A Two-State Model for the Dynamics of the Pyrophosphate Ion Release in Bacterial RNA Polymerase

Lin-Tai Da[1], Fátima Pardo Avila[1], Dong Wang[2], Xuhui Huang[1,3,4]*

1 Department of Chemistry, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, 2 Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California, United States of America, 3 Division of Biomedical Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, 4 Center of Systems Biology and Human Health, Institute for Advance Study and School of Science, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

Abstract

The dynamics of the PP_i release during the transcription elongation of bacterial RNA polymerase and its effects on the Trigger Loop (TL) opening motion are still elusive. Here, we built a Markov State Model (MSM) from extensive all-atom molecular dynamics (MD) simulations to investigate the mechanism of the PP_i release. Our MSM has identified a simple two-state mechanism for the PP_i release instead of a more complex four-state mechanism observed in RNA polymerase II (Pol II). We observed that the PP_i release in bacterial RNA polymerase occurs at sub-microsecond timescale, which is ~3-fold faster than that in Pol II. After escaping from the active site, the $(Mg-PP_i)^{2-}$ group passes through a single elongated metastable region where several positively charged residues on the secondary channel provide favorable interactions. Surprisingly, we found that the PP_i release is not coupled with the TL unfolding but correlates tightly with the side-chain rotation of the TL residue R1239. Our work sheds light on the dynamics underlying the transcription elongation of the bacterial RNA polymerase.

Introduction

The DNA-dependent RNA polymerase is the main enzyme that participates in the transcription process transferring the genetic information from DNA to messenger RNA (mRNA) [1]. Crystallographic structures of the multi-subunit RNA polymerases in eukaryotes [2–4] and bacteria [5–8] engaged in transcription elongation process have been obtained. These atomic-level structures provide static snapshots of the transcription cycle [9–14].

In each nucleotide addition cycle (NAC) of the multi-subunit RNA polymerase, the post-translocation state first allows the substrate NTP to bind to the active site [6]. Then, a critical domain, named trigger loop (TL), can fold then expel the solvent from the active site [15–17], and finally form direct contacts with the substrate NTP. Substitution of a conserved TL histidine can significantly decrease the polymerization rate [18–21]. Recent mutagenesis studies have shed light on the roles of the TL on the nucleotidyl transfer [20,21], and the reverse intrinsic hydrolysis process [22]. Previous MD simulation studies also provided information on TL dynamics and its potential regulatory roles during the translocation process [23,24]. After the catalytic reaction, PP_i forms and releases from the active site [25,26]; then the TL opens and allows the template DNA to translocate so that a new NAC can start. Extensive biochemical and theoretical studies

have been performed to understand the specific steps in the NAC, such as motions of the TL [17], catalysis [26–30], translocation [23,24,31–33] and NTP binding [34,35].

PP_i release in single subunit T7 RNA polymerase is proposed to be tightly coupled with the translocation [36] but the same coupling is not observed in *Escherichia coli* (*E. coli*) RNA polymerase [37]. Interestingly, recent fluorescence and biochemical studies found that the PP_i release in the *E. coli* RNA polymerase occurs shortly before or concurrently with the translocation [33]. Nonetheless, the interplay between the PP_i release step and the TL opening motions at molecular level is still elusive. Previously, we used MD simulations to study the PP_i release in the eukaryotic RNA polymerase II (Pol II) [25]. We proposed a hopping model for Pol II in which PP_i release was coupled with the TL tip motion through the interactions between the TL residue H1085 and the $(Mg-PP_i)^{2-}$ group, and subsequently hopping among several positive charged residues in the secondary channel. Our model further suggested that the PP_i release is a fast dynamic process so that it may not be able to induce the fully TL opening motion.

A comparison of the secondary channel and TL structure between Pol II and bacterial RNA polymerase (RNAP) from *T. thermophilus* (*Tth*) displays substantial differences (See Figure 1) [3,7]. In Pol II, the TL contains a long loop domain (from the Rpb1 residue T1080 to T1095) [3]. However, the TL in RNAP consists of two alpha helices connected by a short turn in the

Author Summary

Pyrophosphate ion (PP$_i$) release is a critical step in the nucleotide addition cycle of transcription elongation. Despite extensive experimental studies, the kinetic mechanism of the PP$_i$ release in bacterial RNA polymerases (RNAP) still remains largely a mystery. As a cellular machine, RNAP contains more than 3000 residues, and thus it is challenging for all-atom molecular dynamics (MD) simulations to directly capture the process of the PP$_i$ release. In this study, we have simulated the dynamics of the PP$_i$ release at microsecond timescale using the Markov State Models (MSMs) built from extensive MD simulations in explicit solvent. MSM is a powerful kinetic network model and can predict long timescale dynamics from many short MD simulations. Our results suggest a simple two-state model for the PP$_i$ release in RNAP, which sharply contrasts with the more complex four-state hopping model in the yeast RNA polymerase (Pol II). We also observe a 3-fold faster dynamics for the PP$_i$ release in RNAP compared to Pol II, consistent with the faster transcription rate in the bacterial systems. Our results greatly improve our understanding of the PP$_i$ release, and also provide predictions to guide future experimental tests.

closed state [6]. This structural difference suggests that the dynamics of the TL folding in these two systems are likely to be different. Moreover, in addition to the conserved *Tth* TL residue H1242, the *Tth* TL residue R1239 also interacts with the substrate NTP [6]; this residue is absent in Pol II and mutation of the counterpart residue in *E. coli* (R933A) can reduce the nucleotide addition rate [20]. Moreover, the secondary channel in *Tth* RNAP is much shorter than that in Pol II (See Figure 1), and exhibits a different layout of the positively charged residues. Specifically, in Pol II, the four residues, K619, K620, K518 and K880 are located at relatively separated sites (See Figure 1A). However, the positively charged residues in *Tth* RNAP: K908, K912, K780 and K1369 are close to each other in a continuous region (See Figure 1B). Given these structural differences, it is of interest to compare the dynamics of PP$_i$ release in RNAP with that in Pol II.

Although conventional all-atom MD simulations can provide the dynamic information for biological macromolecules at atomic resolution, it is still challenging to capture the biologically relevant timescales in microseconds or even longer. Markov State Models (MSMs) constructed from a large number of short simulations provide one way to overcome this timescale gap [38,39]. MSMs have been successfully applied to model the long timescale dynamics that cannot be directly accessed by conventional MD simulations in studying the conformational changes of biological macromolecules [39–41], including our previous study of PP$_i$ release in Pol II [25].

In this study, in order to reveal the mechanism of the PP$_i$ release in RNAP, we constructed a MSM from extensive all-atom molecular dynamics (MD) simulations in explicit solvent with a system size of nearly 300,000 atoms and aggregated simulation time of ~1 µs. Our results reveal that the PP$_i$ release in *Tth* RNAP adopts a simple two-state model with a fast dynamics over a few hundred nanoseconds. Surprisingly, we found that the PP$_i$ release is not coupled with the secondary structure unfolding of TL but only with the side-chain rotation of the TL residue R1239.

