systems are shown in Figs. 6(a)–6(e). For the complimentary GTP and dGTP, there is a narrow distribution area, while for ncNTPs, ATP, CTP, and UTP, there is a much broader distribution area. It shows clearly that the bases of GTP and dGTP stay stably at the active site, while the bases of ncNTPs cannot stay stably at the active site but have moved out of the active site (see Fig. 4 and Fig. A8 in Appendix A). It is worth noting that for the two complimentary NTPs — GTP and dGTP — it is also difficult to discriminate them from the 2D sampling maps. Thus, the discrimination against dC:dGTP base pair is probably most difficult from this point of view.

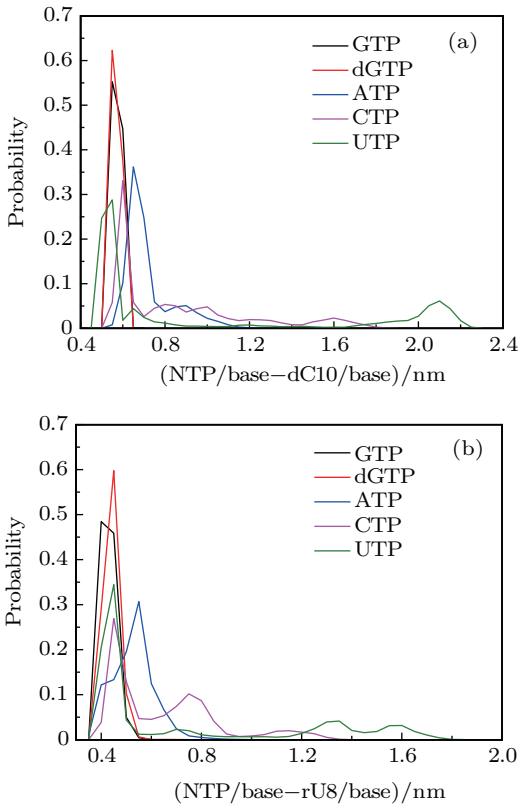


Fig. 5. (color online) Stability of the base of substrate NTP. (a) Distributions of distances between the base of NTP and that of template DNA nucleotide (NTP/base–dC10/base) in systems bound with GTP, dGTP, ATP, CTP, and UTP are shown in black, red, blue, pink, and green, respectively. (b) Distributions of distances between the base of NTP and that of RNA 3′-end nucleotide (NTP/base–rU8/base) in systems bound with GTP, dGTP, ATP, CTP, and UTP are shown in black, red, blue, pink, and green, respectively.

As shown above, the ncNTPs at the active site are more flexible than the cGTP and cdGTP. This feature is expected to be fundamentally related to the ability to form Watson–Crick (WC) hydrogen bonding interaction. For cGTP and cdGTP, they are able to form WC hydrogen bonding with the template DNA nucleotide to stabilize the base of substrate NTP, while for the ncNTPs, they are unable. It is difficult to discriminate the difference between cGTP and cdGTP, perhaps due to their similar ability to form WC interactions with the template nucleotide. Number of hydrogen bonds between the base of substrate NTP and base of template DNA nucleotide was calculated and the number distributions in different NTP-binding systems are shown in Fig. 7(a). It is seen that the numbers of hydrogen bonds between the base of template DNA nucleotide and base of substrate cGTP are almost the same as that between the base of template DNA nucleotide and base of cdGTP, and can reach as much as three. Furthermore, there is a larger probability that three hydrogen bonds could be formed. By contrast, for ncNTPs, the numbers of hydrogen bonds cannot reach three, and there is a larger probability that no hydrogen bonds can be formed between the two bases. As expected, the distributions of the numbers of hydrogen bonds between the base of substrate NTP and base of template DNA nucleotide can explain the stabilities of the base of substrate NTPs well. The stability of the base of substrate NTP along NTP/base–rU8/base distance could also rely on the base-pair stacking interaction with the 3′ terminal RNA (rU8) base. Although all of these NTP could form base pair stacking with the 3′ terminal RNA, this interaction could not favor a stable conformation by itself. The WC hydrogen bonding interaction plays a more important role in the stabilization. If there is not a stable WC interaction, the base pair stacking would easily be destroyed by thermal fluctuations. A broader distribution along NTP/base–rU8/base than along NTP/base–dC10/base in the 2D sampling maps for GTP- and dGTP-binding systems also indicates a less important role of the base pair stacking interaction in the stabilization.

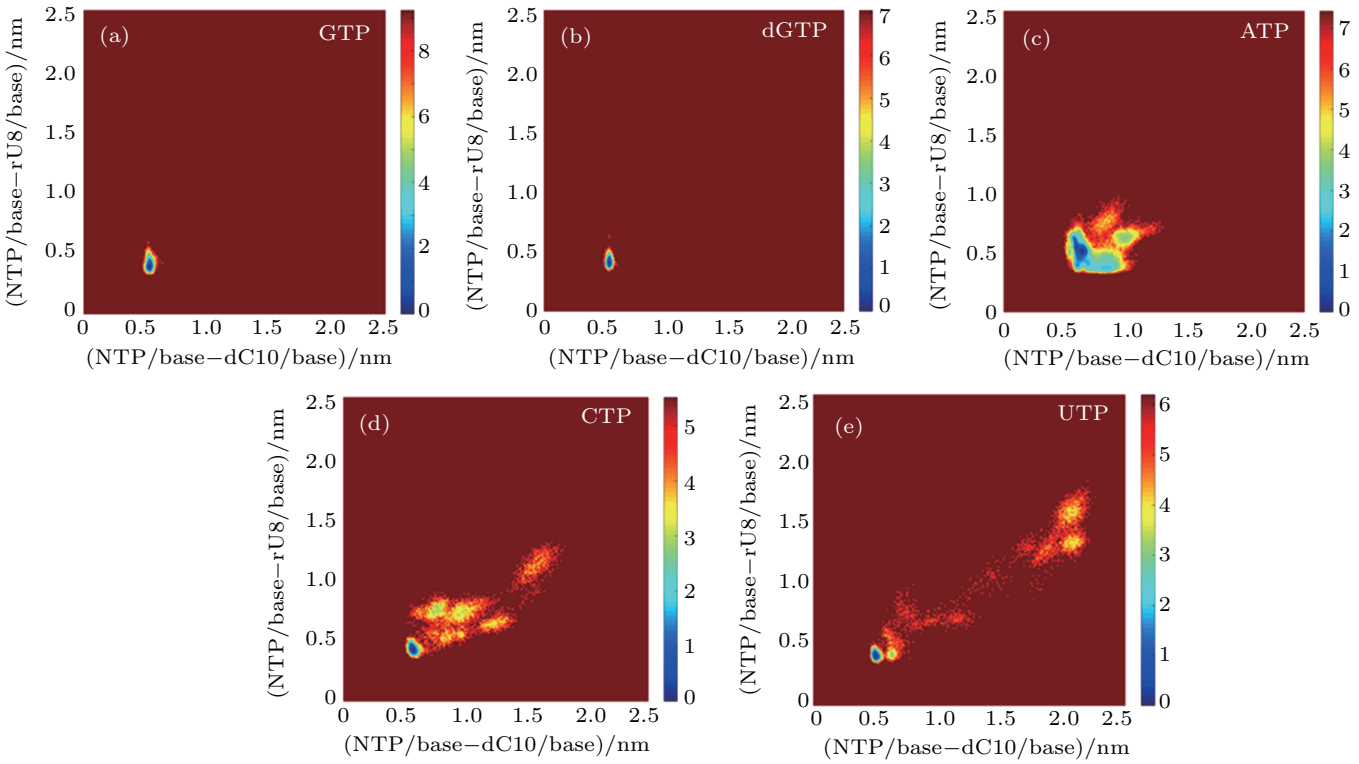


Fig. 6. (color online) Stability of the base of substrate NTP. 2D sampling maps of the base of substrate NTP along NTP/base–dC10/base distance and NTP/base–rU8/base distance in systems bound with GTP, dGTP, ATP, CTP, and UTP are shown in panels (a)–(e), respectively. Color-coding represents relative free energies ($-\ln P$) as indicated with color bar.

