Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Scienze Biochimiche e Biotecnologiche

Ciclo XXIX

Settore Concorsuale di afferenza: 03/D1

Settore Scientifico disciplinare: CHIM/08

Multiscale Simulations to Dissect Enzymatic Processing of Nucleic Acids

Presentata da: Vito Genna

Coordinatore Dottorato	**Relatore**
Prof. Santi Mario Spampinato	**Dr. Marco De Vivo**

Esame finale anno 2017

Acknowledgments

I am grateful to my supervisor Dr. Marco De Vivo who, along these challenging years, shaped – day by day with passion – the scientist I am today. Importantly, I am also grateful to Prof. Matteo Dal Peraro who, in 2013, gave me a chance to become a scientist by offering me the opportunity to join his team at the EPFL (Switzerland) few days after my Master graduation at University of Cagliari.

Doing a PhD requires daily endeavors, passion and selflessness. Personally, I think this is not "a job" rather I firmly believe that doing research is a vocation that needs to be fed in every moment of your life. A researcher never stops to think about new ideas or ways to find out a solution to a given problem, to cure a disease and improve the quality of life of peoples less lucky than us. This is something out of your control; your mind is not yours anymore. It gets spread in honor of social well-being since you are called to work for the humanity and not for yourself. People trust you, they put their hopes in you, they have expectations on you and you cannot disappoint their expectations. Science is a beast that takes possession of you, pervades your soul and to stay alive it strongly asks to face and solve problems every day. Passion and fantasy are the keys. That is why I am grateful to Dr. Marco De Vivo. He was able to push my horizons further day after day.

Sometimes all those efforts annihilate you. Sometimes you feel you made a bad choice and you are led to think to give up. Exactly in those moments you need to be loved by your partner, family and friends. They represent the energy you need to restart after a storm because only them are able to deeply comprehend you and give you the way to start over again and again. Thus, I warmly thank my mom, dad, brothers, grandmother and all my family for having supported me along these tough years with love and patience.

Finally, I would love to warmly thank all those people who joined and left my life. I have always believed that we are today what our social interactions were. Each person that somehow influences our life – unconsciously – is giving us the precious opportunity to change something of our way to think and, sometimes, these changes could lead you toward new perspectives and solutions. In any case, are they bad persons rather than angels; they bring entropy in your life. So, thanks to everyone who has provoked me a smile or a tear.

Contents

Preface

Classical Molecular Dynamics (MD) simulations, based on Newtonian physics to simulate atomic motions, are currently the mainly used approach to capture the dynamics of biological systems with atomistic precision. This approach is herein used to investigate the physical step of human DNA Pol-η catalytic cycle. We unraveled a highly cooperative mechanism, which is here described by an equilibrium ensemble of structures (i.e. Pol-η X-ray crystals) that connect the reactants to the products in Pol-η catalysis. This dynamic mechanism clarifies how Arg61 (a residue invariantly conserved in human Pol-η) and multiple metal ions jointly catalyze DNA editing, and allows decoding the conformational transitions disclosed by the recent experimental data on Pol-η.

Despite the success of classical MD simulations, this technique suffers of an intrinsic limitation related the use of empirical force fields for the study of reactive processes in which chemical bonds are formed or broken. The modeling of such processes can only be studied by quantum mechanics (QM) that permit to investigate formation/breakage of chemical bonds. QM methods aim to solve the Schrödinger equation based on first–principles only, and no empirical data are used differently from MD simulations. This approach has been here adopted to investigate the chemical step performed by Pol-η to incorporate an incoming nucleotide on the growing DNA primer strand.

The studies on the DNA Pol-η mechanism of action discussed in this thesis were performed in collaboration with the group of Prof. Matteo Dal Peraro who hosted me in his laboratory at the Swiss Federal Institute of Technology in Lausanne – CH (EPFL) for 6 months and with the group of Prof. Paolo Carloni who hosted me in his laboratory at the Forschungszentrum Jülich – DE (FZJ) for 12 months.

Thesis outline

This thesis is divided into 5 chapters as follows:

- **Chapter 1.** A general introduction on DNA Polymerases biological function, their catalytic mechanism and pharmaceutical relevance.
- **Chapter 2.** Theoretical methods employed (i.e. classical MD and hybrid QM/MM simulations) are described.
- **Chapter 3.** We focus our attention on the human DNA Polymerase-η, a Y-family member strictly linked to melanoma onset. Classical MD have revealed a crucial role an invariantly conserved positively charged residue within the active site.
- **Chapter 4.** Here we investigate the chemical step performed by Pol-η to elongate DNA primer strand.
- **Chapter 5.** Given the conservation of different structural determinants among DNA Pols from all kingdom of life, we discuss the extension of our finding to the entire Pols superfamily.

Chapter 1

Human DNA Polymerases: biological function and therapeutic relevance

1.1 Enzyme-mediated Processing of DNA and RNA

To accomplish vital tasks such as development, growth and proliferation each living cell has to rely on crucial DNA transactions such as replication, recombination or repairing. All these transactions imply a constant DNA-editing process that is fundamentally based on two main chemical transformations: degradation and formation. As such, these transformations are thermodynamically disfavored under ambient conditions because of the prominent amount of energy required to ensure their accomplishment.[1] In this scenario, enzymes (biological catalysts) increase the rate of those chemical transformations by lowering their activation energy (i.e. minimum energy quantity required to start a chemical reaction) catalyzing therefore their completion. Thus, different and specialized enzymatic classes in prokaryotes, eukaryotes and viruses, by mediating those key chemical reactions, guarantee cell's proficiency in storing, retrieving and transmitting genetic information.

Firstly, to access the genetic information, nuclear double strand DNA (dsDNA) – normally found in an overwound state termed chromatin – must be unwound.[2] This is performed by a set of specific enzymes such helicases and topoisomerases that separate double strand DNA in different manner. Helicases progressively unzip the DNA double helix by disrupting the interbase hydrogen-bonds characteristic of the C:G and A:T (U:T in RNA) pairing model revealed by Thomas Watson and Francis Crick in 1953. Since helicase activity requires energy to be accomplished, they often hydrolyze ATP (adenosine triphosphate) molecules into ADP (adenosine diphosphate) and phosphate to obtain that free energy amount to sustain their activity. Thus, helicases that make use of ATP hydrolysis are defined as an ATP-dependent helicases.[3] Differently, topoisomerases relax dsDNA topology by transiently breaking/forming the sugar-phosphate backbone of nucleic acids in an ATP-independent manner.[4] To do that, they use multiple catalytic metal ions, which assist the cleavage/formation of the DNA sugar-phosphate backbone. Once topoisomerase activity is occurred, the dsDNA is relaxed and ready to be replicated, transcribed or differently manipulated.[1]

DNA replication and transcription are processes mainly based on the de-novo synthesis of DNA/RNA strands. Whether DNA/RNA formation serves to replicate, repair or transcript DNA into messenger-RNA (mRNA), DNA/RNA polymerases (Pols) are the main players in catalyzing the chemical reaction leading to the formation of new nucleic acid strands. DNA Pols are mainly divided into different families based on their sequence identity.[5] Nevertheless, they are grouped two main classes: the so-called high-processivity DNA Pols that ordinary replicate the undamaged dsDNA with efficiency constant of about ~300 base pair (bp) per second ($k_{eff} = k_{cat}/k_M$ ~6.5 $mM^{-1}s^{-1}$)[6] and the low-processivity Pols that process ~15 bp per second ($k_{eff} = k_{cat}/k_M$ ~0.03 $mM^{-1}s^{-1}$).[7,8] Importantly, the latter class is principally involved in repairing damaged DNA.[5]