Results

PP$_i$ release in bacterial RNAP follows a simple two-state model

To study the release mechanism of the (Mg-PP$_i$)$^{2-}$ group in RNAP, we modeled the PP$_i$-bound RNAP complex by directly cleaving the P$_\alpha$-O bond in the ATP-bound RNAP complex that is derived from the *Tth* RNAP crystal structure (See SI Figure S1 for the two structures and the Methods section for the modeling details) [6]. This modeled PP$_i$-RNAP complex was used as the starting structure for the steered MD (SMD) simulations to obtain the initial release pathways. To eliminate the bias in SMD simulations, we have then performed 100 10-ns MD simulations, and these simulations have widely sampled the region in the secondary channel (See SI Figure S2). Finally, we have constructed a MSM from these simulations to obtain the dynamics and other thermodynamic properties of the PP$_i$ release (See the Methods section for details).

Our MSM shows that the PP$_i$ release in *Tth* RNAP adopts a simple two-state model. In addition to the initial state with the PP$_i$ in the active site (S1 state in Figure 2A), only one additional metastable state is identified (S2 state in Figure 2A), and this state is ~7-fold more populated than the S1 state (See Figure 2B). The S2 state locates in an elongated region where several positively
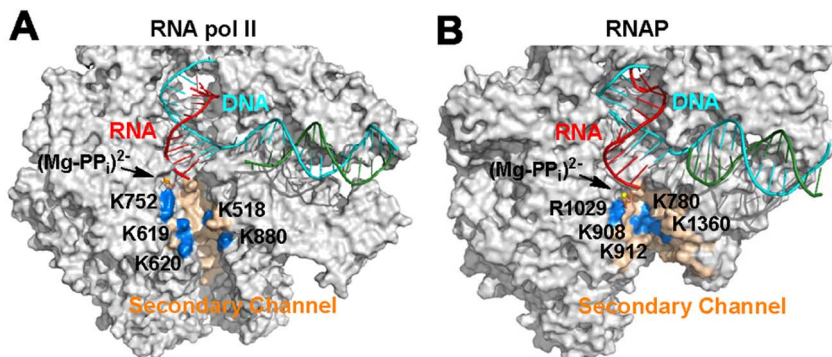
Figure 1. Comparison of the secondary channel (in wheat) of RNA Pol II (A) and RNAP (B). For both structures, RNA, template DNA and non-template DNA are shown in red, cyan and green, respectively. The (Mg-PP$_i$)$^{2-}$ group is represented in stick and sphere models. Several critical residues in the channel: K752, K619, K620, K518 and K880 in RNA Pol II; R1029, K908, K912, K780 and K1360 in RNAP, are highlighted in blue. The Pol II model used to make this figure was taken from our previous study [25].
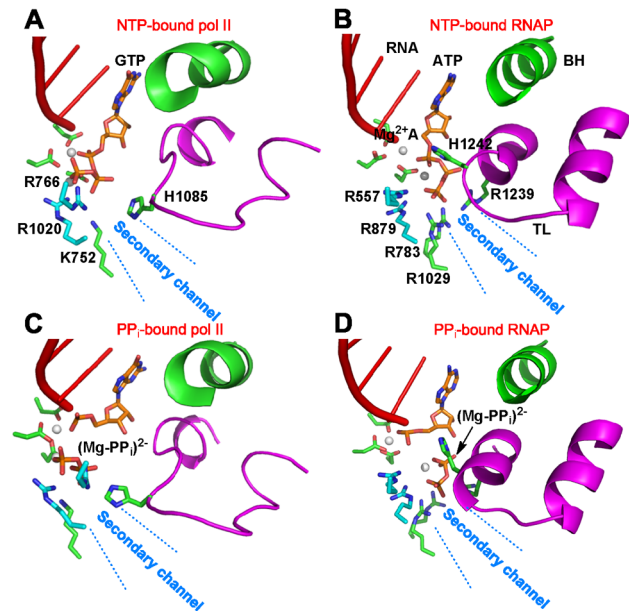doi:10.1371/journal.pcbi.1003020.g001
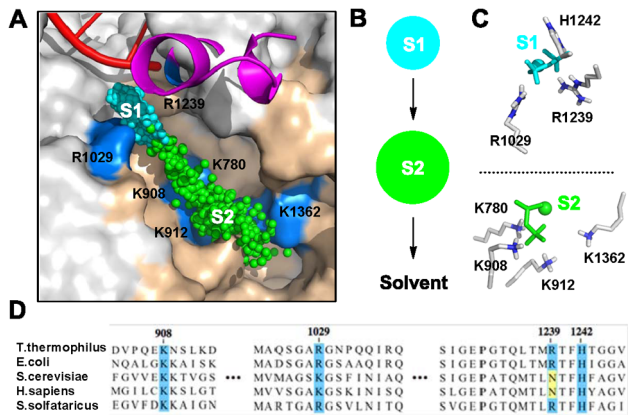
Figure 2. A two-state mechanism for the PP_i release in RNAP revealed by the MSM. (A) Two metastable states (S1 and S2) are identified. 500 randomly selected conformations from each metastable state are superimposed and represented with cyan and green spheres for S1 and S2 respectively. Each sphere indicates the coordinate of the center of mass of the PP_i group. (B) The two metastable states are displayed as two circles, and the size of these circles is proportional to the equilibrium populations of the S1 (12.6%±0.02%) and S2 (87.4%±0.02%) state, (C) Key interactions between $(Mg-PP_i)^{2-}$ group and RNAP in each state are displayed. (D) Conservation analysis of the positively charged residues that interact with the $(Mg-PP_i)^{2-}$ group among different species. The sequence alignment was performed using the online software ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/).
doi:10.1371/journal.pcbi.1003020.g002

Figure 3. Different binding modes between the TL histidine in Pol II (H1085) and RNAP (H1242) with the $(Mg-PP_i)^{2-}$ group. (A) and (B) are the structures of the NTP-bound RNA Pol II and RNAP complexes respectively. (C) and (D) are the corresponding PP_i-bound models. The Bridge Helix (BH, in green), Trigger Loop (TL, in magenta), RNA chain (in red), NTP or PP_i (orange and blue), Mg²⁺ atoms (in white), and selected residues in the active sites are displayed.
doi:10.1371/journal.pcbi.1003020.g003

charged residues can stabilize the $(Mg-PP_i)^{2-}$ group. These results contrast with our previous findings that the $(Mg-PP_i)^{2-}$ group in Pol II hops through four clearly separated metastable states [25].