To further study the discrimination ability of cGTP against cdGTP, we calculated the binding affinity difference of the RNAP complex for cGTP and for cdGTP. To this end, umbrella sampling simulations along the distance between the base of substrate NTP and base of template DNA nucleotide (NTP/base–dC10/base) for dC:GTP and dC:dGTP were performed. The potential of mean force (PMF) results are shown in Fig. 7(b) (see also Fig. A10 in Appendix A). Firstly, the results for both cGTP and cdGTP show a global minimum at about 5.5 AA, which is consistent with the results of the distance distributions for NTP/base–dC10/base (Fig. 5(a)). Secondly, it shows clearly that the binding affinity for cGTP is larger than that for cdGTP. For cGTP, the binding affinity is larger than 13 kcal/mol, while for cdGTP, the binding affinity is about 10 kcal/mol, indicating that the cdGTP is more likely to dissociate from the active site. However, it is worth noting

that even for the smaller energy barrier of cdGTP it is about 10 kcal/mol, which is large enough to ensure its stability in our simulation timescale. Thus, it seems that the discrimination against cdGTP is impossible based solely on the stability of cdGTP at the active site.

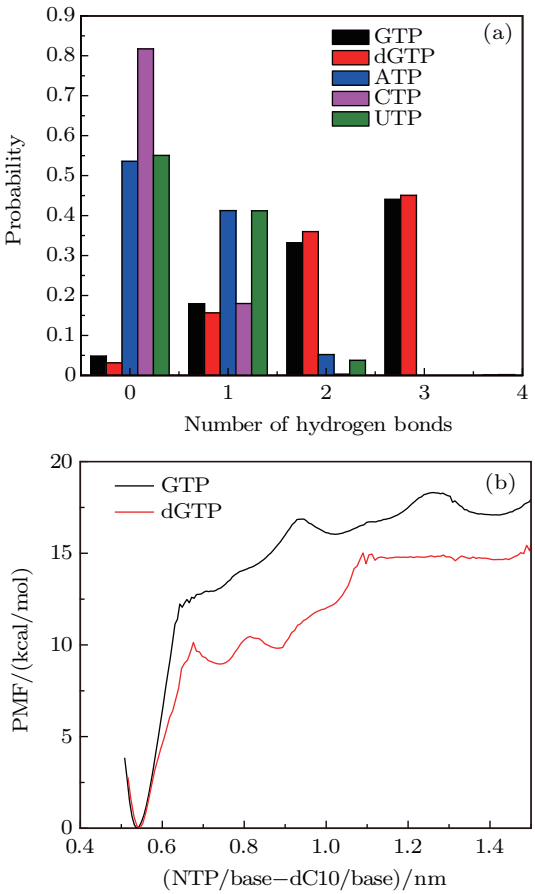


Fig. 7. (color online) Energetic features between the template DNA nucleotide and substrate NTPs. (a) Distributions of numbers of hydrogen bonds between the base of template DNA nucleotide (dC) and that of substrate GTP, dGTP, ATP, CTP, UTP are shown in black, red, blue, pink, and green, respectively. (b) Potential of mean force (PMF) along the distance between the base of template DNA nucleotide and that of substrate GTP and dGTP (NTP/base–dC10/base) are shown in black and red, respectively.

2.2.2. Effect of the stability of the components near the active site on NTP selection

Besides the stability of the base of substrate NTP, the stability of the components near the active site was also suggested to play a role to ensure the high fidelity in both DNAPs and RNAPs. The nascent $(i+1)$-th hybrid DNA–RNA base pair has been demonstrated to play a role in distinguishing cNTPs in both RNAPs and DNAPs.[26,34–36]

Thus, the distance between the two bases in the nascent $(i+1)$-th base pair (dA11/base–rU8/base) was calculated and studied statistically in different NTP-binding systems (Fig. 8). For cGTP-binding system, the distances between dA11/base and rU8/base are narrowly distributed around 6 Å. For systems bound with ncNTPs, the distributions of the distances between dA11/base and rU8/base are broader than that for cGTP-binding system. Interestingly, the distribution of the distances between dA11/base and rU8/base in cdGTP-binding system is also broader than that in cGTP-binding system. Our results thus show that mismatched NTPs (including cdNTP) can lead to conformational deformations of the components near the active site, which may perturb the coordination of magnesium ions and likely affect the ability of chemical catalysis. Thus, it is easier for the cNTP to form a stable state that can proceed to the next catalytically active state. In other words, the energy barrier for ncNTPs and cdGTP to transit to the catalytically active state is larger than that for cGTP.

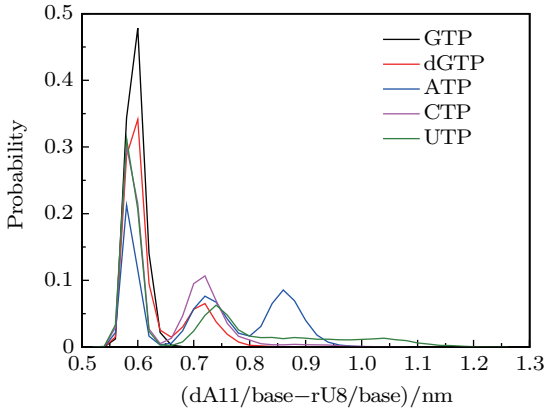


Fig. 8. (color online) Stability of the components (the $(i+1)$-th base pair) near the active site. Distributions of distances between the two bases of the $(i+1)$-th base pair (dA11/base–rU8/base) are shown in black, red, blue, pink, and green lines in systems bound with GTP, dGTP, ATP, CTP, and UTP, respectively.

2.2.3. Role of Y639 in NTP selection

Residue Y639 was suggested to play a critical role in the ability to discriminate against cdNTP by biochemical, structural and MD simulation studies.[11,29,37–39] The role of Y639 in discrimination against cdNTP was also investigated here by mutation simulations. The residue tyrosine 639 resides on the C-terminus of O-helix was mutated to phenylalanine in the mutation systems. The binding affinities of cNTP and cdNTP in the wild type and mutant type systems were studied by steered molecular dynamics (SMD) simulations. In SMD simulations, a harmonic potential was applied to the center of mass of the base of substrate NTP with a force constant of 2000 kJ/mol/nm^2 and a pulling rate of 0.1 nm/ns (see Section 4: Materials and Method for more details). The SMD simulations are actually the initial step of umbrella sampling simulations to obtain a series of configurations along the reaction coordinate. The time series of the distances between the base of the substrate NTP and that of template DNA nucleotide in wild-type systems bound with cGTP and cdGTP and in mutant systems bound with cGTP and cdGTP are shown in Fig. 9. The sharp change of the distance from about 5.5 Å to about 8 Å means the transition of the base of substrate NTP out of the active site and the breakage of the WC hydrogen bonding inter-

action with the base of template DNA nucleotide. The time series show a flat region at the beginning of the SMD simulations in all systems, indicating a stable initial conformation. Consistent with above results, the stable distance between the two bases for the substrate NTP and template DNA nucleotide is about 5.5 Å. With a harmonic potential applied on the center of mass of the base of substrate NTP and a constant pulling rate, the longer time when the distance keeps unchanged represents the larger binding affinity between the base of the substrate NTP and the template DNA nucleotide. For cGTP and cdGTP in the wild-type systems, the results show a clear difference between them: the distance transition in the wild-type system bound with GTP occurs at about 4.7 ns (black line), while the transition in the wild-type system bound with dGTP occurs at about 3.3 ns (red line). The difference of the binding affinity can be estimated to be about $1/2k(\Delta x)^2$, where k is the force constant of 2000 kJ/(mol $\cdot$ nm^2) and Δx is the distance difference equal to $r\Delta t$, with r being the pulling rate of 0.1 nm/ns and $\Delta t = 4.7$ ns–3.3 ns$=$ 1.4 ns. The binding affinity difference was estimated to be about 4.7 kcal/mol, which is quantitatively consistent with the PMF result obtained above. Overall, the SMD results also show a larger binding affinity for cGTP than cdGTP in the wild type systems. When residue tyrosine was mutated to phenylalanine, the time when the distance transition occurs decreases for both GTP and dGTP (blue and pink lines, respectively). Besides, the difference of the transition time between GTP and dGTP in the mutant systems also decreases, with the transition time of the mutant GTP-binding system (blue line) being even slightly smaller than that of the mutant dGTP-binding system (pink line), indicating that the difference of the binding affinity decreases and thus a low efficiency of discriminating against cdNTP. These are consistent with previous biochemical studies on T7 RNAP, which showed that the mutant Y639F worked simultaneously as an RNAP and as a DNAP, losing its ability to differentiate cNTP and cdNTP.[37]