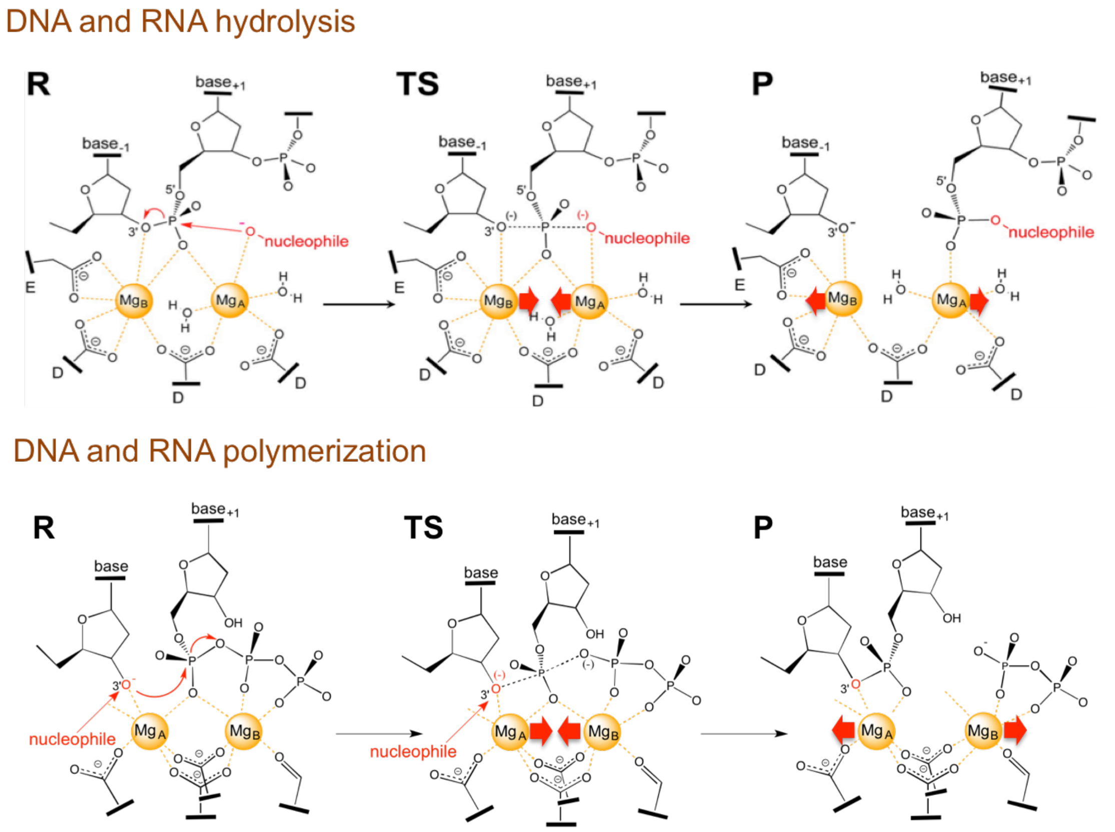
In fact, DNA damages, caused by toxins, radiation, and endogenous/exogenous metabolites, represent a major obstacle for the functionality of the native DNA replication machinery.[9] Indeed, the presence of a chemically modified DNA base or a DNA strand break could cause the stalling or the collapse of the replication fork that would lead cells to death. For instance, daily sunlight exposure can cause the formation of ultraviolet (UV)-induced covalent bonds between adjacent pyrimidine bases within the double helix.[10] This results in the formation of the so-called cyclobutane pyrimidine dimers (CPDs), one of the major forms of DNA damage.[11] Since high-processivity DNA Pols do not handle CPDs, one way to restore the DNA replication machinery and guarantee cell's life is to bypass those hurt regions via the so-called translesion synthesis (TLS).[9,12] TLS is performed by a set of Pols belonging to the Y-family polymerases (such as Pol-η) capable to extend the CPDs-damaged DNA.[13] Once base pairing has been restored beyond the lesion, the replicative polymerase regains the control of DNA replication process. Thus, Y-family Pols are crucial enzymes for life given their pivotal contribution to cell survival by promoting DNA damage tolerance and genome stability.[14-16]

Hence, the plethora of physicochemical transformation underwent by nucleic acid is broad and alterations of these processes could severely impact on cell function and life. Recently, deregulation of those enzymes has been linked to dramatic diseases spanning from neurodegenerative diseases such as X-linked-α-thalassemia-mental retardation (ATR-X syndrome) or Cockayne syndrome to cancer onset.[3,5,17] Hence, Pols have recently become targets of pharmaceutical interest to develop different therapeutic strategies to treat cancer, viral and bacterial infections and neurodegenerative diseases. Extending our comprehension on how enzyme-mediated processing of nucleic acids strand is performed is therefore vital for drug discovery efforts.

1.2 Polymerases operate through the two-metal-ion mechanism

In 1993, a seminal paper of Steitz and Steitz suggested that two metal ions are essential for RNA splicing process [i.e. intron excision to obtain mature messenger-RNA (mRNA)] catalyzed by self-cleaving ribozyme (RNA enzymes).[18] The chemical reactions at the basis of such convoluted process are the (i) breakage of phosphodiester bond linking two consecutive nucleobases and (ii) the formation of a new phosphodiester bond to join them with other parts of the RNA assembly yielding the mature form of mRNA. Interestingly, the ribozyme-mediated RNA splicing process is similar to that of the enzyme-mediated DNA cleavage albeit the different biological scope.[19] This proposal was based on the crystal structures of two DNA enzyme-substrate complexes: alkaline phosphatase and the exonuclease domain of Klenow fragment (active part of DNA Pol-I). The active-site architecture of these two protein/RNA enzymes consists of conserved carboxylates (i.e. DED-motif) coordinating two catalytically competent Mg^{2+} ions, namely MgA and MgB. These are normally found ~ 4 Å apart and assist the proper coordination of the sugar-phosphate backbone of the DNA substrate to efficiently generate the Michaelis-Menten complex.

Observing the particular coordination geometry of both Mg ions, Steitz and Steitz proposed that MgA reduces the pKa of a water molecule (mediating its deprotonation) to trigger nucleophilic attack on the scissile phosphate via the formation of a hydroxide ion. On the other hand, MgB has been proposed to stabilize the pentacovalent intermediate species (i.e. typical S_N2-like phosphoryl-transfer transition state, TS in Scheme 1) while contextually facilitates departure of the 3′ oxyanion group (see Scheme 1, hydrolysis).[20] Thus, RNA and DNA hydrolysis occurring via two divalent metals has been proposed as a general mechanism and has been defined as "2M-mechanism". This general 2M-mechanism for nucleic acids hydrolysis has been corroborated and further dissected by several X-ray structures of different nucleic acid-processing metalloenzymes during the following years.[1,13,21] For example, the X-ray structures of ribozymes such as group-II introns, ribonuclease H and several topoisomerases show a common metal-binding site, despite differences in the protein tertiary structure and cellular function.



Scheme 1. Reaction mechanisms at the basis of nucleic acid hydrolysis (upper part) or polymerization (lower part). Catalytic metals are depicted in orange while the attacking species in red. In both cases the reaction is the S_N2-type phosphoryl-transfer. Both metals actively participate in catalysis by shortening in transition-state or increasing (product state) their internuclear distance in order to alleviate the highly-charged environment raised up in transition state or to promote departure of leaving group in product state.

However, a very large part of nucleic acid-editing processes imply the formation of a phosphodiester bond between free-floating nucleotides leading to the de-novo formation of new DNA/RNA strands. This process is catalyzed by a complex and finely tuned replication machinery in which DNA- and RNA-Polymerases (Pols) are key players. Over the last years, supported by the constant increasing high-resolution structural data, it has been observed that all DNA and RNA Pols from all kingdom of life operate through the two-metal (Mg^{2+})-ion mechanism to catalyze the incorporation of an incoming nucleotide triphosphate [(d)NTP] into the growing nucleic acid strand (i.e. primer strand). As well as for nucleic acid hydrolysis, also DNA/RNA elongation occurs via the typical S_N2-like phosphoryl-transfer reaction albeit in this case a pyrophosphate (PP_i) molecule is generated as leaving group at each enzymatic cycle. The two-metal-aided phosphoryl transfer reaction for nucleotide addition in Pols is preceded by deprotonation of the 3′-hydroxyl (3′-OH) of the 3′-end deoxyribose (see Scheme 1, polymerization). This generates the activated nucleophilic 3′-hydroxide ion. In Pol's catalysis, the formation of the 3′-hydroxide ion is the very first chemical step to trigger a nucleophilic attack on the incoming nucleotide, which is bound to the enzyme thanks to a large conformational change for Watson–Crick nascent base pairs, as explained well through this thesis (see Chapter 4).

Hence, the 2M-mechanism has been widely observed in numerous nucleic-acid processing enzymes performing both hydrolysis and polymerization of DNA and RNA strands. Intriguingly, other nucleic-acid processing metalloenzymes have been observed retaining more than two metal binding sites.[8,22,23] For instance, theoretical studies of endonuclease IV, an enzyme involved in DNA repair, have shown how three Zn^{2+} ions assist catalysis,[24] finding that structural rearrangements in this multinuclear enzyme's active sites are essential for reaction. Moreover, the restriction endonuclease EcoRV has been observed to possess three metal-binding sites whereas only two of them result critical in modulating its catalytic activity.[25] Multiple metal binding-site where also detected in diffeent Pols. Indeed, human Pol-β and Pol-η — both investigated by time-depended X-ray crystallography — have been observed to make use of a transient third Mg^{2+} ion appearing in the active site only once phosphoryl-transfer reaction occurred.[8,26] However, whether these functionally related enzymes work with two or three metal ions is still uncertain. Despite the difference in the number of observed ions in the X-ray structures, these observations suggest that a third metal ion may have a common role in enhancing the canonical 2M-mechanism.