When the $(Mg-PP_i)^{2-}$ group is in the active site (See Figure 2C), three positively charged β′ residues R1029, H1242 and R1239 can interact with the negatively charged $(Mg-PP_i)^{2-}$ group. The residue R1029 locates at the exit of the active site, and thus it may play similar roles on the PP_i release with its corresponding residue K752 in Pol II (See Figure 2D) [25]. Interestingly, the location of the conserved TL residue H1242 is different from its counterpart residue H1085 in Pol II, though both of them are in direct contact with the $(Mg-PP_i)^{2-}$ group. Both before and after chemistry, H1242 interacts with the $P_α$-O atom of the NTP in RNAP, whereas H1085 is in contact with $P_β$-O atom in Pol II (See Figure 3) [3,6]. To achieve this, the H1242 in RNAP has to locate deeper in the active site compared to H1085. Finally, R1239 in RNAP locates at the same position as H1085 in Pol II, suggesting that these two residues may play similar roles in the PP_i release.

After escaping from the active site, the $(Mg-PP_i)^{2-}$ group reaches the S2 state with an elongated shape. In this state, multiple positively charged residues on the secondary channel (K780, K908, K912 and K1362) can provide favorable electrostatic interactions with the negatively charged $(Mg-PP_i)^{2-}$ group (See Figure 2C). In contrast, the $(Mg-PP_i)^{2-}$ group in Pol II is found to transfer through several hopping sites where groups of positively charged residues are spatially well separated (See Figure 1A) [25]. From the S2 state, the $(Mg-PP_i)^{2-}$ group will directly enter the solvent. In order to elucidate the specific roles of the three important residues: R1029, H1242 and R1239 in the PP_i release (See Figure 2D), we performed additional mutant simulations starting from several different conformations from the S1 and S2 states.

Specific roles of several positively charged residues in PP_i release revealed by mutant simulations

The potential of mean force (PMF) profile along the distance between the $(Mg-PP_i)^{2-}$ group and the Mg²⁺A is displayed in Figure 4A. The PMF plot shows two major free energy basins that are consistent with the two metastable states identified by our MSM. The starting structures chosen for the mutant simulations fall into two different regions in the PMF profile (P1 and P2 sites in Figure 4A). The P1 site is located in the S1 state, while the P2 site is located in the S2 state but near to the boundary between the S1 and S2 states. Initial conformations from these two sites allow us to examine the roles of the residues involved in different stages of the PP_i release.

The mutant simulation results indicate that both residues R1239 and R1029 can facilitate the escape of the $(Mg-PP_i)^{2-}$ group from the active site to S1 state (See Figure 4B). Here, we use the distance between the $(Mg-PP_i)^{2-}$ group and the $P_α$ atom of the 3′-terminal nucleotide of the RNA chain ($d_{αβ}$) to describe the extent of the PP_i release from the active site. In the WT simulations (See P1 in Figure 4A), the $(Mg-PP_i)^{2-}$ group can move towards the exit of active site with the $d_{αβ}$ value increasing from 6 Å to around 8 Å (See the left panel of Figure 4B). However, the R1239A and R1029A mutants lead to a weaker tendency for the $(Mg-PP_i)^{2-}$ group to escape the active site (the $d_{αβ}$ value fluctuates around 5.5 Å, middle and right panels in Figure 4B). On the other hand, the R1029K mutant is shown to have a similar effect to help the $(Mg-PP_i)^{2-}$ group to leave the active site as in WT(see Figure S4A). These results indicate that positively charged residues play a crucial role to facilitate the PP_i to release from the active site.

Notably, the H1242A mutant can dramatically promote the PP_i release from P1 site (See SI Figure S4C), suggesting that H1242 may prevent the PP_i release from the active site. In contrast, the TL residue H1085 in Pol II was previously found to facilitate the
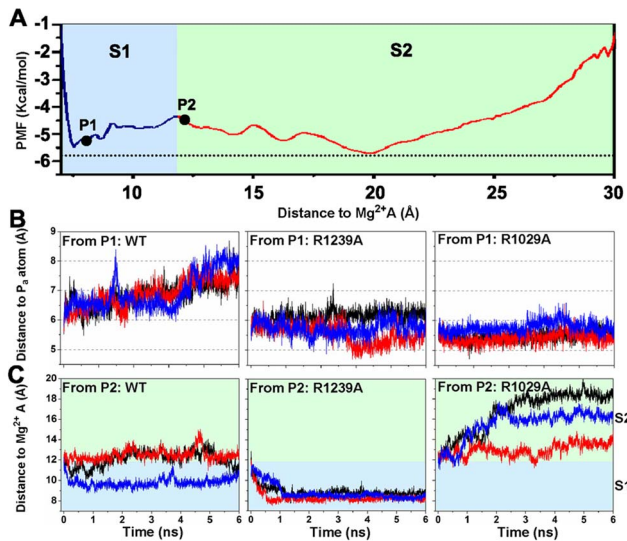
Figure 4. Single mutant simulations reveal the roles of the critical residues H1242, R1239 and R1029 in the PP$_i$ release. (A) Potential of mean force (PMF) plot along the distance between the PP$_i$ group and Mg^{2+}A. The initial conformations of the mutant simulations are highlighted as black spheres (P1 and P2). (B) The distance between the PP$_i$ group and the P$_\alpha$ atom as a function of the simulation time for WT (left panel), R1239A (middle panel), and R1029A (right panel) simulations initiated from P1. We chose this reaction coordination because this distance can directly measure the relative motions between the terminal RNA nucleotide and the PP$_i$ group before it leaves the active site. (C) The same as (B) except that all the simulations were initiated from P2, and the distance between the PP$_i$ group and the Mg^{2+}A was shown. In (A) and (C), the S1 and S2 state are highlighted in blue and light green respectively.
doi:10.1371/journal.pcbi.1003020.g004

PP$_i$ release from the active site [25]. This difference may be due to the different locations of these two residues in the active site. Compared with H1085 in Pol II, H1242 in RNAP locates significantly deeper inside of the active site (See Figure 3D). Therefore, it will be more difficult for H1242 to rotate and help the (Mg-PP$_i$)$^{2-}$ group to leave the active site. Instead, H1242 can provide an attractive interaction to prevent the PP$_i$ release.