For cGTP-binding systems, the transition time decreases dramatically from 4.7 ns in the wild type to 2.8 ns in the mutant type. The only difference between residues tyrosine and phenylalanine is that the former has a hydroxyl group (–OH) at the end of the side chain, whereas the latter has an instead hydrogen atom (–H) at the corresponding position. It seems that the hydroxyl group has an interaction with substrate, whereas the hydrogen atom does not have the interaction. However, for the cdGTP-binding systems, the transition time changes slightly from 3.3 ns in the wild type to 3.1 ns in the mutant type. The main difference between GTP and dGTP is that it is a hydroxyl group that attacks the 2′ carbon of the ribose in GTP, whereas the hydroxyl group is replaced by a hydrogen atom in dGTP. It seems that when the hydroxyl group of GTP is replaced by a hydrogen atom, whether there is a hydroxyl group or there is a hydrogen atom at the end of the side chain of residue 639, the interaction disappears. It is noted that these transition times of dGTP in the wild-type and mutant-type systems (3.3 ns and 3.1 ns, respectively) are similar to that of GTP-binding mutant system (2 8 ns). Thus, it seems that when both residue 639 and substrate NTP have a hydroxyl group at the corresponding position, the interaction between them is strong, otherwise it is weak.

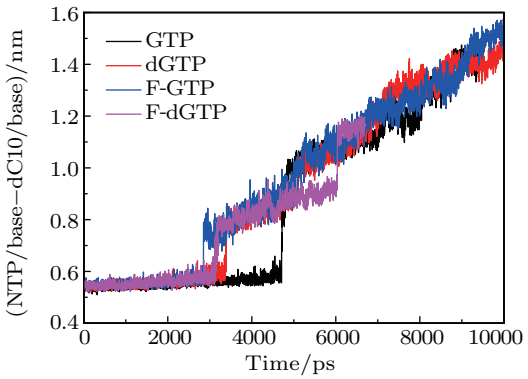


Fig. 9. (color online) Role of Y639. Time series of the NTP/base–dC10/base distance in four different SMD simulation trajectories. Time series in wide-type systems bound with GTP and dGTP are shown in black and red lines, respectively. Time series in the Y to F mutant systems bound with GTP and dGTP are shown in blue and pink, respectively.

3. Discussion

Based on the crystal structural data, it has been proposed that during transcription elongation stage of single-subunit T7 RNAP an NAC consists mainly of four states: pre-translocation, post-translocation, pre-insertion and insertion states (see Fig. 10). The translocation process occurs during the transition from pre-translocation to post-translocation state, while the NTP-selection process occurs during transition from pre-insertion to insertion to pre-translocation state.

Our all-atom MD simulations showed that between the pre- and post-translocation conformations an intermediate conformation also exists, where the O helix C-terminal residue Y639 is located between its positions at pre- and post-translocation conformations, with the intermediate position being about 0.1 nm away from that in the post-translocation conformation (Fig. 2(b) and Fig. 2(d)). The existence of this intermediate conformation is also consistent with the inference from the MD simulations by Duan *et al.*[29] showing that Y639 is stabilized marginally by the 3′-end base pair of DNA–RNA hybrid at post-translocation state. Furthermore, we found that in the intermediate conformation the side chain of TN $i+2$ can sometimes squeeze into the active site and then stay there steadily. This interesting intermediate state with the side chain of TN $i+2$ locating at the active site is much similar to the post-translocation state of DNAP in the DNA replication cycle.[25,26] Thus, similar to the case in DNAP where the post-translocation state can make dNTP selection, we showed here

that in T7 RNAP the intermediate state can also make section of cNTP against ncNTPs by relying on the Watson–Crick interaction and the effect of the stability of the components near the active site. Furthermore, we showed that besides the effect of the stability of the components near the active site, residue Y639 also plays an important role in discriminating an NTP from dNTP in the intermediate state.[37] These results thus suggest that there is another NTP-selection pathway starting from the intermediate state (State IMT1) upon the entry of substrate NTP (State IMT2) besides the main pathway starting from the post-translocation state upon the entry of substrate NTP (pre-insertion state), as proposed in Fig. 10. After state ITM2 transits to insertion state, the O helix becomes closed, where the interaction between the bound NTP and closed O helix could make a further contribution to the selection of the bound NTP, which is similar to the case that after pre-insertion state transits to insertion state. In other words, it can be said that the NTP selection in state ITM2 in the new pathway proposed here is equivalent to that in pre-insertion state in the main pathway proposed before.

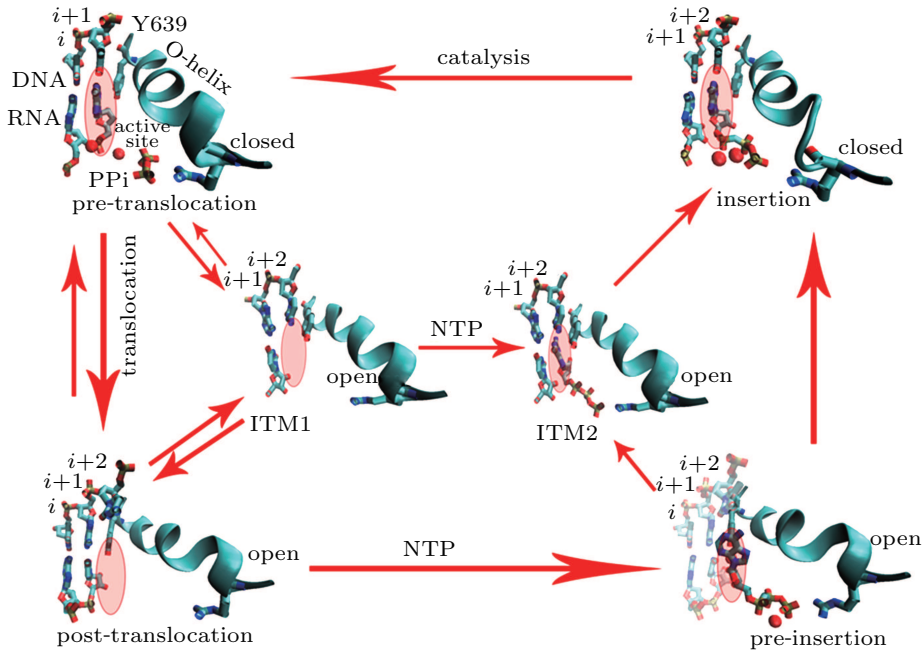


Fig. 10. (color online) A revised pathway of nucleotide addition cycle in T7 RNAP transcription elongation.

In the new pathway, ITM2 state arises from ITM1 state after a substrate NTP came to the active site. In the traditional pathway, during the transition from pre-insertion state to insertion state, the newly formed base pair between NTP and template DNA nucleotide needs to insert into the active site and the fingers domain (O-helix) becomes closed to proceed with catalysis. Whether these two processes occur simultaneously or not still need to be investigated. If they are not simultaneous, ITM2 state is also probably an intermediate state during the transition of pre-insertion to insertion state, where the newly formed base pair has inserted into the active site while the fingers domain (O-helix) is not closed. In ITM2 state, when the fingers domain becomes closed the complex transforms to the insertion conformation.