1.3 Y-family Polymerases structure and function

As mentioned above, a finely-tuned and complex multimolecular machinery replicates DNA with very high efficiency and shocking fidelity. However, this multimolecular assembly is easily disturbed by damages on the DNA template that heavily distort the native nucleic acid topology. Indeed, despite the plethora of mechanisms for repairing damaged-DNA, it is likely that the replication machinery will encounters lesions in the DNA template during each cell division cycle. The most important part of the replication machinery is represented by Pols that directly extend the growing primer strand. However, the catalytic site of the high-processive DNA polymerases is narrow and is not prone to accommodate the vast majority of DNA lesions.[5,6,27] As a consequence, DNA synthesis arrests at most forms of DNA damage. Such arrest poses a vital problem for cells because they must replicate the damaged DNA before mitosis in order to integrally copy the genome to transmit to both daughter cells. In this scenario one possible solution is known as DNA damage tolerance process. Here, the DNA is synthesized past the damaged bases, which can be subsequently excised after the replication is occurred and they are safely located within duplex DNA. Direct replication past DNA damage in these circumstances — a process known as translesion synthesis (TLS) — is mediated by specialized DNA Pols, the most abundant class of which are those belonging to the Y-family (see Fig. 1).[9,13,14,16]

Y-Family polymerases are divided into six major groups based on their own amino acid sequence. They are represented by E. coli DNA Pol-IV (also known as DinB) and Pol-V (known also as UmuC) and four human enzymes: Pol-η, Pol-ι, Pol-κ, and Rev1. All Y-family members are constituted by two functional parts, (i) the polymerase catalytic region made of 350–500 residues and (ii) by a regulatory region spanning from 10 to 600 residues. Among Y-family members, fingers, palm and thumb domains directly participate in the formation of the two-metal-centered active site (see Fig. 2).[15,16,28] The double-strand DNA (dsDNA) is accommodated between the palm and the little finger domain, with the sugar-phosphate backbones of both the primer and template strands being contacted across the minor groove by the thumb and across the major groove by the little finger domain. The 5′-end DNA strand (i.e. template) is channeled

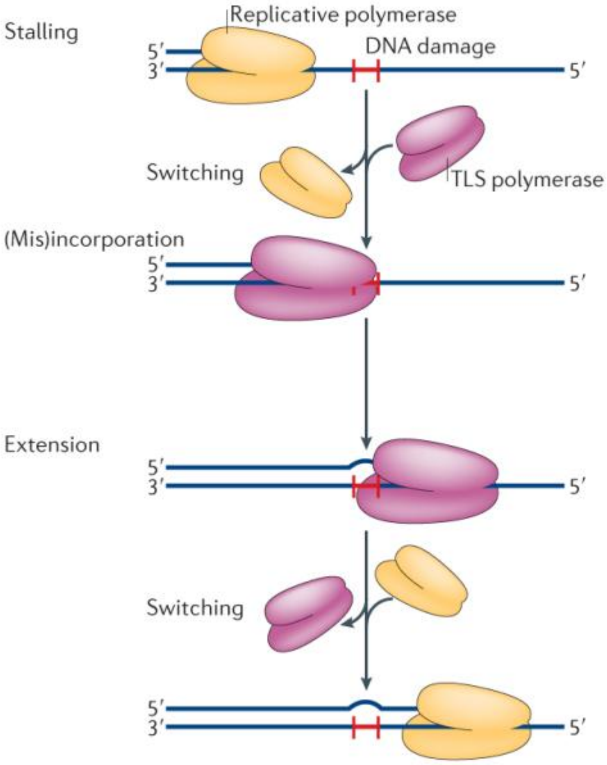


Figure 1. (From Sale, E. J., et al., Nat. Rev. Mol. Cell. Biol., 2012, 13, 141-152) Translesion synthesis (TLS) is a multi-step process. The replicative DNA polymerase stalls at a site of DNA damage (red `H'), then a TLS polymerase (or polymerases) is recruited to the primer terminus and correct, or incorrect, bases are inserted opposite the lesion. The inserted base is subsequently extended to complete TLS. All TLS polymerases are distributive and dissociate from the primer terminus after synthesizing a short tract of DNA. They are then quickly replaced by the cell's replicative polymerase, which completes genome duplication.

into the polymerase active site through contacts with the fingers and/or little finger domain (see Fig. 2).[28] Despite the high degree of structural similarities, each Y-family member shows outstanding substrate specificity towards specific DNA lesions. For instance, UmuC and its eukaryotic analogue Pol-η, primarily bypass UV-induced lesions such as cyclobutane pyrimidine dimers (CPD) one of most widely diffused DNA lesion.[29] It has been recently shown that Pol-η binds CPD-containing DNA with higher specificity than native DNA and extends primer strands more processively once CPDs are located in the active site.[15]

Unfortunately, Pol-η bypasses also DNA crosslinking damages generated by anticancer drugs, such as cisplatin, allowing cancer cells to survive and proliferate.[15,27,30,31] Indeed, it has been reported that the expression level of Pol-η is further induced by cisplatin treatment.[30] Such an increased Pol-η expression level further reduces the effectiveness of chemotherapy and the survival time of patients with non–small-cell lung cancer or metastatic gastric adenocarcinoma.[32] Moreover, a breaking through work has been reported that Pol-η is yet the only one of the fifteen human DNA polymerases in which defects are unequivocally associated with cancer development.[8] In fact, mutations of the human POLH gene, which encodes Pol-η, are directly linked to a variant form of the cancer predisposition syndrome xeroderma pigmentosum (XP-V).[29,33] XP-V patients are deficient in TLS of UV-induced CPDs and are hypersensitive to sunlight. Taken together, these findings point up Pol-η as a promising pharmaceutically relevant target to treat skin carcinogenesis and overcome cancer drug resistance in other kinds of solid tumors.

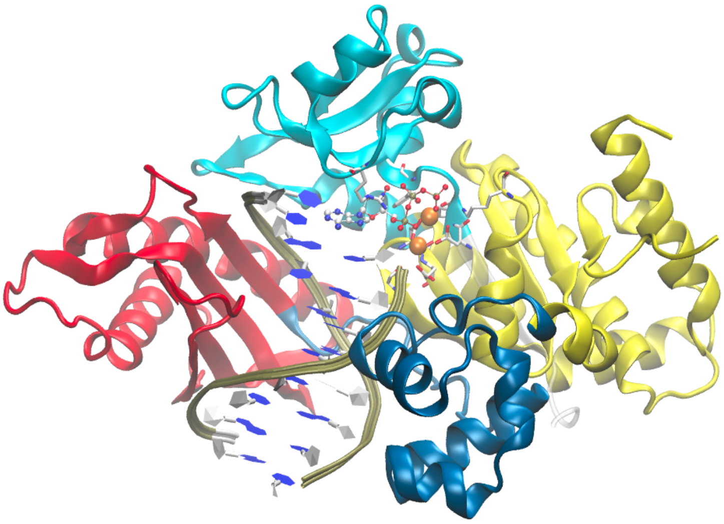


Figure 2. Human Pol-η structural features. Graphic of the ternary Pol-η/dsDNA/dATP complex. Protein in graphic representation, dsDNA in ribbon representation; orange spheres indicate the two catalytic Mg2+ ions. The following domains are depicted: palm (yellow), thumb (blue), fingers (cyan) and little finger (red). All the Y-family members possess the same subdomain arrangement.

1.4 Convergent evolution has preserved key structural determinants to guarantee enzyme-mediated processing of DNA and RNA

Over the years, new X-ray structures have revealed additional important catalytic elements among different two-metal-ion enzymes. For example, in human DNA Pol-η and human Pol-β a third transiently-bound divalent cation (namely MgC) has been detected in the two-metal-centered active site immediately after the transition state (TS) of phosphoryl-transfer reaction for nucleotide incorporation.[8,22,26] This transient cation has been investigated by both experimental and theoretical studies and hyphothesis put forward are competitive and still elusive. Wei Yang and collaborators have suggested that in Pol-η, DNA synthesis does not occur solely via the canonical 2M-mechanism.[22] By using time-resolved X-ray crystallography, they have shown that the phosphoryl-transfer reaction starts only after the ES-complex (enzyme + substrate) recruits a third divalent cation that is found orienting α- and β-phosphate groups of the incoming nucleotide.[8] On the other hand, Samuel Wilson and co-workers have initially proposed that this transiently bound MgC serves to trigger pyrophosphorolysis reaction (reverse process to DNA synthesis) by which Pols are able to excise wrong paired nucleotides reconstituting, from products, the triphosphate form of the nitrogenous base.[23] In any case, the necessity of a third divalent metal ion for catalysis is hard to discredit on the basis of these recent studies due to their controversial conclusions.

Contextually, a second intriguing shared structural feature comes from a crystallographic work focused on the reaction mechanism performed by ribozymes such group II introns. Recent crystal structures of self-splicing group-II intron ribozymes revealed that, besides MgA and MgB forming the canonical 2M-architecture, two conserved potassium ions (K1 and K2) are essential for catalysis.[21,34] Interestingly, the group-II intron active site is strikingly similar to that of type-II restriction endonuclease BamHI (see Fig. 3), suggesting the existence of enzymatic amino-acid counterparts of K1 and K2. Such structural analogy between two-metal-ion protein and RNA enzymes would infer a more extended set of functional components instrumental for proficiently process DNA and RNA via a multi-metal-centered active site. This intriguing similarity between RNA- and protein-enzyme represents an appealing example of convergent evolution and emphasizes the fact that catalytic strategy is universal and independent of biopolymer scaffold.