Next, we evaluated the roles of residues R1239 and R1029 in PP$_i$ release when the (Mg-PP$_i$)$^{2-}$ group is at the S2 state (P2 in Figure 4A). In the WT system, the (Mg-PP$_i$)$^{2-}$ group fluctuates around its initial location within our simulations at a few nanoseconds, which was also observed in the R1029K mutant simulations initiated from P2 (Figure 4C). Intriguingly, R1029A and R1239A substitutions lead to dramatic, but opposite, effects. The R1029A substitution facilitates the PP$_i$ release toward the solvent (Figure 4C, right panel). Combined with the previous observations, we conclude that the R1029 may facilitate the PP$_i$ release from the active site but prevents the PP$_i$ release when it arrives at the S2 state. Thus R1029 plays a similar role as the corresponding residue K752 in Pol II (See Figure 2D) [25]. In contrast, the R1239A substitution drives the (Mg-PP$_i$)$^{2-}$ group back to the S1 state, suggesting that R1239 is critical for the (Mg-PP$_i$)$^{2-}$ group to escape the active site (middle panel in Figure 4C). This indicates that R1239, rather than H1242 residue, plays the role in PP$_i$ release most equivalent to that played by H1085 in Pol II. Compared with its counterpart residue H1085 in Pol II, the R1029 has a longer and more flexible side chain. In addition, it can form a stronger salt bridge with the (Mg-PP$_i$)$^{2-}$ group. Therefore, the side-chain rotation of R1239 alone may be sufficient to facilitate the PP$_i$ release.

PP$_i$ release does not induce the TL backbone unfolding, but is tightly coupled with the side-chain rotation of the TL residue R1239. In order to reveal if the PP$_i$ release is coupled with the TL unfolding in RNAP, we have monitored the structural changes of the TL during the PP$_i$ release. We calculated the RMSD values of the heavy atoms (non-hydrogen) for both the complete TL domain (β' residue Q1235 to G1255) and its tip part (residues R1239 to A1249). RMSD values for the complete TL mostly fluctuate between 1 and 3 Å (See SI Figure S5A), indicating that TL does not unfold during the PP$_i$ release. On the other hand, for the tip part of the TL, the RMSD values mostly fluctuate between 1 and 2.5 Å when the (Mg-PP$_i$)$^{2-}$ group is in the S1 state (See SI Figure S5B). But RMSD values increase to between 2 and 4.5 Å when the (Mg-PP$_i$)$^{2-}$ group reaches the S2 state, indicating that the tip region of the TL becomes more flexible after the (Mg-PP$_i$)$^{2-}$ group leaves the active site. The Mean First Passage Time (MFPT, the average transition time from S1 state to S2 state) for the PP$_i$ release is around 0.5 μs (See SI Table S1 and the Methods section for MFPT calculation details), which is three-fold faster than that in Pol II (∼1.5 μs) [25]. These results further support the idea that PP$_i$ release occurs too fast to lead directly to unfolding of the TL helices.

More interestingly, we found a direct correlation between the PP$_i$ release and the side chain rotation of the TL residue R1239. The PMF profile in Figure 5A clearly shows that the transition of the (Mg-PP$_i$)$^{2-}$ group from the S1 to S2 state correlates with the rotational motion of the residue R1239, with its distance to the Mg^{2+}A ion increasing from 10 to 14 Å (See Figure 5B and C). These results are consistent with our previous observation that R1239 can facilitate the PP$_i$ release. When the (Mg-PP$_i$)$^{2-}$ group arrives at the S2 state (See Figure 5D), its interaction with the residue R1239 is lost, and this increase the fluctuations of the residue R1239 (See Figure 5A).
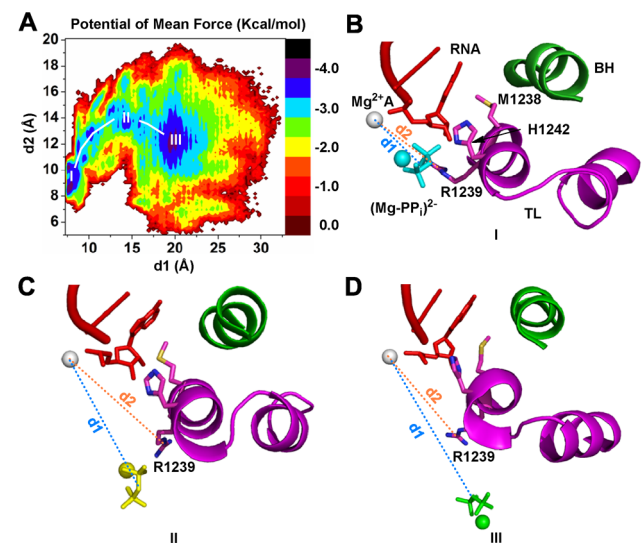
Figure 5. Side chain rotation of the TL residue R1239 facilitates the PP$_i$ release. (A) Potential of mean force (PMF) of the (Mg-PP$_i$)$^{2-}$ conformations projected on two reaction coordinates: d$_1$ (distance between the PP$_i$ group and Mg^{2+}A) and d$_2$ (distance between the Guanidine C atom from the R1239 and Mg^{2+}A). Representative structures from three free energy minima labeled as I, II and III in the PMF plot are shown in (B), (C) and (D) respectively. The free energy minimum I belongs to the S1 state, while the other two belong to the S2 state. The structural presentation is the same as Figure 3.
doi:10.1371/journal.pcbi.1003020.g005

We did not observe the backbone unfolding of the TL during the PP$_i$ release in our model. However, since the timescale for the PP$_i$ release is an order of magnitude longer than our individual seeding MD simulations, there exists a possibility that the seeding MD simulations may be biased by initial conformations obtained from steered MD, where the TL is always folded. We thus performed control simulations of both isolated TL domain and a truncated model in which all the motifs surrounding the TL domains were included to further investigate its folding. For an isolated TL domain (from A1225 to A1265) in free solution, it can quickly unfold at ~210-ns (See SI Figure S6A–C). However, the inclusion of the RNAP motifs surrounding the TL domain can shield not only TL helix facing the active site but also part of the other TL helix facing the secondary channel from the solvent. This may greatly stabilize the secondary structure of the TL domain and prevent it from unfolding. Indeed, no unfolding events were observed in two independent 300-ns MD simulations, and the secondary structure of the TL domain was well preserved (See SI Figure S7A–C). These control simulations suggest that the TL domain must be exposed to the solvent before its secondary structures can be unfolded. We speculate that the side chain rotation of the TL residue R1239 may initiate and allow the overall motion of the TL domain, and this will in turn make the TL more exposed to the solvent and eventually unfold. Interestingly, one recent crystal structure of the RNAP captures an open state of the TL, and in this structure the completely solvent-exposed segment (TL residues from 1246 to 1254) is unfolded but, the segment that is not fully exposed to the solvent remains folded (TL residues from 1235 to 1243) [7]. Based on the control simulations, we conclude that the full opening motion of the TL likely occurs at a timescale longer than the timescale of PP$_i$ release.