NTP selection in the traditional pathway was studied recently with MD simulations.[29] The difference between our study and that in Ref. [29] is in that in our study the TN $i+2$ has already inserted into the active site before the substrate nucleotide bound to it, whereas in the study of Ref. [29] the TN $i+2$ could insert into the active only after the correct substrate nucleotide bound to it. The study in Ref. [29] showed that in the nucleotide pre-insertion state an incorrect nucleotide substantially stabilizes the residue Y639 at the active site, blocking the insertion of the TN $i+2$ and substrate NTP into the active site, whereas the correct RNA nucleotide does not affect the local Y639 stabilization and TN $i+2$ could insert into the active site.[29] With this mechanism the correct NTP can be discriminated against incorrect NTPs. By contrast, in our study the active site is occupied by the TN $i+2$ and thus the correct substrate can form WC interaction with TN $i+2$ where incorrect NTPs cannot. With this WC interaction as well as the stability of the nascent DNA–RNA hybrid the correct NTP can be discriminated against incorrect NTPs. Both simulations with mutation of Y639 to F639 in our work and those in Ref. [29] showed that the hydroxyl group of residue Y639 plays a critical role in discriminating NTP and dNTP, just as the biochemical study shows.[37] However, the two studies indicated that Y639 used different mechanisms to

discriminate cNTP against cdNTP. In Ref. [29], in the case of cdNTP residue Y639 would be stabilized at its initial position blocking the active site, whereas in the case of cNTP Y639 would be much easier to move out and allow the insertion of substrate cNTP. With this mechanism Y639 can discriminate cNTP against cdNTP. Here, our study indicated that even when the residue Y639 has moved out of the active site, it can still play a critical role in the discrimination of NTP from dNTP. The hydroxyl group at the end of the side chain of Y639 has the interaction with the hydroxyl group of the $2'$ carbon of the ribose in NTP (e.g., GTP), whereas such interaction with dNTP (e.g., dGTP) disappears. With this mechanism Y639 can still discriminate cNTP against cdNTP.

As shown elsewhere,[9] with the consideration that the NAC pathway corresponds to the transition from pre-translocation to post-translocation to pre-insertion to insertion to pre-translocation state (Fig. 10), the available single-molecule optical trapping data[40,41] on transcription rate versus external force on T7 RNAP can be fitted well by using Brownian ratchet model A of the translocation (see Appendix A (A1)) and red lines in Fig. A12 in Appendix A). In model A, before NTP binding to the post-translocation state, the enzyme can move reversibly between the pre- and post-translocation positions (the Brownian motion). The binding of NTP to the post-translocation state prevents the enzyme from moving backward to the pre-translocation state (the ratchet). As the distance between the pre- and post-translocation positions along the translocation direction is about 0.34 nm, in the calculations for (see Fig. A12 in Appendix A) (red lines) we took the translocation distance in an NAC to be $p = 0.34$ nm. Similarly, with the consideration that the NAC pathway corresponds to the transition from pre-translocation to IMT1 to IMT2 to insertion to pre-translocation state, the available single-molecule optical trapping data[40,41] can also be fitted well by using Brownian ratchet model B of the translocation (see Appendix A (A2)) and black lines in Fig. A12 in Appendix A). In model B, before NTP binding to state IMT1, the enzyme can move reversibly between the pre-translocation state and state IMT1 (the Brownian motion). The binding of NTP to state IMT1 prevents the enzyme from moving backward to the pre-translocation state (the ratchet). As our MD simulation showed, the distance between the post-translocation state and state IMT1 is about 0.1 nm, i.e., the distance between the pre-translocation state and state IMT1 is about 0.24 nm. Thus, in the calculations for (see Fig. A12 in Appendix A) (black lines) we took the translocation distance to be $p = 0.24$ nm. In reality, both the transition from pre-translocation to post-translocation to pre-insertion to insertion to pre-translocation state and the transition from pre-translocation to IMT1 to IMT2 to insertion to pre-translocation state can occur in an NAC. Thus, the theoretical data should be in between the red and black lines, which are in good agreement with the single-molecule optical trapping data. Thus, our proposed NAC pathway (Fig. 10) based on our MD simulations presented in this work is also consistent with the single-molecule data.

Finally, it is mentioned that with the high-resolution cryo-electron microscopy technique our simulated intermediate states (ITM1 and ITM2 in Fig. 10) could be observed, testing the new pathway of transition from ITM1 to ITM2 to insertion state. Besides, it is hoped to use the high-temporal-resolution single molecule fluorescence resonance energy transfer (smFRET) to test the new pathway, as described as follows. As it is noted, during the transition from ITM2 to insertion state (Fig. 10) the distance between the base of the newly added GTP and residue 780 in RNAP is kept unchanged, whereas during the transition from pre-insertion to insertion state (Fig. 10) the distance between the base of the newly added GTP and residue 780 in RNAP is changed from about 1.33 nm to 0.78 nm. Thus, by labeling Cy3 to the base of GTP and Cy5 to residue 780 in RNAP, it is expected that in the new pathway of transition from ITM1 to ITM2 state, the high FRET value would not be changed. By contrast, in the traditional pathway of transition from pre-insertion to insertion state, the FRET value would change from a low value to a high value. If transition from pre-insertion to ITM2 to insertion takes place, the FRET value would also change from a low value to a high value. Consequently, from the measured temporal evolution of the FRET value one can test if both traditional and new pathways coexist or only the traditional pathway exists.

4. Materials and methods

4.1. Initial structures

The crystallographic structure of single-subunit T7 RNAP (PDB ID: 1H38.[10]) at post-translocation state was used for MD simulations. All data regarding crystal structures used in these simulations were downloaded from the RCSB protein data bank. Missing residues 365–371 in the post-translocation state structure (PDB ID: 1H38) were modeled based on structure with PDB ID: 2PI4,[42] and missing residues 756–765 were modeled based on structure with PDB ID: 1S76.[9] First, we make alignments of the two structures (PDB ID: 2PI4 and PDB ID: 1S76) with the post-translocation state structure (PDB ID: 1H38), separately, and then add the location coordinates of the missing residues in the post-translocation structure (PDB ID: 1H38) according to those of the two structures (PDB ID: 2PI4 and PDB ID: 1S76). The parameters used for GTP, dGTP, ATP, CTP, and UTP were built

by combining the existing parameter for ATP[43] and nucleic acids.[26,29] The coordinates of different (d)NTPs in ITM2 state were obtained by superposing the intermediate structure with insertion complex (PDB ID: 1S76).

4.2. MD simulations

We followed a similar procedure to that done in our previous works.[27,44–46] to do the simulations. All of the MD simulations were performed using GROMACS 4.6[47] and the Amber03 Force Field.[48] To avoid the edge effect, the distance between the complex and the boundary of the box was set to be at least 1 nm. The complexes were fully solvated in cuboid boxes with explicit TIP3P water,[49] in which there were 5-mM magnesium ions and neutralizing counter-ions. All chemical bonds were constrained using the LINCS algorithm.[50] The cutoff for van der Waals interaction and short-range electrostatics interaction was 1 nm. The Particle Mesh Ewald (PME) algorithm[51] was used for calculations of long-range electrostatics. Velocity-rescaling temperature coupling[52] and Berendsen pressure coupling[53] were used. The reference coordinates were scaled using the scaling matrix of the pressure coupling. Energy minimization was performed by using the steepest descent method for 5×10^4 steps. The system was then equilibrated for 100 ps at 300 K and 1 bar (1 bar = 10^5 Pa) pressure for 100 ps successively. The final configurations from equilibration were then used for production simulations. The size of each step in the simulation was 2 fs and the neighbor list was updated every 5 steps. Position restraints on heavy atoms of the protein and DNA and RNA chains were imposed at the beginning of the simulation. After the constrained simulation, unconstraint MD simulations were performed for analysis.