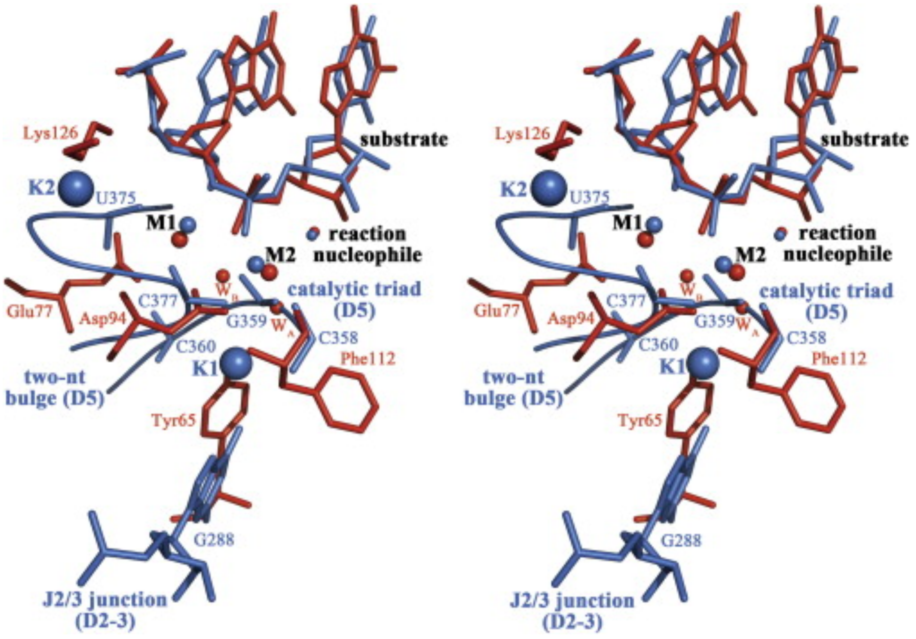


Figure 3. (From Marcia, M. and Pyle, A. M., Cell, 2012, 151, 3, 497-507) Structural convergence between group-II intron and protein endonuclease BamHI. Stereo representation of the intron active site (K^+/Ca^{2+} oligonucleotide-bound structure, blue) superimposed over the R active site of endonuclease BamHI (PDBid.: 2BAM, red). The carboxyl groups of amino acids Glu77 and Asp94 occupy the same location as the phosphates of nucleotides U375 and C377 (D5 bulge), the carbonyl oxygen of Phe112 and other solvent molecules (W_A and W_B) replace the phosphates of C358-G359 (D5 catalytic triad), Lys126 is analogous to the K2 ion, and Tyr65 plays the same role as the K1 ion. As a consequence, the catalytic ions M1 and M2 and also the reaction nucleophile superimpose precisely.

1.5 Scope of the present thesis

Despite the wealth of structural and biochemical data of DNA/RNA Pols, many details concerning Pols structure, function and catalysis are still elusive and much debated. First, by coupling classical MD simulations and enhanced sampling techniques, we will structurally and energetically investigate human Pol-η ternary complex (i.e. Pol-η/dsDNA/dNTP) containing two Mg^{2+} in the active site. Briefly, we will discuss the role of a present flexible and positively-charged residue within the active site that has been shown to actively participate in the phosphoryl-transfer reaction for nucleotide addition. Later on, using the same computational techniques, we will analyze the role of the transiently-bound third metal ion initially detected by time-dependent crystallography in postreactive state of human Pol-η and Pol-β. On the basis of our results, we propose of our findings to entire Pols superfamily.

Second, once having investigated the plasticity of human Pol-η, we will focus our attention on the chemical step that the enzyme catalyzes. In particular, by using ab initio simulations and metadynamics in the context of Car-Parrinello simulations, we will capture atomistic details of the (i) activation of the nucleophilic specie (i.e. 3'OH end of the primer DNA strand), a mandatory step to trigger the phosphoryl-transfer reaction for nucleotide incorporation, (ii) the role of the invariantly conserved residue (i.e. Arg61) along catalysis and (iii) the energetic characterization of the S_N2-type reaction performed by Pol-η. Notably, having also observed a previously unrecognized structural element invariantly present among all known X-ray structures of Pols ternary complexes, we will propose a general reaction mechanism for nucleotide addition performed by DNA and RNA Pols coming from all kingdoms of life.

Lastly, intrigued by the high structural similarities recently found between group-II introns and endonuclease BamHI active sites, we used multiple sequence and structural alignments, molecular modeling and electrostatic potential maps to analyze a large number of DNA/RNA-processing enzymes. Since all oh them use 2M-mechanis, we will discuss how 2M-architecture is transversally adopted by nucleic acid-editing catalysts despite their different biopolymer scaffolds, size and biological function.

Chapter 2

Methods and Materials

2.1 Molecular Dynamics (MD)

Molecular dynamics (MD) simulation is a computational technique that allows investigating atoms movements by solving classical equation of motion [2.1]. Thus, being Newtonian equation time-dependent, MD simulations provide a time-depended description of particle motion of a system under investigation. This calculation returns a trajectory representing the computed configurations of the molecular system as a function of time. Contextually, it provides insights concerning the dynamic and thermodynamic properties of the system under investigation. The method was first developed more than 50 years ago[36] and, given its accuracy in describing diverse chemical and physical problems, it is in continuous development.[35,36] To describe the time evolution of a molecular system by MD simulations, Newton's equation of motion [2.1] has to be integrated.

$$F_i = -\frac{\partial V}{\partial R_i} = M_i \frac{d^2 R_i}{dt^2} \qquad [2.1]$$

F_i is the force acting on atom i with position R_i and mass M_i, while V represents the potential energy of the system. Since there is no analytical solution for integrating Eq. [2.1], numerical algorithms based on time discretization have to be used. A frequently used integration algorithm is the velocity-Verlet algorithm, which uses Taylor expansion truncated beyond the quadratic term of the coordinates.

$$R(t+\Delta t) = R(t) + v(t)\Delta t + \frac{F(t)}{2M}\Delta t^2 \qquad [2.2]$$

The update for the velocities is given by:

$$v(t+\Delta t) = v(t) + \frac{F(t+\Delta t)) + F(t)}{2M}\Delta t \qquad [2.3]$$

The size of the timestep depends on the characteristic dynamical time scale of the system and spans typically between 0.1-0.5 fs for ab initio MD to 1-2 fs for classical MD. The thermodynamic state of a system is defined in terms of macroscopic parameters that are constant during MD simulation. In a microcanonical ensemble the number of particles N, the volume V and the energy E are fixed and a constant energy ensemble (NVE) is sampled.[37] When the MD time approaches infinity, the system has sampled all possibile configurations Γ (ergodic hypothesis) and statistical thermodynamics relates the ensemble average to the true average of an experimentally measurable quantity.

$$A_{obs} = \lim_{T\to\infty} \frac{1}{T} \int_0^T A(\,\Gamma(\mathrm{t}))dt \qquad [2.4]$$

Most chemical and biological processes occur at constant temperature and constant pressure. Therefore, by using thermostat or/and barostat algorithm the canonical ensemble (NVT) or the isothermal-isobaric ensemble (NPT) can be sampled in MD simulations.[37]

2.2 Classical MD

In classical MD simulations the potential energy is determined by an empirical force field (FF) in order to faithfully reproduce experimental or ab initio data. In FF-based methods, the potential energy is calculated as a parametric function of the nuclear while to determine the electronic energy for a given nuclear configuration of an atomic system, quantum mechanical calculations are required. Several force fields have been developed for biomolecular applications such as AMBER,[38] GROMOS[39] and CHARMM.[40] In this thesis, we use the AMBER/parm99,[38] and its parmbsc0 modifications for nucleic acids.[41] Its pairwise-additive potential is of the form:

$$V = \sum_{bonds} K_R(R-R_{eq})^2 + \sum_{angles} K_\theta(\theta-\theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi-\gamma)] + \sum_{i<j}[\frac{A_{ij}}{R_{ij}^{12}} - \frac{AB_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}] \qquad [2.5]$$

The interactions are divided into bonded interactions, acting within a molecule, and non-bonded interactions, acting between all atoms except from bonded neighbors. E. [2.5] contains three terms for the bonded interactions, namely one for the chemical bonds between two neighboring atoms, one for the bond angles between three atoms and one for the dihedral angles between four consecutive atoms. Additionally, improper dihedral-angle terms can be applied to maintain planar or tetrahedral conformations. The functional form of the bond and angle terms is quadratic, while the dihedral term uses a trigonometric function. In Eq. [2.5] R is the distance between two atoms i and j that are bound together via a covalent bond, R_{eq} and θ_{eq} refer to equilibrium bond lengths and angles, K_R and K_θ are the vibrational constants. V_n is the torsional barrier corresponding to the n^{th} barrier of a given torsional angle with phase γ. The last term in Eq. [2.5] refers to the non-bonded interactions, which are composed of a Lennard-Jones term for the van-der-Waals interactions and a Coulomb term for the electrostatic interactions between atoms i and j that are involved in direct (1-2) or indirect (1-3, 1-4) bonded interactions. These are excluded from the potential energy of system that therefore is rescaled. Eq. [2.6] represents the van-der-Waals term that describes the repulsive forces at small interatomic distances due to the Pauli repulsion between electrons (decaying exponentially, modeled with R_{ij}^{-12}), and the attractive forces at intermediate distances due to instantaneous dipole-induced dipole interactions (modeled with R_{ij}^{-6}).