Discussion

Based on extensive unbiased MD simulations, we built a MSM for the PP$_i$ release in RNAP to elucidate its long timescale dynamics. The MSM identified a two-state model for the PP$_i$ release (See Figure 6A). The mutant simulations indicate that the β' residues R1239 and R1029 can facilitate the escape of the (Mg-PP$_i$)$^{2-}$ group from the active site after the catalytic reaction (See Figure 4B). Then the (Mg-PP$_i$)$^{2-}$ group transfers to the S2 state, where it forms favorable interactions with four positively charged residues on the secondary channel: K908, K912, K780 and K1369 (See Figure 6A). More strikingly, our work suggests that the PP$_i$ release does not induce the TL unfolding but tightly couples to the side-chain rotation of the TL residue R1239, which in turn makes the TL tip region more flexible. Furthermore, our control simulations show that TL is stable in Pol II, but can quickly unfold (within 200 ns) when exposed to the solvent. We thus speculate that the rotation of R1239 that accompanies the PP$_i$ release may allow solvent to re-enter the active site and promote the overall movements of the TL domain; This TL movements would further lead to its exposure to solvent and eventually allow TL unfolding. However, the timescales for this solvent-induced TL unfolding may be significantly longer than that of the PP$_i$ release so that we didn't observe it in our simulations.

We found that the TL in RNAP may be more difficult to unfold than that of Pol II, since its secondary structures barely unfold upon the PP$_i$ release. Therefore, if the open state of the TL is a pre-requisite step for the translocation as recently suggested by both experimental [33] and computational studies [24], it is intriguing that the transcription rate for bacterial RNAP is much faster than that of Pol II [27]. Despite that the more stable secondary structure of the TL in RNAP may slow down its opening motion, its reverse closing motion may be spontaneous and fast. This fast closing motion may further accelerate the nucleotide addition process to achieve an overall fast transcription rate.

Finally, our MSM indicates that the PP$_i$ release in bacterial RNAP is faster than that in Pol II [25]. This faster dynamics is due to several factors: First, the secondary channel of RNAP is shorter than the that of Pol II due to the absence of the funnel region, therefore the (Mg-PP$_i$)$^{2-}$ release path is shorter, which leads to a faster PP$_i$ release from the active site of RNAP. Next, in RNAP, the PP$_i$ only needs to overcome a single free energy barrier before it can be released to the solvent (See Figure 6A). In contrast, the PP$_i$ release in Pol II was found to go over multiple free energy barriers sequentially before it could be released (See Figure 6B). Furthermore, in our model for RNAP, the S2 state (in the pore) has a population over seven times larger than that of the S1 state (in the active site). Thermodynamically, this difference will favor release of PP$_i$ from the active site. However, in Pol II, the equilibrium population of the S2 state (the first state in the pore) is comparable to that of the S1 state (in the active site) [25]. This difference may be due to the fact that the S2 state in bacterial RNAP is greatly stabilized by four positively charged residues that are spatially close to each other (K908, K912, K780 and K1369), but in Pol II, these positively charged residues locate at relatively
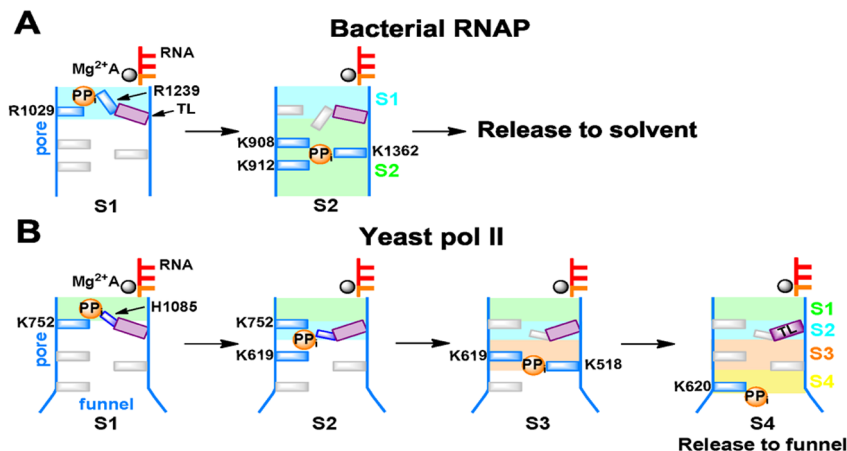
Figure 6. Cartoon models of the PP$_i$ release in RNAP (A) and Pol II (B).
doi:10.1371/journal.pcbi.1003020.g006

separate sites (See Figure 1A). Finally, R1239 in RNAP can substantially facilitate the (Mg-PP$_i$)$^{2-}$ release from the active site all the way to the solvent due to its longer and more flexible side chain (See Figure 5). However, its counterpart residue for PP$_i$ release from Pol II, H1085, only promotes the PP$_i$ escape from the active site to the first metastable state, S1, rather than all the way to solvent [25].

Methods

We constructed the MSM to study the PP$_i$ release in RNAP, and our algorithm consists of the following steps: (1) Model the PP$_i$ bound complex. (2) Generate initial release pathways using SMD simulations. (3) Seed unbiased MD simulations from these initial pathways, and (4) Construct the MSM to identify metastable intermediate states and obtain both thermodynamics and kinetics of the PP$_i$ release.

1. System setup and MD simulations

Setup of the bacterial RNAP elongation complex (EC). The RNAP model was built from the crystal structure of the *T. thermophilus* RNAP bound with a non-hydrolysable substrate analogue, AMPCPP (PDB ID: 2O5J) [6]. Two missing motifs in the β′ subunit, from the residues 208 to 390 and from 1272 to 1328, were replaced with three GAG residues. Since these two motifs are largely exposed to the solvent, we thought they would not affect the dynamics of the PP$_i$ release. Other parts of the EC, including four subunits (β, α1, α2 and ω domains), the downstream DNA, DNA-RNA hybrid, two Mg^{2+} ions, two Zn^{2+} ions, five crystal waters in the active site and the AMPCPP molecule, were retained.

Modeling the ATP- and PP$_i$- bound RNAP complexes. We created ATP-bound RNAP complex by replacing the bridge carbon atom that connects the P$_\alpha$ and P$_\beta$ atoms of AMPCPP with an oxygen atom. The minimized ATP-bound RNAP complex exhibits reasonable deviations from the crystal structure (See SI Figure S1A). Based on this ATP-bound RNAP complex, we built the PP$_i$-bound RNAP complex by cleaving the P$_\alpha$-O bond of the ATP to form the PP$_i$ group product.

The AMBER03 force field [42] was used to describe the protein residues, DNA, RNA, and metal ions. The parameters of ATP were taken from our previous study [43]. The (Mg-PP$_i$)$^{2-}$ group was treated as one group due to the significant internal charge transfer and its parameters were adopted from our previous study [25].