Additional umbrella sampling simulations[27,54,55] were carried out to determine the free energy change along the distance between the base of substrate NTP and the base of template DNA nucleotide. A steered MD (SMD) simulation was performed to obtain the initial structures for umbrella sampling simulations with a spring constant of 2000 $\mathrm{kJ\cdot mol^{-1}\cdot nm^{-2}}$ and a pulling rate of 1×10^{-4} $\mathrm{nm\cdot ps^{-1}}$, wherein the palm domain and the DNA and RNA strands were restrained. The spacing of the umbrella sampling simulation windows was less than 0.2 nm and the simulation period in each window lasted 20 ns, so that the histograms of the configurations would sufficiently overlap with their neighboring windows and the simulation time was long enough to ensure the convergence (see Fig. S10 in Appendix A). For GTP- and dGTP-bound systems, there are 8 and 9 windows, respectively. The last 19 ns samplings were used for calculating the potential of mean force (PMF) through the weighted histogram analysis method (WHAM).[56]

Appendix A: Fitting to the single-molecule optical trapping data with Brownian ratchet models A and B

A1: Model A

In Model A, the NAC pathway corresponds to the transition from pre-translocation to post-translocation to pre-insertion to insertion to pre-translocation state (Fig. 10). Before NTP binding to the active site of the enzyme, driven by thermal noise the enzyme can move reversibly between the pre-translocation position, which is represented by the n-th site, and the post-translocation site, which is represented by the $(n+1)$-th site (the Brownian motion). The interacting potential between the enzyme and the DNA–RNA is schematically shown in Fig. A11. At the n-th site, the potential has an energy depth E_n while at the $(n+1)$-th site the potential has a energy depth E_{n+1}. At the n-th site (the pre-translocation state) NTP cannot bind to the active site, but at the $(n+1)$-th site (the post-translocation state) NTP can bind to the active site. The binding of NTP to the post-translocation state prevents the enzyme from moving backward to the n-th site (the ratchet).

As done in the single-molecule experiment using optical trapping,[57] consider a backward external force, F, acting on the enzyme. From the potential (Fig. A11), the ratio of the mean time for the enzyme to move from the n-th site to the $(n+1)$-th site, $T_{n\to(n+1)}$, to the mean time for the enzyme to move from the $(n+1)$-th site to the n-th site, $T_{(n+1)\to n}$, approximately has the following form[58,59]

$$\frac{T_{n\to(n+1)}}{T_{(n+1)\to n}}\approx\exp\left(\frac{E_n-E_{n+1}}{k_{\mathrm{B}}T}\right)\exp\left(\frac{Fp}{k_{\mathrm{B}}T}\right), \tag{A1}$$

where p is the distance between the n-th site and the $(n+1)$-th site, i.e., the distance of Y639 between the pre-translocation and the post-translocation state along the translocation direction. As the ratio of the mean time, T_n, for the enzyme to stay at the nth site to the mean time, T_{n+1}, to stay at the $(n+1)$-th site satisfies $T_n/T_{n+1}=T_{n\to(n+1)}/T_{(n+1)\to n}$, from Eq. (A1) we have the following:

$$\frac{T_n}{T_{n+1}}\approx\exp\left(\frac{E_n-E_{n+1}}{k_{\mathrm{B}}T}\right)\exp\left(\frac{Fp}{k_{\mathrm{B}}T}\right). \tag{A2}$$

Based on the ratchet mechanism, only when the active site is positioned at the $(n+1)$-th site, i.e., only during the time period T_{n+1}, can NTP bind to the active site. Thus, based on Eq. (A2), the ratio of the NTP-binding rate, $k_{\mathrm{b}}(F)$, under the external force F to that, $k_{\mathrm{b}}(0)$, under no external force has the following form[58,59]

$$\frac{k_{\mathrm{b}}(F)}{k_{\mathrm{b}}(0)}=\frac{T_{n+1}(F)/[T_n(F)+T_{n+1}(F)]}{T_{n+1}(0)/[T_n(0)+T_{n+1}(0)]}, \tag{A3}$$

where $T_n(F)$ and $T_{n+1}(F)$ represent T_n and T_{n+1} under the external force F, respectively, and $T_n(0)$ and $T_{n+1}(0)$ represent

T_n and T_{n+1} under $F = 0$, respectively. With Eq. (A2), equation (A3) can be rewritten as follows:

$$k_{\mathrm{b}}(F) = \frac{\exp\left(\dfrac{\Delta E}{k_{\mathrm{B}}T}\right)+1}{\exp\left(\dfrac{\Delta E}{k_{\mathrm{B}}T}\right)\exp\left(\dfrac{Fp}{k_{\mathrm{B}}T}\right)+1}k_{\mathrm{b}}(0), \tag{A4}$$

where $\Delta E = E_n - E_{n+1}$.

Considering that the residues of the enzyme acted by the external force F are far away from the active site, F has a negligible effect on the polymerase activity of the active site. Then, the transcription velocity has the Michaelis–Menten form:

$$k = p\frac{k_{\mathrm{c}}[\mathrm{NTP}]}{K_{\mathrm{m}}^{(F)}+[\mathrm{NTP}]}, \tag{A5}$$

where k_{c} is independent of F, and $K_{\mathrm{m}}^{(F)} = k_{\mathrm{c}}/k_{\mathrm{b}}(F)$ is the Michaelis–Menten constant, which is dependent on F. With Eq. (A4), equation (A5) is rewritten as follows:

$$k = p\frac{k_{\mathrm{c}}[\mathrm{dNTP}]}{[\mathrm{dNTP}]+K_{\mathrm{m}}^{(0)}\left[\exp\left(\dfrac{\Delta E}{k_{\mathrm{B}}T}\right)\exp\left(\dfrac{Fp}{k_{\mathrm{B}}T}\right)+1\right]/\left[\exp\left(\dfrac{\Delta E}{k_{\mathrm{B}}T}\right)+1\right]}, \tag{A6}$$

where $K_{\mathrm{m}}^{(0)} = k_{\mathrm{c}}/k_{\mathrm{b}}^{(0)}$ is the Michaelis–Menten constant under no external force, i.e., $F = 0$.

As structural data showed, the distance between the pre-translocation and the post-translocation positions along the translocation direction is about 0.34 nm. Thus, in the calculations we took $p = 0.34$ nm. Using Eq. (A6), the experimental data[57] on the transcription velocity of T7 RNA polymerase as a function of NTP concentration under different values of the external force F can be fitted well with a fixed set of parameters $k_{\mathrm{c}} = 130\ \mathrm{s}^{-1}$, $K_{\mathrm{m}}^{(0)} = 165\ \mu\mathrm{M}$, and $\exp(\Delta E/k_{\mathrm{B}}T) = 0.1$, as shown by red lines in Fig. A12.

A2: Model B

In Model B, the NAC pathway corresponds to the transition from pre-translocation to IMT1 to IMT2 to insertion to pre-translocation state (Fig. 10). Before NTP binding to the active site of the enzyme, driven by thermal noise the enzyme can move reversibly between the pre-translocation position, which is represented by the n-th site, and state IMT1, which is still represented by the $(n+1)$-th site (the Brownian motion). The interacting potential between the enzyme and the DNA–RNA is still schematically shown in Fig. A11. At the n-th site, the potential has an energy depth E_n while at the $(n+1)$-th site the potential has a energy depth E_{n+1}. At the n-th site (the pre-translocation state) NTP cannot bind to the active site, but at the $(n+1)$-th site (state IMT1) NTP can bind to the active site. The binding of NTP to state IMT1 prevents the enzyme from moving backward to the n-th site (the ratchet).

The transcription velocity versus [NTP] and F can still described by Eq. (A6), but with p being the distance of Y639 between the pre-translocation state and state IMT1 along the translocation direction. As our MD simulations showed, the distance between the post-translocation state and state IMT1 is about 0.1 nm, i.e., the distance between the pre-translocation state and state IMT1 is about 0.24 nm. Thus, in the calculations we took $p = 0.24$ nm. Still with $k_{\mathrm{c}} = 130\ \mathrm{s}^{-1}$, $K_{\mathrm{m}}^{(0)} = 165\ \mu\mathrm{M}$, and $\exp(\Delta E/k_{\mathrm{B}}T) = 0.1$, with Eq. (A6) the calculated data of the transcription velocity as a function of

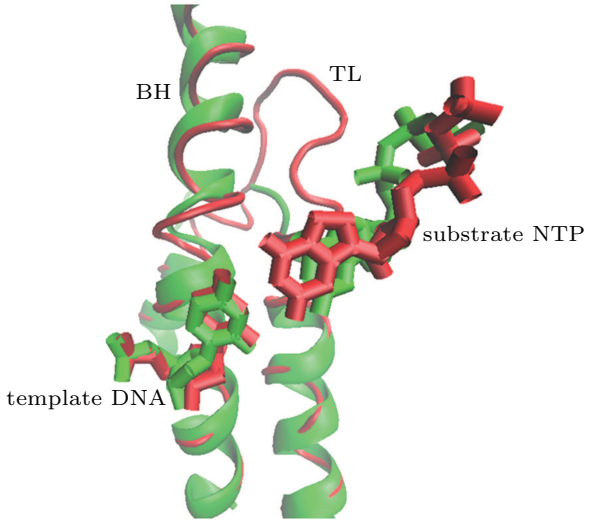


Fig. A1. (color online) Structural alignment between the pre-insertion (green) and insertion (red) structures for multi-subunit RNAP. The positions of the template DNA nucleotide in pre-insertion and insertion states are overlapped, implying that the template DNA nucleotide has moved into the active site even in the pre-insertion state.