$$V_{vdW} = \sum_{i<j} [\frac{A_{ij}}{R_{ij}^{12}} - \frac{AB_{ij}}{R_{ij}^{6}}] \qquad [2.6]$$

The Coulomb term [2.7] takes into account the electrostatic interactions between charged particles.

$$V_{el} = \sum_{i<j} [\frac{q_i q_j}{\epsilon R_{ij}}] \qquad [2.7]$$

The calculation of the electrostatic Coulomb interactions between charged particles is computationally the most demanding operation as it represents a long-range interaction and converges particularly slowly. In the AMBER/parm99 FF, the electrostatic properties of the electron density of the system are reproduced by atom-centered point charges derived from quantum chemical calculations within a restrained electrostatic potential (RESP) fitting scheme.[42] Non-bonded interactions represent a major part of the computational cost and their treatment requires special care. To start with, in order to avoid artifacts due to finite-size effects, the boundary is usually completely removed by treating the system with periodic boundary conditions (PBC). This places the system in an environment similar to an infinite solution. The most widely used algorithm for a fast and accurate treatments of the Coulomb term employs the exact periodicity of the system and includes all electrostatic interactions (also with periodic copies) by lattice sum techniques based on Ewald summation, among which the particle mesh Ewald (PME) method is widely used.[43] Despite the high reliability of recent FFs, they suffer from intrinsic limitations like limited transferability, neglect of polarization and their inability to describe chemical bond breaking and formation. Classical MD simulations have been widely applied in this thesis, as reported in Chapter 3.

2.2.1 Thermostats

In order to obtain a realistic reproduction of the biological conditions, constant temperature can be obtained by coupling the system to a thermostat. The main used methods are: (*i*) the weak coupling method (Berendsen thermostat)[44] and (*ii*) the coupling to an external bath (Nosé-Hoover thermostat).[45] In the Berendsen thermostat, the temperature of the system is kept close to the target value by the equation [2.8] where T is the instantaneous value of the temperature and T_T is the temperature coupling time.

$$\frac{dT}{dt} = \frac{1}{T_t}[T_0 - T(t)] \qquad [2.8]$$

The instantaneous temperature T of a system with N_{df} degrees of freedom in related to the kinetic energy $E_{kin}(t)$:

$$E_{kin}(t) = \sum_{k=1}^{N} m_k \dot{q}_k^2(t) = \frac{1}{2} N_{df} k_b T(t) \qquad [2.9]$$

The atomic velocities can be scaled by a factor λ:

$$\delta E_{kin}(t) = [\lambda(t))^2 - 1]\frac{1}{2} N_{df} k_b T(t) \qquad [2.10]$$

Here, the λ(t) factor is applied to scale the velocities $\dot{q_k}$ at each step of integration, so to relax the temperature towards the target value T_0. The time coupling constant T_T controls the relaxation rate.

The Nosé-Hoover thermostat extends the ensemble by introducing a thermal reservoir and a fiction term in the equations of motion. The friction force is proportional to the product of the velocity of each particle and fiction parameter ξ, which corresponds to a dynamic quantity featured by its own equation of motion. The time derivative is calculated from the difference between the current kinetic energy and the reference temperature. Thus, the equation of motion of each particle becomes [2.11], whereas the equation of motion for the bath parameter becomes ξ [2.12]. Here T_0 is the reference temperature and Q controls the strength of coupling to the heat bath.

$$\frac{d^2 r_i}{dt^2} = \frac{F_i}{m_i} - \xi \frac{dr_i}{dt} \qquad [2.11]$$

$$\frac{d\xi}{dt} = \frac{1}{Q}(T - T_0) \qquad [2.12]$$

2.2.2 Barostats

The pressure of the system can be controlled by coupling the system to a barostat as done in the Berendsen approach.[44] This method scales the coordinates and the box vectors, obtaining a first-order kinetic relaxation of the instantaneous pressure *p(t)* toward the reference pressure p_0 with a time constant τ_p. Analogously to the Berendsen thermostat:

$$\frac{dp(t)}{dt} = \frac{p_0 - p(t)}{\tau_p} \qquad [2.13]$$

the scaling factor μ(t) for atomic position is:

$$\mu_{\alpha\beta} = \delta_{\alpha\beta} - \beta_{\alpha\beta} \frac{\Delta t}{3\tau_p}[p_{0,\alpha\beta} - p_{\alpha\beta}(t)] \qquad [2.14]$$

where β is the isothermal compressibility, *p(t)* is the instantaneous pressure, p_0 is the target pressure and τ_p is the pressure coupling time constant.

2.3 Quantum Mechanics (QM)

Classical mechanics do not allow the study of chemical reactions, since atoms and bonds are parameterized via an empirical FF and electrons are not taken explicitly into account. The quantum mechanical description allows the inclusion of electrons. In quantum mechanics (QM) each particle (e.g. the electron) is described by a wave function $\Psi(r, t)$, where $\Psi \cdot \Psi$ represents the probability of finding the particle at position r at time t (e.g., the electron density). To each classical observable (i.e., position or momentum), corresponds a QM operator that, acting on the wave function, returns the expectation value of this operator. The Hamilton operator was introduced by Erwin Schrödinger and allows calculating the energy of a system, formed of M nuclei and N electrons. The time-independent Schrödinger equation is [2.15].

$$H\Psi = E\Psi \qquad [2.15]$$

where the Hamilton operator can be divided in the terms:

$$H = T + V_{ext} + V_{ee} \qquad [2.16]$$

where T is the kinetic energy operator, V_{ext} is the external potential produced by the nucleic acting on the electrons and V_{ee} is the electron-electron potential. Substituting the classical form of observables with their QM operator, the precise form of the Hamilton is given.

$$x = \hat{x} \qquad [2.17]$$

$$p_x = \frac{\hbar}{i}\frac{\partial}{\partial x}$$

2.3.1 The variational principle

The wave-function of an interacting many-body system is not know *a priori*. Thus, the Schrödinger equation is calculated from a trivial wave-function Ψ_T, which is a linear combination of basis functions:

$$\Psi_T = \sum_n c_n \psi_n \qquad [2.18]$$

where ψ_n are wave-functions orthonormal and normalized, such that $\sum_n |c_n|^2 = 1$. In order to approximate the true ground-state wave-function, the variational principle is applied. Accordingly, "The expectation value of an Hamiltonian, $\widehat{\mathcal{H}}$, calculated with the trial wave-function Ψ_T is never lower in value than the true ground-state energy, ε_0, calculated with the true ground-state wave function Ψ_0".

2.4 Density Functional Theory

The main idea of the density functional theory (DFT), proposed by Hohenberg and Kohn in 1964, is that *the ground-state energy of a many-electron system is a unique functional of its electron density and the correct ground-state electron density minimizes this energy functional.* This idea represented a revolution for the ab initio methods; given the advantages in terms of computational efficiency and broad applicability for many chemical/physical problems.

DFT is a widely applied quantum chemical method for the investigation of biological systems. It scales favorably with the number of electrons and the accuracy of the employed exchange-correlation functionals, which contains all the intricacies of the many-body problems, is constantly improving.

2.4.1 Kohn-Sham DFT

In DFT the pivotal quantity is the electronic (spin) density that depends only on 3 spatial coordinates plus the spin coordinate in case we are referring on spin-polarize DFT. This provides a sensible simplification compared to wave-function based methods, in which the *N*-electron wave-function depends on 3 × *N* spatial plus *N* spin coordinates. The two Hohenberg-Kohn (HK) theorems,[46] describe how to determine the properties of a system on the basis of its electron density only. The first one states that the external potential V_{ext} of a system is (up to a constant) solely determined by the ground-state density n_0. Since $\hat{\mathcal{H}}$depends on V_{ext} and on the density n_0, the electronic ground-state and its relative properties are determined by n_0.

$$n_o \Rightarrow \{N, Z_A, R_A\} \Rightarrow \hat{\mathcal{H}} \Rightarrow \Psi_0 \Rightarrow E_0 \qquad [2.19]$$

The ground-state energy E_0 can be written as a functional of n_0.

$$E_0[n_o] = E_{Ne}[n_o] + T[n_o] + E_{ee}[n_o] \qquad [2.20]$$

that can be split into a system dependent part E_{Ne}, defined by the position R_A and charges Z_A of the nucleic through the external potential V_{ext} and a system independent part, containing the kinetic energy of the electrons $T[n]$ and the electron-electron interaction $E_{ee}[n]$.The system independent functional is called the universal or HK functional.