Molecular Dynamics (MD) simulation details. We used GROMACS 4.5 to conduct all the MD simulations [44]. Each EC was solvated with SPC water [45] in a cubic box and the minimum distance from the protein to the wall was 7.0 Å. To neutralize the system, 77 Na$^+$ ions were added. There are 297,944 atoms in the final PP$_i$ bound RNAP complex. Van der Waals and short-range electrostatic interactions were cut off at 10 Å. Long-range electrostatic interactions were treated with the Particle-Mesh Ewald (PME) summation method [46,47]. The MD simulations were run at 1 bar and 310K using the Berendsen barostat [48] and the velocity rescaling thermostat [49], respectively. The LINCS algorithm was used to constrain all the chemical bonds [50]. The time-step was 2 fs and we updated the neighbor list every 10 steps. The solvated system was minimized with the steepest decent minimization method followed by a 120 ps MD simulation with position restrains on the heavy atoms of the proteins, DNA and RNA chains. The minimized PP$_i$-bound RNAP complex displays minor fluctuations for the active site residues comparing to the crystal structure (See SI Figure S1B), indicating that our model is a good starting point.

2. Generating initial PP$_i$ release pathways using SMD

In order to obtain the initial PP$_i$ release pathway, we applied steered MD simulations [51] to pull the (Mg-PP$_i$)$^{2-}$ group out of the active site. The pulling was performed along three directions with the aim of considering all the possible PP$_i$ release pathways. Three groups of residues were used to determine the pulling directions: β′ subunit residues 1136–1145, 908–914 and 1246–1253 (named as group I, II and III respectively). Two sets of pulling simulations were along the wall of the secondary channel: one was pulled towards the center of the C$_\alpha$ atoms of group I residues, and the other was directed to the center of C$_\alpha$atoms of both group I and group II residues. The third set of pulling simulations pointed to the center of the C$_\alpha$atoms of group II and group III residues, and toward the solvent. The external force was only applied on the center of mass of the PP$_i$ group with a force constant of 0.5 kJ mol^{-1} Å^{-2} and pulling rate of 0.01 Å/ps. For each pulling direction, five independent steered MD simulations were performed starting from the final snapshot derived from 5 parallel MD simulations of the PP$_i$-bound RNAP complex.

3. Seeding unbiased MD simulations

We first divided the conformations from SMD simulations into 20 clusters using the K-center clustering algorithm [52]. In the clustering, the distance between a pair of conformations was set to be the RMSD value of three PP$_i$ atoms (the bridge oxygen and two phosphate atoms). To compute RMSD, the structure was aligned to the energy minimized PP$_i$-RNAP complex by the C$_\alpha$atoms of the bridge helix domain. We then randomly selected 5 conformations from each cluster (a total of 100 conformations) to conduct unbiased MD simulations. Each simulation was run for 10 ns and the snapshots were saved every 2 ps. Altogether, we obtained an aggregation of ~1 μs simulations with 500,000 conformations.

4. Constructing the Markov State Model (MSM)

In MSM, the conformational space is divided into a number of metastable macrostates and the fast motions are integrated out by coarse graining in time with a discrete unit of Δt. The model is markovian if Δt is longer than the intra-state relaxation time. In other words, the probability for the system to be at a given state at time $t + \Delta t$ only depends on the state at time t. In MSM, the long timescale dynamics can be modeled by the first-order master equation.

$$P(n\Delta t) = [T(\Delta t)]^n P(0) \qquad (1)$$

Where $P(n\Delta t)$is the state populations vector at time $n\Delta t$, and T is the transition probability matrix. $\Delta t$is the lag time of the model. To calculate T, one can normalize the transition count matrix generated by counting the number of transitions between each pair of states at the observation interval of Δt from MD trajectories. MSM has been successfully applied to model conformational changes that occur at timescales that cannot be directly accessed by conventional MD simulations such as protein folding [39,40,53–55].

To construct the MSM, we have followed a splitting and lumping procedure [52]:

Splitting MD conformations into microstates. We have divided all the conformations from our seeding MD simulations (500,000 conformations) into 200 microstates by employing the K-center clustering algorithm [52]. The distance between a pair of conformations was set to be the RMSD value of three PP$_i$ atoms (the bridge oxygen and two phosphate atoms). To compute RMSD, the structure was aligned to the modeled PP$_i$-bound RNAP complex according to the C$_\alpha$ atoms of the BH residues.

The microstates are small, and the average RMSD values to its central conformation in each state are only ~2 Å.

The transition probability matrix T_{ij} from state i to state j was obtained by counting the transition numbers N_{ij} observed in the MD trajectories within a certain lag time T, and normalizing each row by:

$$T_{ij} = N_{ij} / \sum_j N_{ij}$$

We used the sliding-window to count the transitions due to the limited samplings. To avoid the re-crossing events, we count the transitions from i to j only if the state j can stay in j at least within 50 ps without transferring to other states.

Lumping microstate into macrostates. We applied the Robust Perron Cluster Cluster Analysis (PCCA+) algorithm [56] to lump the 200 microstates obtained above into 2 macrostates. The number of macrostates was determined from the implied timescale plot of the 200 microstate. The plot levels off starting from the lag time of 4 ns and exhibits one clear gap, suggesting that two macrostates exist (See SI Figure S3A). Finally we chose a lag time of 4.5 ns to calculate the equilibrium populations and the Mean First Passage Time (MFPT).

We calculated the MFPTs to estimate the average transition rates between each macrostates pair. As described before [25,54], the MFPT can be determined using the following formula:

$$MFPT_{if} = \sum_j P_{ij} \times (t_{ij} + MFPT_{jf})$$

where P_{ij} is the transition probability from state i to state j, t_{ij} is the lag time used to construct the transition probability matrix $\boldsymbol{T}_{ij}$, t_{ij} is equal to 4.5 ns in our model, and $MFPT_{if}$ is the mean first passage time of the state j to final state f. For each transition, a set of linear equations can be solved under the boundary condition in which $MFPT_{ff} = 0$. The uncertainties of the MFPT were obtained by bootstrapping the MD trajectories for 100 times.

5. Validating the Markov State Model

In order to check if the model is markovian, we have plotted the implied timescales (τ_k) as a function of the lag time τ:

$$\tau_k = -\frac{\tau}{\ln \mu_k(\tau)} \qquad (2)$$

where μ_k is the eigenvalue of the transition probability matrix T with the lag time τ. The implied timescales correspond to the average transition times between two groups of states, and thus indicate the dynamics of the system. If τ is sufficiently large, the model is markovian, and the predicted implied timescales will not change upon the further increase of the lag time. In our system, the implied timescale plots reach the plateau at the lag time of ~4 ns (See SI Figure S3A). Therefore, we select the lag time of 4.5 ns to construct the final MSM.

To further validate the model, we predicted the probability for a given macrostate to stay within it after a certain lag-time based on our MSM, and this predicted values are in good agreement with those obtained from the original MD simulations (See SI Figure S3B).