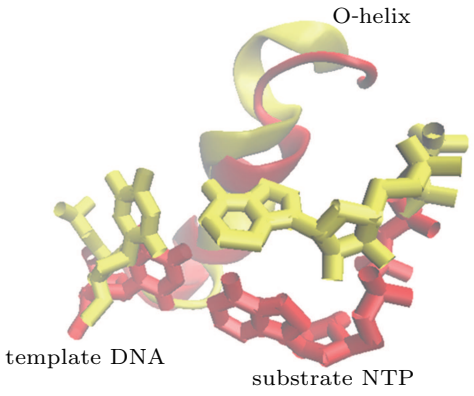


Fig. A2. (color online) Structural alignment between the pre-insertion (yellow) and insertion (red) structures for T7 RNAP. The positions of the template DNA nucleotide in pre-insertion and insertion states are not overlapped, implying that the template DNA nucleotide has not moved into the active site in the pre-insertion state.

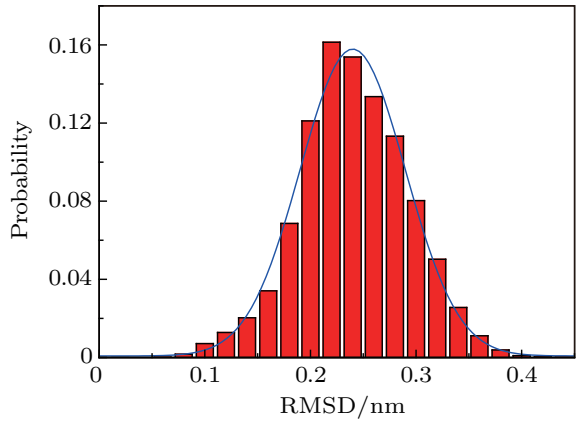


Fig. A3. (color online) Distribution of the root mean square deviations (RMSDs) of the newly formed base pair.

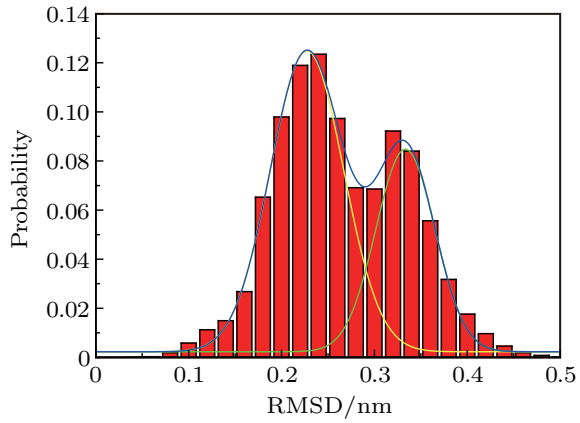


Fig. A4. (color online) Distribution of the root mean square deviations (RMSDs) of CA atom of residue Y639. Yellow and green lines represent Gaussian fits; the blue line represents the sum of two Gaussians.

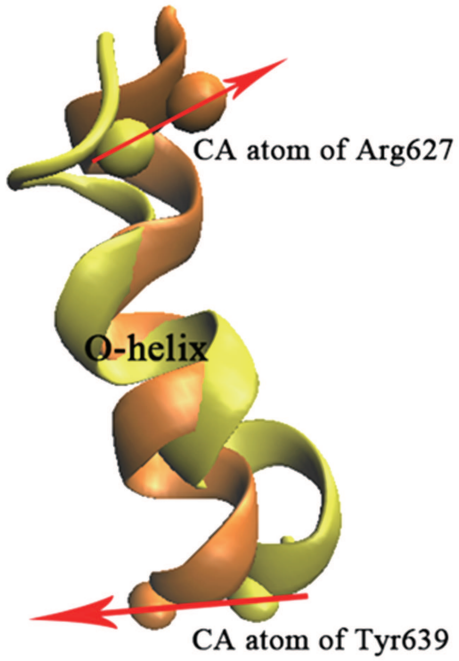


Fig. A5. (color online) Illustration of the two coordinates to describe the displacement of CA atoms of Y639 and R627. The O-helix in pre- and post-translocation conformations is shown in yellow and orange, respectively. CA atoms of Y639 and R627 are depicted as balls. The positive directions are shown with red arrows.

the NTP concentration under different values of the external force F are shown by black lines in Fig. A12. It is seen that the calculated data are still in agreement with the experimental data.

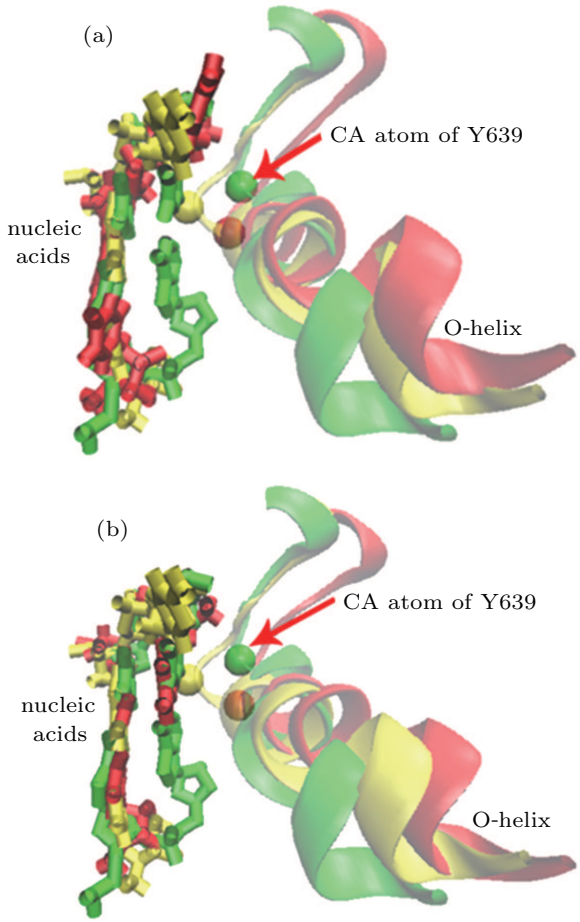


Fig. A6. (color online) The intermediate position of CA atom of Y639. Components near the active site of pre- and post-translocation conformations are shown in green and yellow, respectively. The red one is a snapshot derived from our MD trajectories. Nucleic acids are drawn with bonds method. CA atoms of Y639 in different conformations are depicted as balls. When CA atom of Y639 stays at the intermediate position, the template DNA nucleotide could be in either pre-insertion position (a) or insertion position (b).

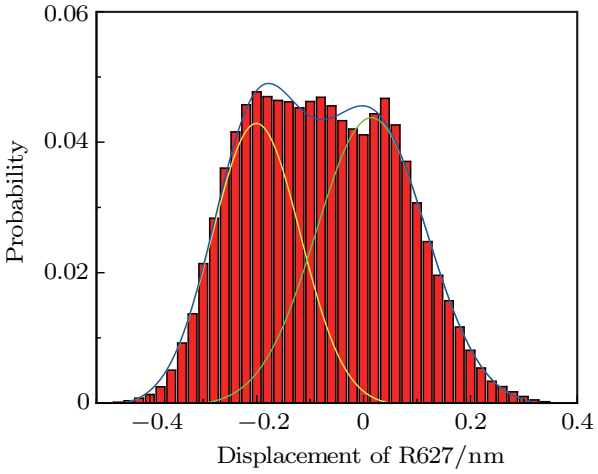


Fig. A7. (color online) Distribution of displacements of CA atom of residue R627. Yellow and green lines represent Gaussian fits; the blue line represents the sum of two Gaussians.