$$F[n] = T[n] + E_{ee}[n] \qquad [2.21]$$

The second Hohenberg-Kohn theorem provides a variational principle for the ground state of a system in order to determine the true ground-state density n_0. Bases on the variational principle, it states that the trial density $\tilde{n}$, which minimizes the ground-state energy functional E_0, is equal to the true ground state density n_0 of the system. Therefore, any density $\tilde{n}$, that satisfies specific boundary conditions, provides an upper bound to the true ground-state energy [2.22].

$$E[\tilde{n}] = \int V_{ext}(r)\tilde{n}(r)dr + F[\tilde{n}] \geq E_0 = E[\tilde{n}_0] \qquad [2.22]$$

In principle, by minimizing $E[\tilde{n}]$, we are able to search for the ground-state energy. However, the exact dependence of the electronic kinetic energy on the density is unknown and, therefore, no exact relationship between the non-classical electron-electron interaction and the density is known. The Kohn-Sham (KS) approach provides a way of evaluating the main unknown terms. In this approach, $T[n]$ and $E_{ee}[n]$ are split into two parts:

$$T[n] = T_S[\{\phi_i\}] + T_C[n] \qquad [2.23]$$

$$E_{ee}[n] = J[n]] + E_{ncl}[n] \qquad [2.24]$$

$T_S[\{\phi_i\}]$ is the electronic kinetic energy of a non-interacting reference system and E_{ncl}stands for the non-classical electron-electron interaction terms. The electrons of the non-interacting system are described by orbitals ϕ_i, satisfying [2.25] where it is assumed that the density of the reference system n is equal to the true density n_0 of the system (n=n_0).

$$\sum_{i}^{N} |\phi_i(r)|^2 = n(r) \qquad [2.21]$$

T_C is defined as the difference between the kinetic energy of the interacting system and T_S and raises due to the interaction of the electron including all quantum effects. By introducing non-interacting orbitals ϕ_i that integrate to n, T_S can be determined as follows:

$$T_S[\phi_i] = -\frac{1}{2}\sum_i^N \langle\phi_i|\nabla^2|\phi_i\rangle \quad [2.26]$$

$$J[n] = \frac{1}{2}\iint \frac{n(r)n(r')}{|r - r''|}drdr' \quad [2.27]$$

$J[n]$ in the second equation is the classical Coulomb energy.

The $E_{ncl}[n]$ and $T_C[n]$ in Eq. [2.23 and 2.24] are quantities unknown in the KS approach and relies on the assumption that these two are rather small compared to the other terms and that the can be therefore approximated. By defining the exchange-correlation functional $E_{xc}[n]$ as:

$$E_{ex}[n] = T_C[n] + E_{ncl}[n] \quad [2.28]$$

the HK functional becomes:

$$F[n] = T_S[\{\phi_i\}] + J[n]] + E_{xc}[n] \quad [2.29]$$

Given the approximation for E_{xc}, the ground0state energy E_0 can be obtained by minimizing the functional

$$\begin{aligned} E_0[n] &= E_{Ne}[n] + T_S[\phi_i] + J[n] + E_{xc}[n] \\ &= -\sum_i \int \sum_A \frac{Z_A}{|R_A - r|}|\phi_i|^2 dr \quad [2.30] \\ &-\frac{1}{2}\sum_i \langle\phi_i|\nabla^2|\phi_i\rangle \\ &+\frac{1}{2}\sum_i\sum_j \iint |\phi_i(r))|^2 \frac{1}{r_{ij}} |\,\phi_j(r')|^2 drdr'' + E_{xc}[n] \end{aligned}$$

Minimization of E_0 can be obtained by solving self-consistently a set of equations, for non-interacting electrons moving in an effective potential V_{eff} [2.31] that yields the optimal non-interacting orbitals. The optimized orbitals ϕ_i are called Kohn-Sham orbitals.

$$\left[-\frac{1}{2}\nabla^2 + V_{eff}\right](r)\phi_i(r)) = \epsilon_i\phi_i(r) \qquad [2.31]$$

The form of the effective potential V_{eff} is:

$$V_{eff}(r) = V_{xc}(r) + \int \frac{n(r')}{|r'' - r|}dr'' - \sum_A \frac{Z_A}{|R_A - r|} \qquad [2.32]$$

where:

$$V_{xc} = \frac{\delta E_{xc}[n]}{\delta n} \qquad [2.33]$$

2.4.2 Exchange-Correlation Functionals (xc-functionals)

In principle the DFT in the KS formulation is an exact theory, but in practice approximations have to be made for the unknown exchange and correlations (xc) functional. A large variety of xc-functionals exist based on different assumptions. The fist and simplest assumption is the Local Density Approximation (LDA). In the LDA the xc-correlations energy per particle is equal to the one of a uniform electron gas of the same density at every point in space with a given density $\rho(r)$. It is called LDA the approximation for $\varepsilon_{xc}(\rho)$ depends on the electron density at a given point only. E_{xc} is defined as:

$$E_{xc}^{LDA}[\rho] = \int \rho(r)\,\varepsilon_{xc}(\rho)dr \qquad [2.34]$$

where $\varepsilon_{xc}(\rho)$ is the sum of the exchange and correlation energy of electrons in a homogeneous electron gas of density ρ [2.35]

$$\varepsilon_{xc}(\rho) = \varepsilon_x(\rho) + \varepsilon_c(\rho) \quad [2.35]$$

with

$$\varepsilon_x(\rho) = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} \times(\rho)^{1/3} \quad [2.36]$$
$$\varepsilon_c(\rho) = \varepsilon_c{}^{VWN}(\rho)$$

For the correlation term ε_c, no analytical expression is available, but accurate values have been obtained from quantum Monte Carlo calculations.[47] LDA yields good results for solid state systems with slowly varying charge densities, but for more complex molecules the homogeneous electron gas approximation is in general too simple.

Better results are obtained with functionals that do not only depend in the local density $\rho(r)$, but also on the local gradient $\nabla\rho(r)$ of the density. These functionals are summarized under the name Generalized Gradient Approximations (GGA) and are usually divided into their exchange and correlation contributions.

$$E_{xc}^{GGA}[\rho] = \int \varepsilon_{xc}(\rho(\vec{r}), \nabla\rho(\vec{r}))d\vec{r} = E_x^{GGA} + E_c^{GGA} \quad [2.37]$$

The BLYP functional, widely applied for biological systems, use the exchange part due to Becke combined with the correlation part due to Lee, Yang and Parrinello (LYP). The Becke exchange functional is given by [2.38] where the β parameter is determined by a fit on exact Hartree-Fock data of several systems, and was fixed by Becke as β = 0.0042 *au.*[49,50]

$$E_x^{Becke}[\rho] = E_x^{LDA} - \int \rho^{\frac{4}{3}}(\vec{r}) \frac{\beta x^2}{1 + 6\beta x sinh^{-1}(x)} d\vec{r} \quad [2.38]$$

with:

$$x(\vec{r}) = \frac{|\nabla\rho(\vec{r})|}{\rho^{\frac{4}{3}}(\vec{r})} \quad [2.39]$$

The LYP correlation functional[48] has an analytical form which contains one empirical parameter and is not based on the uniform electron gas but was derived by Colle-Salvetti from an expression for the correlation energy of the He atom. The BLYP xc-functional has been used in DFT calculations presented in this thesis (see Chapter 4), as it has been shown to provide reasonable results for organic and biological systems, without being too much computationally demanding.

A main drawback of the GGA functionals is their poor description of long-range exchange and correlation effects. The introduction so-called hybrid functionals in which a certain amount of the exact non-local Hartree-Fock exchange is mixed with the GGA exchange-correlation functional ($E_{xc} = E_x^{exact} + E_c^{KS}$), has been an important step towards a higher accuracy of DFT calculations for molecules. The most popular hybrid functional is the Becke B3LYP.[49,50] However, its use for ab initio MD simulations is limited by the increased computational cost of calculating the two-electron integral of non-local Hartree-Fock exact exchange term. More accurate functionals are still under development and they constitute one of the main topics of research for the improvement of DFT.[51]

2.5 Ab initio MD (AIMD)

Force field-base MD simulations suffer from their lack to describe bond formation and bond breaking. Ab-initio MD (AIMD) simulations overcome this issue, as allow to derive the forces acting on the investigated system from electronic structure calculations, which are performed on-the-fly as the MD trajectory is propagated. AIMD simulations represent a straightforward parameter-free MD, where the equilibrium and transport properties can be calculated with the accuracy and reliability of an ab initio method. The main advantages of AIMD can be summarized are: (i) Bond formation and breakage can be calculated, (ii) effect of finite temperature on complex chemical processes are taken into account and (iii) non trivial reaction coordinates can be employed to describe the event under investigation.