6. Control simulations of the truncated systems

In order to investigate the stability of the TL in free solution, we have performed a 300 ns control simulation with the isolated TL domain (A1225 to A1265) in solution (with ~6900 atoms, See SI

Figure S6B). However, it is difficult to extend individual MD simulations of the complete transcription complex (nearly 300 K atoms) to hundreds of nanoseconds due to its high computing cost. Therefore, we have also performed simulations with a truncated RNAP complex containing all the motifs surrounding the TL domain (See SI Figure S7A), including β subunit residues 381–569, 831–1049, β′ residues 604–794, 901–1470, 10 upstream hybrid DNA-RNA base pairs, 6 downstream DNA base pairs and Mg^{2+}A in the active site. The final solvated system only contains ~118 K atoms, but it still takes more than 2 months to perform one 300-ns simulation using 24 CPU cores.

The explicit SPC water model was used for the MD simulations, 1 and 29 Na$^+$ ions were added to neutralize the isolated TL and truncated RNAP model respectively. The other setups for the MD simulations were the same as the seeding MD simulations. We have performed one 300-ns simulation for the isolated TL and the other two 300-ns simulation for the truncated RNAP. For the truncated RNAP model, several terminal residues that are truncated from the complete model were fixed in the simulations in order to avoid undesired unfolding.

Supporting Information

Figure S1 Energy minimized structures of the ATP-bound (A) and PP$_i$-bound RNAP complexes (B). The two complexes are superimposed with the crystal structure of the AMPCPP bound RNAP complex (PDB ID: 2O5J, gray). The BH, TL, RNA chain, Mg^{2+} ions and the the substrate are shown in green, magenta, red, yellow, and organe/red, respectively. Several residues around the active site are also highlighted.
(TIF)

Figure S2 Our 100 unbaised MD simulations widely sampled the secondary channel. Three conformations from each simulation at 0 ns, 5 ns and 10 ns are shown as spheres, and connected by sticks.
(TIF)

Figure S3 (A). Implied timescale plot as the function of the lag time for the 200-microstates MSM (left panel) and 2-macrostates MSM (right panel). (B). Validation of our MSM. The probability for a given macrostate to stay within it after a certain lag-time can be predicted from our MSM (blue dashed lines), and this predicted values are comparable to the direct counts from the MD simulations (red dashed lines). The lag time we used is 4.5 ns.
(TIF)

Figure S4 (A) The distance between the PP$_i$ group and the P$_\alpha$ as the function of the simulation time for R1029K mutant MD simulation initiated from P1 conformation. (B) R1029K mutant MD simulations initiated from P2, the distance between PP$_i$ group and Mg^{2+}A was shown. (C) Same as (A) but for R1029K mutant MD simulation. Please refer to the caption of Figure 4A for additional details.
(TIF)

Figure S5 Potential of mean force (PMF) plots for: (A) the complete TL (Q1235 to G1255) and (B) the TL tip (R1239 to A1249). Both of the PMF plots were projected on two reaction coordinates: the distance between the PP$_i$ group and the Mg^{2+}A (d$_1$), and the heavy-atom RMSD comparing to the energy minimized PP$_i$-bound RNAP complex.
(TIF)

Figure S6 MD simulation of the isolated TL in solution. (A) The heavy-atom RMSD of the TL residues (from Q1255 to G1275) as a function of the simulation time. (B) Structures of two snapshots

from the MD simulation at 0ns and 300ns. (C) The secondary structure analysis of the TL domain along the simulation time. (TIF)

Figure S7 MD simulations of the truncated RNAP complex. (A) The initial structure of the truncated RNAP complex. (B) The heavy-atom RMSD of the TL residues (from Q1255 to G1275) as a function of the simulation time for two independent MD simulations. (C) The secondary structure analysis of the TL domain along the simulation time. (TIF)

Table S1 Mean First Passage Time (MFPT) obtained from our MSMs for transitions between two metastable states. See Methods section for details of the MFPT calculations. (DOCX)

Author Contributions

Conceived and designed the experiments: LTD FPA DW XH. Performed the experiments: LTD. Analyzed the data: LTD FPA DW XH. Wrote the paper: LTD FPA DW XH.