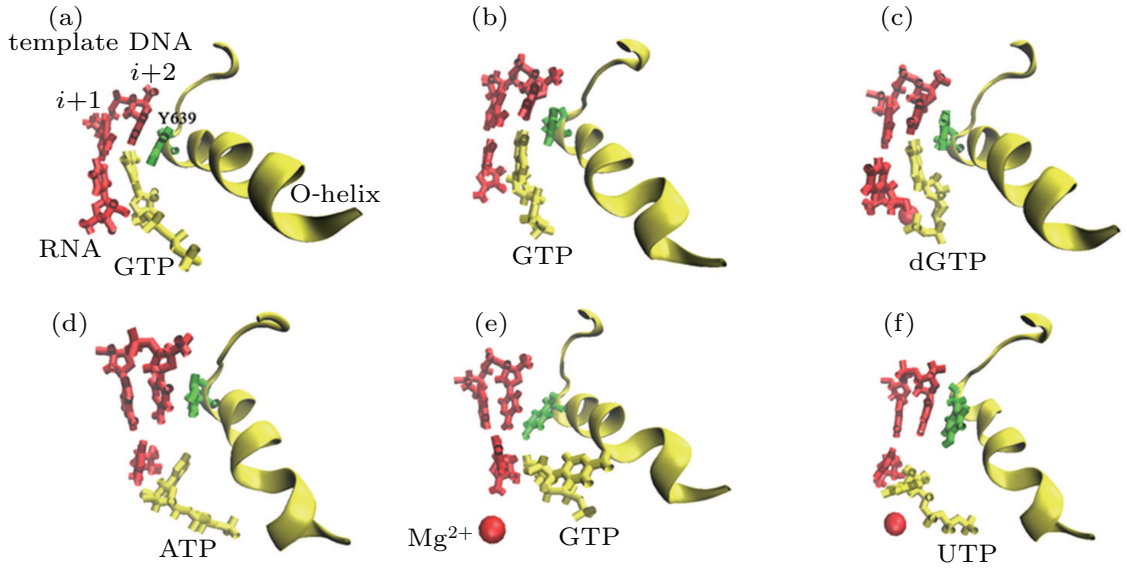


Fig. A8. (color online) Molecular views of the active site bound with different (d)NTPs and the side chain of Y639. Nucleic acids and amino acids are drawn with bonds and cartoon method, respectively. DNA and RNA are shown in red; and protein and substrate NTP are shown in yellow. Magnesium ions are shown in red balls and the CA atom of Y639 in green ball. (a) The conformation in GTP-bound system after energy minimization. Representative conformations in systems bound with GTP, dGTP, ATP, CTP, and UTP are shown in panels (b)–(f), respectively.

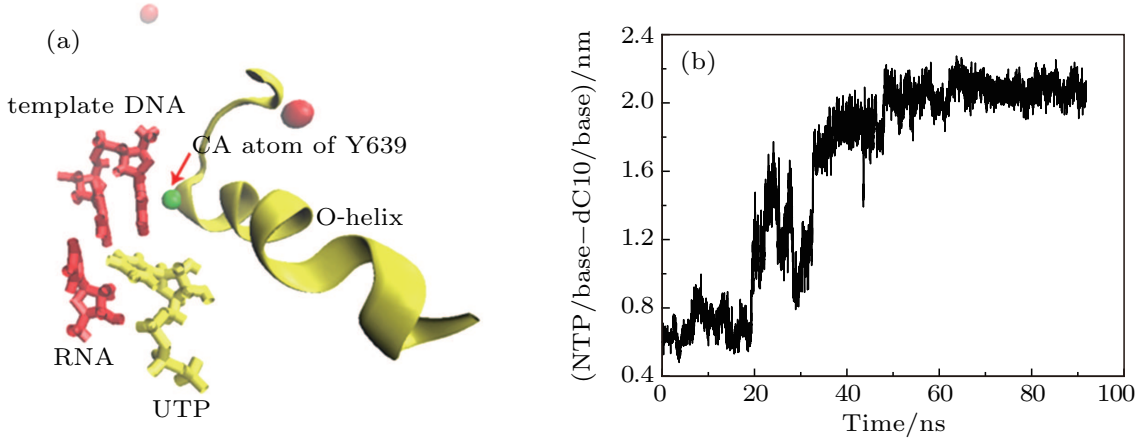


Fig. A9. (color online) Shorter distances between the base of UTP and that of template DNA nucleotide. (a) A snapshot when the shorter distances occur. Template DNA strand and product RNA strand are depicted with bonds method, while amino acids are depicted with cartoon method in yellow. Substrate UTP is depicted with bonds method in yellow. CA atom of Y639 is shown as a green ball. (b) Time series of the distance between the base of UTP and that of template DNA nucleotide in one of UTP bound trajectory.

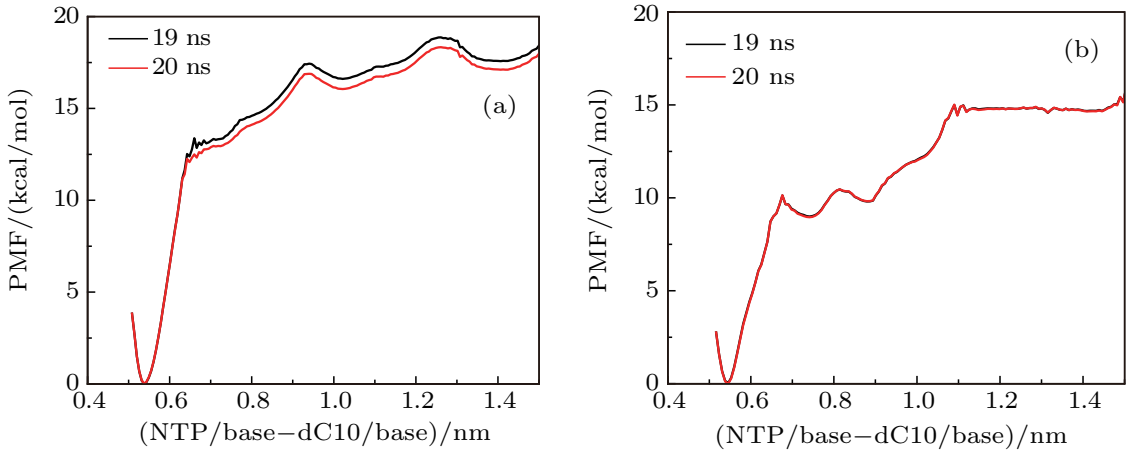


Fig. A10. (color online) Convergence analysis of the PMF results of GTP-bound system (a) and dGTP-bound system (b). Black and red lines are results analyzed with data from 1ns through 19 ns and data from 1ns through 20 ns, respectively. Results showed that the simulation time of 20 ns in each window is long enough to ensure the convergence.

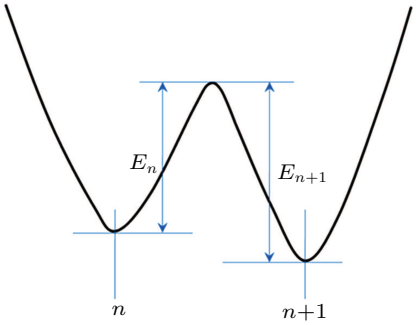


Fig. A11. (color online) Free energy landscape characterizing the transition of the elongation complex.