2.5.1 Car-Parrinello MD (CPMD)

The method of Car and Parrinello[52] has been developed in the 80's obtaining a great success in the scientific community.[53,54] The Car-Parrinello (CP) method combines DFT and MD in an efficient scheme in which the Newtonian equations of motion for both wavefunction and atomic coordinates are integrated. Therefore, the electronic and nuclear degrees of freedom evolve simultaneously according to a modified set of classical equations of motion, in which a fictitious electron mass is assigned to the electronic KS orbitals ψ_i. Forces are evaluated both on the nucleic and the electrons applying the KS equations. The basic equation of the CP method is given by the Lagrangian $\mathcal{L}$ as a functional of the wave-function ψ_l (where *l* represents the electronic state) and the atomic position R_I (where *R* is the atom index):

$$\mathcal{L} = \sum_l \frac{\mu}{2} \int d^3r \left|\dot{\psi}_l\right|^2 + \frac{1}{2} \sum_I M_I \left|\dot{R}_I\right| - \varepsilon[\{\psi_l\}, \{R_I\}] \qquad [2.40]$$

where, μ is a fictitious electron mass that controls the motion of the wave-function, and $\varepsilon[\{\psi_l\}, \{R_I\}]$ includes the electronic potential and the classical interaction between the nuclei. In the CP scheme, the system has no energy dissipation and the total energy, including the fictitious kinetic energy (KE) $\mu_l|\psi_l|^2$ of the electron is conserved. Therefore, the atomic oscillation f this system is preserved permanently without damping. The equation of motion for the wave-function are derived from CP $\mathcal{L}$:

$$\mu \frac{d^2}{dt^2} \psi_l = -\mathcal{H}\psi_l + \sum_v \Lambda_{lv} \psi_v \qquad [2.41]$$

where the Λ_{lv} are the Lagrange multipliers that ensure the orthogonality of the wave-function ψ_l, determined to orthonormalize the wave-functions ψ_l. To ensure the Born-Oppenheimer (BO) adiabatic approximation in the CP formalism, the inertia parameter μ assigned to the orbital degrees of freedom can be tuned in order to be small enough to ensure adiabaticity. However, even with an appropriate choice of μ, the electronic wave-function oscillate rapidly around the BO surface, following the nuclear motion. This rapid oscillation of the electronic states is an artifact but, for sufficiently small μ values, the amplitude of this oscillation is very small.

Importantly, the nuclear motion follows the Newtonian equations:

$$M_i \frac{d^2}{dt^2} R_I = \nabla_I E \qquad [2.42]$$

where E is the sum of the Coulomb potential between the nucleic and the total energy of the electron system for nuclear positions (i.e., the physical total energy + $\sum(\mu/2)|\psi_l|^2$) which is conserved. Therefore, obtaining the force on the right hand side of Eq. [2.42], this equation can be integrated to follow the electronic and nuclear degrees of freedom "on-the-fly" as the MD trajectory is generated.

2.6 Hybrid Quantum Mechanics/Molecular Mechanics QM/MM

Hybrid QM/MM techniques have become very famous during the last years.[53-55] In these schemes the biological system of interest is divided into two parts. The region of the biological system in which the chemical event takes place (the active site of a protein or crucial DNA/protein residues) is treated at the ab-initio level (QM–part) while the remaining part of the system (MM–part) is treated by parameterized Hamiltonians. This methodology dates back to the work of Warshel and Levitt in 1976,[56] and became popular due to significant improvements in the QM/MM interface and fast growing of the computational power. In this thesis we use the method developed in the group of Prof. U. Rothlisberger,[57] in which the QM–part is treated at the DFT/BLYP level of theory, whereas the MM–part is treated with the AMBER FF. Car-Parrinello and Born–Oppenheimer MD have been performed as implemented in the CPMD program and the CP2K program (freely available at the URL http://www.cp2k.org, released under GPL license), featuring an efficient interface[58] between the QM program QUICKSTEP and the MM–MD driver FIST. Details are reported in Chapters 3 and 4.

In QM/MM approach, the Hamiltonian $\mathcal{H}$ for a hybrid system contains Hamiltonians describing the quantum ($\mathcal{H}_{QM}$) and the classical ($\mathcal{H}_{MM}$) systems and that describing the interacting part between the QM and MM regions ($\mathcal{H}_{QM/MM}$).

$$\mathcal{H} = \mathcal{H}_{QM} + \mathcal{H}_{MM} + \mathcal{H}_{QM/MM} \qquad [2.43]$$

Special attention has to be done on the treatment of the QM and MM interface region. The definition of the $\mathcal{H}_{QM/MM}$ part of the Hamiltonian has been challenging. Here, we apply the scheme developed by Laio et al.,[57] in which bonded, non-bonded and long-range electrostatic interactions are treated as explained below.

2.6.1 Bonded interactions

The bonded interactions at the QM and MM interface are included in the classical Force Field, as the stretching, bending and torsional terms. Although these terms allow the stability of the system at the QM/MM interface, their parameterization does not take into account changes due to chemical reactions. Therefore, if the QM/MM interface region undergoes crucial distortions upon chemical rearrangements, these terms could lack in accuracy. Importantly, QM atoms directly linked to MM atoms by chemical bonds which are left with unsaturated valence orbitals. Different methods are available to saturate these valence orbitals. The simplest way is to use a link-atom pseudopotential. In this case it is required to constrain the distance between the link atom and the QM neighbor atom to the QM equilibrium distance, to preserve the electronic structure in the center of the QM subsystem. This is a rather crude approximation in which the constraint creates a small imbalance in the forces between the QM and MM sub-systems that can result in a drift in the total energy, if the length of the constraint is badly chosen. A more rigorous approach would be to use an optimized pseudopotential constructed with the method described in von Lilienfeld et al.[59] that should take care of the need for the constraint. An alternative method consists in the use of "capping" dummy–hydrogen atoms, which can fill the valences, but their interactions with the MM part have to be excluded.

2.6.2 Non-bonded interactions

The non-bonded Hamiltonian takes into account the electrostatic and steric contributions between non-bonded atoms:

$$\mathcal{H}_{QM/MM}^{non-bonded} = \sum_{i \in MM} q_i \int \rho(r) v(|r - R_i|) dr + \sum_{i \in MM} \sum_{j \in QM} V_{Vdw}(r_{ij}) \qquad [2.44]$$

where ρ I is the total charge of the QM-system and $V_{VdW}(r_{ij})$ is the van der Waals interaction between the MM-atom I and the QM-atom j. The electrostatic term takes into account the interaction of the classical point-charges with the total charge-density ϱ, which is the sum of the electronic density ρ and the positive charge of the nuclei, multiplied by the screened Coulomb potential $v(|r - R_i|)$. The calculation of the electrostatic interactions has been of difficult implementation, given two main problems:

- Positive classical charges deprived of the Pauli repulsion term act as electron traps when close to the QM-region. This leads to the electron spill-out effect in which there is an anomalous rearrangement of the electron density of the QM-part that is attracted by the MM point charges of the nearest QM/MM interface region. This effect is particularly pronounced when the wave-functions are expanded in a delocalized plane wave basis set.
- The full evaluation of the electrostatic interaction is computationally expensive.

Therefore, different treatments have been developed for short- and long-range interactions.

2.6.3 Short-range electrostatic interactions

A simple solution for electron spill-out effect is given by modifying the Coulomb potential in the vicinity of MM atoms: whereas the 1/r behaviors is maintained for large r, for values of r shorter than the covalent radius, the Coulomb potential goes to a finite value. The electrostatic term becomes:

$$\mathcal{H}_{QM/MM}^{el} = \sum_{i \in MM} q_i \int \varrho(r) \frac{r_{c_i}^4 - r^4}{r_{c_i}^5 - r^5} dr \qquad [2.45]$$

where r_c are the atomic covalent radii. This choice of the smoothing function has been tested to produce accurate results for the structural properties of many test system without any ad hoc re-parametrization of the force field.[57]

2.6.4 Long-range electrostatic interactions

The explicit calculation of the long–range electrostatic interaction energy term is too expensive in a delocalized plane wave based approach. This problem is minimized, applying a hierarchical scheme in which the classical system is divided into three shells around the QM–part and the degree of accuracy of the calculation is lowered, as the MM atoms get further away from the QM region. This approach allows a fully Hamiltonian description of the electrostatic interactions. The MM region is partitioned into three regions with two cut–off radii r_1 and r_2. The cut–off radii r_1 and r_2 are the only free parameters in this approach and have to be carefully chosen following the method of Laio et. al.[57] In the first shell closest to the QM-part ($r < r_1$), the electrostatic interaction energy of the QM charge distribution with all MM atoms within r_1 (nearest neighboring NN atoms) is explicitly calculated.