References

1. Fuda NJ, Ardehali MB, Lis JT (2009) Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature 461: 186–192.
2. Kornberg RD (2007) The molecular basis of eukaryotic transcription. Proc Natl Acad Sci USA 104: 12955–12961.
3. Wang D, Bushnell DA, Westover KD, Kaplan CD, Kornberg RD (2006) Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. Cell 127: 941–954.
4. Cramer P, Bushnell DA, Fu J, Gnatt AL, Maier-Davis B, et al. (2000) Architecture of RNA polymerase II and implications for the transcription mechanism. Science 288: 640–649.
5. Tagami S, Sekine S, Kumarevel T, Hino N, Murayama Y, et al. (2010) Crystal structure of bacterial RNA polymerase bound with a transcription inhibitor protein. Nature 468: 978–982.
6. Vassylyev DG, Vassylyeva MN, Zhang J, Palangat M, Artsimovitch I, et al. (2007) Structural basis for substrate loading in bacterial RNA polymerase. Nature 448: 163–168.
7. Vassylyev DG, Vassylyeva MN, Perederina A, Tahirov TH, Artsimovitch I (2007) Structural basis for transcription elongation by bacterial RNA polymerase. Nature 448: 157–162.
8. Vassylyev DG, Sekine S, Laptenko O, Lee J, Vassylyeva MN, et al. (2002) Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 Å resolution. Nature 417: 712–719.
9. Cheung A, Cramer P (2012) A Movie of RNA Polymerase II Transcription. Cell 149: 1431–1437.
10. Sydow JF, Cramer P (2009) RNA polymerase fidelity and transcriptional proofreading. Curr Opin Struc Biol 19: 732–739.
11. Svetlov V, Nudler E (2009) Macromolecular micromovements: how RNA polymerase translocates. Curr Opin Struc Biol 19: 701–707.
12. Landick R (2009) Transcriptional pausing without backtracking. Proc Natl Acad Sci USA 106: 8797–8798.
13. Brueckner F, Ortiz J, Cramer P (2009) A movie of the RNA polymerase nucleotide addition cycle. Curr Opin Struc Biol 19: 294–299.
14. Svetlov V, Nudler E (2012) Basic mechanism of transcription by RNA polymerase II. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms 1829: 20–28.
15. Palangat M, Larson MH, Hu X, Gnatt A, Block SM, et al. (2012) Efficient reconstitution of transcription elongation complexes for single-molecule studies of eukaryotic RNA polymerase II. Transcription 3: 146–153.
16. Larson MH, Zhou J, Kaplan CD, Palangat M, Kornberg RD, et al. (2012) Trigger loop dynamics mediate the balance between the transcriptional fidelity and speed of RNA polymerase II. Proc Natl Acad Sci USA 109: 6555–6560.
17. Huang X, Wang D, Weiss DR, Bushnell DA, Kornberg RD, et al. (2010) RNA polymerase II trigger loop residues stabilize and position the incoming nucleotide triphosphate in transcription. Proc Natl Acad Sci USA 107: 15745–15750.
18. Kaplan CD, Jin H, Zhang IL, Belyanin A (2012) Dissection of Pol II Trigger Loop Function and Pol II Activity–Dependent Control of Start Site Selection In Vivo. PLoS Genet 8: e1002627.
19. Kaplan CD (2012) Basic mechanisms of RNA polymerase II activity and alteration of gene expression in Saccharomyces cerevisiae. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1829: 39–54.
20. Zhang J, Palangat M, Landick R (2009) Role of the RNA polymerase trigger loop in catalysis and pausing. Nat Struct Mol Biol 17: 99–104.
21. Kaplan CD, Larsson KM, Kornberg RD (2008) The RNA Polymerase II Trigger Loop Functions in Substrate Selection and Is Directly Targeted by a-Amanitin. Mol Cell 30: 547–556.
22. Yuzenkova Y, Zenkin N (2010) Central role of the RNA polymerase trigger loop in intrinsic RNA hydrolysis. Proc Natl Acad Sci USA 107: 10878–10883.
23. Kireeva ML, Opron K, Seibold SA, Domecq C, Cukier RI, et al. (2012) Molecular dynamics and mutational analysis of the catalytic and translocation cycle of RNA polymerase. BMC Biophysics 5: 11.
24. Feig M, Burton ZF (2010) RNA Polymerase II with Open and Closed Trigger Loops: Active Site Dynamics and Nucleic Acid Translocation. Biophys J 99: 2577–2586.
25. Da LT, Wang D, Huang X (2012) Dynamics of Pyrophosphate Ion Release and Its Coupled Trigger Loop Motion from Closed to Open State in RNA Polymerase II. J Am Chem Soc 134: 2399–2406.
26. Carvalho ATP, Fernandes PA, Ramos MJ (2011) The Catalytic Mechanism of RNA Polymerase II. J Chem Theory and Comput 7: 1177–1188.
27. Maiuri P, Knezevich A, De Marco A, Mazza D, Kula A, et al. (2011) Fast transcription rates of RNA polymerase II in human cells. EMBO Rep 12: 1280–2185.
28. Castro C, Smidansky ED, Arnold JJ, Maksimchuk KR, Moustafa I, et al. (2009) Nucleic acid polymerases use a general acid for nucleotidyl transfer. Nat Struct Mol Biol 16: 212–218.
29. Sosunov V, Sosunova E, Mustaev A, Bass I, Nikiforov V, et al. (2003) Unified two-metal mechanism of RNA synthesis and degradation by RNA polymerase. EMBO J 22: 2234–2244.
30. Seibold SA, Singh BN, Zhang C, Kireeva M, Domecq C, et al. (2010) Conformational coupling, bridge helix dynamics and active site dehydration in catalysis by RNA polymerase. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms 1799: 575–587.
31. Hein PP, Palangat M, Landick R (2011) RNA Transcript 3′-Proximal Sequence Affects Translocation Bias of RNA Polymerase. Biochemistry 50: 7002–7014.
32. Feig M, Burton ZF (2010) RNA polymerase II flexibility during translocation from normal mode analysis. Proteins: Structure, Function, and Bioinformatics 78: 434–446.
33. Malinen AM, Turtola M, Parthiban M, Vainonen L, Johnson MS, et al. (2012) Active site opening and closure control translocation of multisubunit RNA polymerase. Nucleic Acids Res 40: 7442–7551.
34. Batada NN, Westover KD, Bushnell DA, Levitt M, Kornberg RD (2004) Diffusion of nucleoside triphosphates and role of the entry site to the RNA polymerase II active center. Proc Natl Acad Sci USA 101: 17361–17364.
35. Gong XQ, Zhang C, Feig M, Burton ZF (2005) Dynamic error correction and regulation of downstream bubble opening by human RNA polymerase II. Mol Cell 18: 461–470.
36. Yin YW, Steitz TA (2004) The structural mechanism of translocation and helicase activity in T7 RNA polymerase. Cell 116: 393–404.
37. Abbondanzieri EA, Greenleaf WJ, Shaevitz JW, Landick R, Block SM (2005) Direct observation of base-pair stepping by RNA polymerase. Nature 438: 460–465.
38. Bowman GR, Beauchamp KA, Boxer G, Pande VS (2009) Progress and challenges in the automated construction of Markov state models for full protein systems. J Chem Phys 131: 124101.
39. Noé F, Fischer S (2008) Transition networks for modeling the kinetics of conformational change in macromolecules. Curr Opin Struc Biol 18: 154–162.
40. Bowman GR, Voelz VA, Pande VS (2010) Taming the complexity of protein folding. Curr Opin Struc Biol 21: 4–11.
41. Chodera JD, Singhal N, Pande VS, Dill KA, Swope WC (2007) Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. J Chem Phys 126: 155101–155117.
42. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, et al. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 24: 1999–2012.
43. Meagher KL, Redman LT, Carlson HA (2003) Development of polyphosphate parameters for use with the AMBER force field. J Comput Chem 24: 1016–1025.
44. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, et al. (2005) GROMACS: fast, flexible, and free. J Comput Chem 26: 1701–1718.
45. Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J (1981) Interaction models for water in relation to protein hydration. In B. Pullman, editor. Intermolecular forces. Dordrecht: Reidel Publishing Company.

46. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, et al. (1995) A smooth particle mesh Ewald method. J Chem Phys 103: 8577–8593.

47. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. J Chem Phys 98: 10089–10092.

48. Berendsen HJC, Postma JPM, Van Gunsteren WF, DiNola A, Haak J (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81: 3684–3690.

49. Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. J Chem Phys 126: 014101.

50. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: a linear constraint solver for molecular simulations. J Comput Chem 18: 1463–1472.

51. Isralewitz B, Gao M, Schulten K (2001) Steered molecular dynamics and mechanical functions of proteins. Curr Opin Struc Biol 11: 224–230.

52. Bowman GR, Huang X, Pande VS (2009) Using generalized ensemble simulations and Markov state models to identify conformational states. Methods 49: 197–201.

53. Zhuang W, Cui RZ, Silva DA, Huang X (2011) Simulating the T-Jump-Triggered Unfolding Dynamics of trpzip2 Peptide and Its Time-Resolved IR and Two-Dimensional IR Signals Using the Markov State Model Approach. J Phys Chem B 115: 5415–5424.

54. Silva DA, Bowman GR, Sosa-Peinado A, Huang X (2011) A Role for Both Conformational Selection and Induced Fit in Ligand Binding by the LAO Protein. PLoS Comput Biol 7: e1002054.

55. Huang X, Bowman GR, Bacallado S, Pande VS (2009) Rapid equilibrium sampling initiated from nonequilibrium data. Proc Natl Acad Sci USA 106: 19765–19769.

56. Deuflhard P, Weber M (2005) Robust Perron cluster analysis in conformation dynamics. Linear Algebra Appl 398: 161–184.