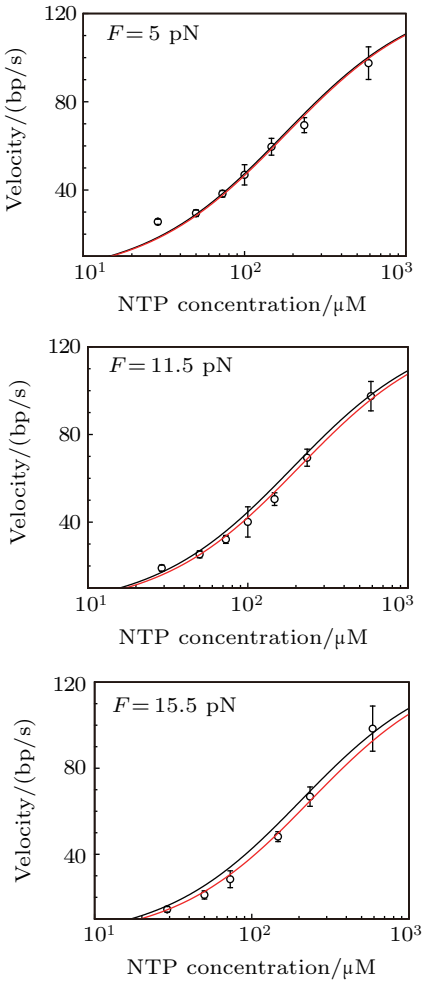


Fig. A12. (color online) Transcription velocity versus NTP concentration under different values of external force F. Circles are experimental data taken from Thomen *et al.*[57] Lines represent theoretical data calculated by using Eq. (A6) with $k_c = 130\ s^{-1}$, $K_m^{(0)} = 165\ \mu M$, and $\exp(\Delta E/k_B T) = 0.1$. Red lines are calculated with $p = 0.34$ nm, while black lines are calculated with $p = 0.24$ nm.

References

[1] Sekine S, Tagami S and Yokoyama S 2012 *Curr. Opin. Struc. Biol.* **22** 110
[2] Vannini A and Cramer P 2012 *Mol. Cell* **45** 439
[3] Svetlov V and Nudler E 2013 *Bba-Gene Regul. Mech.* **1829** 20
[4] Landick R 1997 *Cell* **88** 741
[5] Uptain S M, Kane C M and Chamberlin M J 1997 *Ann. Rev. Biochem.* **66** 117
[6] von Hippel P H 1998 *Science* **281** 660
[7] Doublie S and Ellenberger T 1998 *Curr. Opin. Struc. Biol.* **8** 704
[8] Steitz T A 1999 *J. Biol. Chem.* **274** 17395
[9] Yin Y W and Steitz T A 2004 *Cell* **116** 393
[10] Tahirov T H, Temiakov D, Anikin M, Patlan V, McAllister W T, Vassylyev D G and Yokoyama S 2002 *Nature* **420** 43
[11] Temiakov D, Patlan V, Anikin M, McAllister W T, Yokoyama S and Vassylyev D G 2004 *Cell* **116** 381
[12] Holmes S F, Santangelo T J, Cunningham C K, Roberts J W and Erie D A 2006 *J. Biol. Chem.* **281** 18677
[13] Svetlov V, Vassylyev D G and Artsimovitch I 2004 *J. Biol. Chem.* **279** 38087
[14] Wang D, Bushnell D A, Westover K D, Kaplan C D and Kornberg R D 2006 *Cell* **127** 941
[15] Alic N, Ayoub N, Landrieux E, Favry E, Baudouin-Cornu P, Riva M and Carles C 2007 *Proc. Natl. Acad. Sci. USA* **104** 10400
[16] Kireeva M L, Nedialkov Y A, Cremona G H, Purtov Y A, Lubkowska L, Malagon F, Burton Z F, Strathern J N and Kashlev M 2008 *Mol. Cell* **30** 557
[17] Sydow J F, Brueckner F, Cheung A C M, Damsma G E, Dengl S, Lehmann E, Vassylyev D and Cramer P 2009 *Mol. Cell* **34** 710
[18] Vassylyev D G, Vassylyeva M N, Zhang J W, Palangat M, Artsimovitch I, and Landick R 2007 *Nature* **448** 163
[19] Kaplan C D, Larsson K M and Kornberg R D 2008 *Mol. Cell* **30** 547
[20] Kettenberger H, Armache K J and Cramer P 2004 *Mol. Cell* **16** 955
[21] Depken M, Parrondo J M R and Grill S W 2013 *Cell Rep.* **5** 521
[22] Huang J B, Brieba L G and Sousa R 2000 *Biochemistry-Us* **39** 11571
[23] Anand V S and Patel S S 2006 *J. Biol. Chem.* **281** 35677
[24] Sydow J F and Cramer P 2009 *Curr. Opin. Struc. Biol.* **19** 732
[25] Silva D A, Weiss D R, Avila F P, Da L T, Levitt M, Wang D and Huang X H 2014 *Proc. Natl. Acad. Sci. USA* **111** 7665
[26] Wang B B, Opron K, Burton Z F, Cukier R I and Feig M 2015 *Nucleic Acids Res.* **43** 1133
[27] Wang Z F, Zhang Z Q, Fu Y B, Wang P Y and Xie P 2017 *Chin. Phys. B* **26** 030201
[28] Yu J and Oster G 2012 *Biophys. J.* **102** 532
[29] Duan B G, Wu S G, Da L T and Yu J 2014 *Biophys. J.* **107** 2130
[30] Sousa R, Rose J and Wang B C 1994 *J. Mol. Biol.* **244** 6
[31] Woo H J, Liu Y and Sousa R 2008 *Proteins* **73** 1021
[32] Jeruzalmi D and Steitz T A 1998 *Embo. J.* **17** 4101
[33] Li Y, Korolev S and Waksman G 1998 *Embo. J.* **17** 7514
[34] Kool E T 2001 *Ann. Rev. Bioph. Biom.* **30** 1
[35] Kellinger M W, Ulrich S, Chong J N, Kool E T and Wang D 2012 *J. Am. Chem. Soc.* **134** 8231
[36] Xu L, Plouffe S W, Chong J, Wengel J and Wang D 2013 *Angew Chem. Int. Ed.* **52** 12341
[37] Sousa R and Padilla R 1995 *Embo. J.* **14** 4609
[38] Yoon H and Warshel A 2016 *Proteins* **84** 1616
[39] Brieba L G and Sousa R 2000 *Biochemistry-Us* **39** 919
[40] Thomen P, Lopez P J and Heslot F 2005 *Phys. Rev. Lett.* **94**
[41] Thomen P, Lopez P J, Bockelmann U, Guillerez J, Dreyfus M and Heslot F 2008 *Biophys. J.* **95** 2423
[42] Kennedy W P, Momand J R and Yin Y W 2007 *J. Mol. Biol.* **370** 256
[43] Meagher K L, Redman L T and Carlson H A 2003 *J. Comput. Chem.* **24** 1016
[44] Duan Z W, Xie P, Li W and Wang P Y 2012 *Plos One* **7** e36071
[45] Fu Y B, Wang Z F, Wang P Y and Xie P 2016 *Sci. Rep.-Uk* **6**
[46] Wang Z F, Fu Y B, Wang P Y and Xie P 2017 *Proteins* **85** 614
[47] Hess B, Kutzner C, van der Spoel D and Lindahl E 2008 *J. Chem. Theory Comput.* **4** 435
[48] Duan Y, Wu C, Chowdhury S, Lee M C, Xiong G M, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J M and Kollman P 2003 *J. Comput. Chem.* **24** 1999
[49] Price D J and Brooks C L 2004 *J. Chem. Phys.* **121** 10096
[50] Hess B, Bekker H, Berendsen H J C and Fraaije J G E M 1997 *J. Comput. Chem.* **18** 1463
[51] Essmann U, Perera L, Berkowitz M L, Darden T, Lee H and Pedersen L G 1995 *J. Chem. Phys.* **103** 8577
[52] Bussi G, Donadio D and Parrinello M 2007 *J. Chem. Phys.* **126** 014101
[53] Berendsen H J C, Postma J P M, Vangunsteren W F, Dinola A and Haak J R 1984 *J. Chem. Phys.* **81** 3684
[54] Torrie G M and Valleau J P 1974 *Chem. Phys. Lett.* **28** 578
[55] Torrie G M and Valleau J P 1977 *J. Comput. Phys.* **23** 187
[56] Kumar S, Bouzida D, Swendsen R H, Kollman P A and Rosenberg J M 1992 *J. Comput. Chem.* **13** 1011
[57] Thomen P, Lopez PJ, Bockelmann U, Guillerez J, Dreyfus M, *et al.* 2008 *Biophys. J.* **95** 2423
[58] Xie P 2012 *Proteins-Structure Function and Bioinformatics* **80** 2020
[59] Xie P 2012 *J. Mol. Model.* **18** 1951