- I the second shell (for MM atoms with $r_1 < r < r_2$), the electrostatic interactions are calculated between the classical point charges and QM D-RESP[60] point charges, obtained by a fit to reproduce the electrostatic potential on the *NN* atoms. Thus the electrostatic Hamiltonian becomes:

$$\mathcal{H}^{el}_{QM/MM} = \sum_{i,j}^{r_1 \le r \le r_2} \frac{q_i Q_j^{D-RESP}}{|R_i - R_j|} \qquad [2.46]$$

- In the outer shell (for MM atoms with $r > r_2$), the electrostatic interactions are calculated via a multipolar expansion of the QM charge density and the classical point charges contained in this last shell. The Hamiltonian reads

$$\mathcal{H}^{el}_{\frac{QM}{MM}M} = C \sum q_i \frac{1}{|r - R_i|} \qquad [2.47]$$
$$+ \sum_{\alpha} D^{\alpha} \sum q_i \frac{(R_i^{\alpha} - \overline{r}^{\alpha})}{|r - R_i|^3}$$
$$+ \frac{1}{2} \sum_{\alpha\beta} Q^{\alpha\beta} \sum \frac{(R_i^{\alpha} - \overline{r}^{\alpha})(R_i^{\beta} - \overline{r}^{\beta})}{|r - R_i|^5} + \cdots$$

whit $\bar{r}$ is the origin of the multipolar expansion (i.e. the geometrical center of the QM-system). C, D and Q are the total charge, the dipole moment and the quadrupole moment QM charge distribution, respectively.

2.6.5 D-RESP charges

As described above, atomic charges on the QM-atoms are used in the second shell in the evaluation of the electrostatic interaction. These charges, namely *dynamically restrained electrostatic potential derived charges* (D-RESP charges) are obtained by a fit to the electrostatic field to the corresponding Hirshfield charges with a quadratic penalty function.[60] D-RESP charges reproduce the electrostatic potential due to the QM-charge density, which is polarized by the MM-part including the polarization effects. Furthermore, since the explicit contribution has to be computed in any case at every time-step, D-RESP charges are evaluated "on-the-fly" at each MD-step with no additional computational cost.

The instantaneous electrostatic field on the i^{th} MM atom generated by the electron density is:

$$V_i(R_i) = \int \rho\,(r) v(|r - R_i|) dr \qquad [2.48]$$

where $v(|r - R_i|)$ is the screened Coulomb potential used in Eq. [2.44], and the collection of V_i's is used as target for least-square fit. D-RESP charges can be defined as the set of point charges $\{q_i^D, i \in QM\}$ located on the QM nuclei which reproduce the electrostatic field on the MM atoms. Therefore, D-RESP charges can be derived minimizing the norm:

$$E = \sum_{j \in MM} \left(\sum_{i \in QM} \frac{q_i^D}{|R_i - R_j|} - V_j \right)^2 + W(\{q^D\}) \qquad [2.49]$$

where W is a quadratic restraining function:

$$W(\{q^D\}) = w_q \sum_{i \in QM} (q_i^D - q_i^H)^2 \qquad [2.50]$$

where w_q is a free parameter fixed to 0.1, and $\{q^H\}$ are the Hirshfeld charges of the QM-atoms that are defined as:

$$q_i^H = \int \rho\,(r) \frac{\varrho_i^{at}(|r - R_i|)}{\sum_j \varrho_j^{at}\left(|r - R_j|\right)} dr - Z_i \qquad [2.51]$$

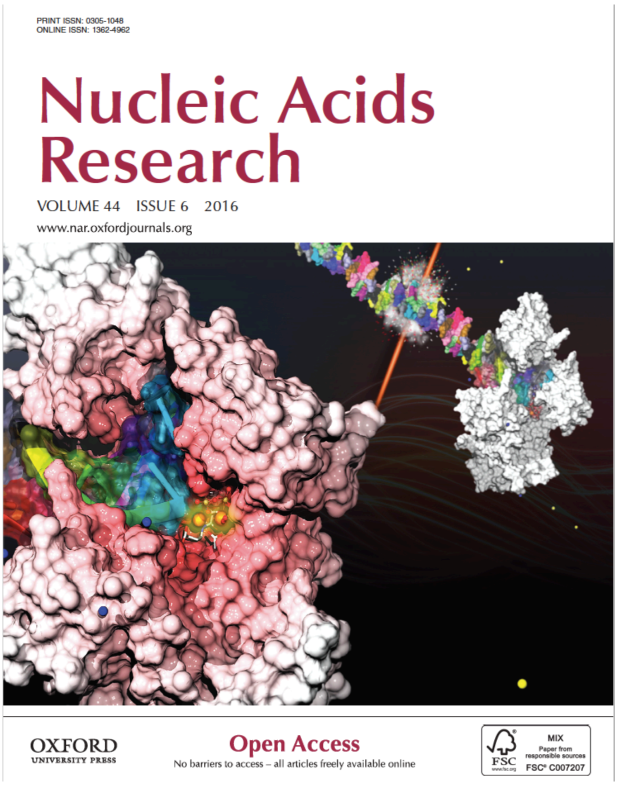
Where $\varrho^{at}$is the pseudo-valence charge density of the i^{th} atom, and $Z_i = \int \varrho_i^{at}(r)dr$ is its valence.

Chapter 3

Cooperative motion of a key positively charged residue and metal ions for DNA replication catalyzed by human DNA Polymerase-η

Vito Genna, Roberto Gaspari, Matteo Dal Peraro and Marco De Vivo

Nucleic Acids Res., **2016**, 44, 6, 2827-2836 (Cover Article)



Legend:

DNA Polymerase-h (Pol-h) bypasses UV-induced DNA damages via a two-metal-ion mechanism that assures DNA strand elongation and correct genetic inheritance to the new cells generation. Two identical ternary complexes formed by Pol-h, double strand DNA and the incoming nucleotide, are shown. The coordination to the key two metal ions, in the active site, is displayed in the protein on the front. The complex in the back suggests a DNA-damage caused by UV light. Thanks to Valentino Genna for post-processing graphical editing.

Abstract

Trans-lesion synthesis polymerases, like DNA Polymerase-η (Pol-η), are essential for cell survival. Pol-η bypasses ultraviolet-induced DNA damages via a two-metal-ion mechanism that assures DNA strand elongation, with formation of the leaving group pyrophosphate (PPi). Recent structural and kinetics studies have shown that Pol-η function depends on the highly flexible and conserved Arg61 and, intriguingly, on a transient third ion resolved at the catalytic site, as lately observed in other nucleic acid-processing metalloenzymes. How these conserved structural features facilitate DNA replication, however, is still poorly understood. Through extended molecular dynamics and free energy simulations, we unravel a highly cooperative and dynamic mechanism for DNA elongation and repair, which is here described by an equilibrium ensemble of structures that connect the reactants to the products in Pol-η catalysis. We reveal that specific conformations of Arg61 help facilitate the recruitment of the incoming base and favor the proper formation of a pre-reactive complex in Pol-η for efficient DNA editing. Also, we show that a third transient metal ion, which acts concertedly with Arg61, serves as an exit shuttle for the leaving PPi. Finally, we discuss how this effective and cooperative mechanism for DNA repair may be shared by other DNA-repairing polymerases.

3.1 Introduction

The DNA replication machinery guarantees the correct genetic inheritance to the new cells generation. However, DNA alterations, caused by endogenous and/or exogenous agents, represent a major obstacle for DNA replication. For example, daily sunlight exposure can cause the formation of ultraviolet (UV)-induced covalent bonds between adjacent pyrimidine bases within the double helix, resulting in cyclobutane pyrimidine dimers (CPDs). These are one of the major forms of DNA damage,[1,15,61-63] which are not handled by most replicative polymerases, causing the stall of the replication fork.[13,64] One way to restore the DNA replication machinery is to bypass those hurt regions via the so-called translesion synthesis (TLS) process, during which specialized translesion polymerases are capable to extend the damaged DNA. Once base pairing has been restored beyond the lesion, the replicative polymerase regains the control of DNA replication.[28,65]

One of those specialized TLS polymerases is human DNA Polymerase-η (Pol-η), member of the Y-family polymerases. Pol-η extends the damaged DNA primer strand in the presence of UV-induced CPDs. It inserts bases opposite these DNA defects, ensuring correct DNA replication.[8,15,61-63] However, Pol-η bypasses also DNA crosslinking damages generated by anticancer drugs, such as cisplatin, allowing cancer cells to survive and proliferate.[8,66-69] Pol-η's function is also critical for DNA elongation in the somatic hypermutation mechanism, which is a programmed base substitution in the variable regions of immunoglobulin genes.[70,71] For these reasons, Pol-η is a promising target to treat skin carcinogenesis[72,73] and overcome cancer drug resistance.[31,74]

Like other Y-family members,[13] Pol-η contains a two-metal-ion catalytic site, where MgA and MgB are coordinated by the so-called *DED-motif* (see Fig. 3.1). Recent time-resolved X-ray structures of Pol-η in complex with a double-strand DNA (dsDNA) and a 2′-deoxyadenosine triphosphate (dATP) have revealed the enzymatic structural evolution during catalysis.[8] The enzymatic reaction of Pol-η catalyzes the two-metal-aided formation of a new bond between the 3′-OH of the primer DNA strand and the α-phosphate of dATP, with the concomitant cleavage of the phosphodiester bond between the α- and β-phosphates of dATP.[20,75,76] Consequently, the primer DNA strand is extended by one nucleotide, with departure of the leaving group pyrophosphate (PPi, Fig. 3.1).[77-79]

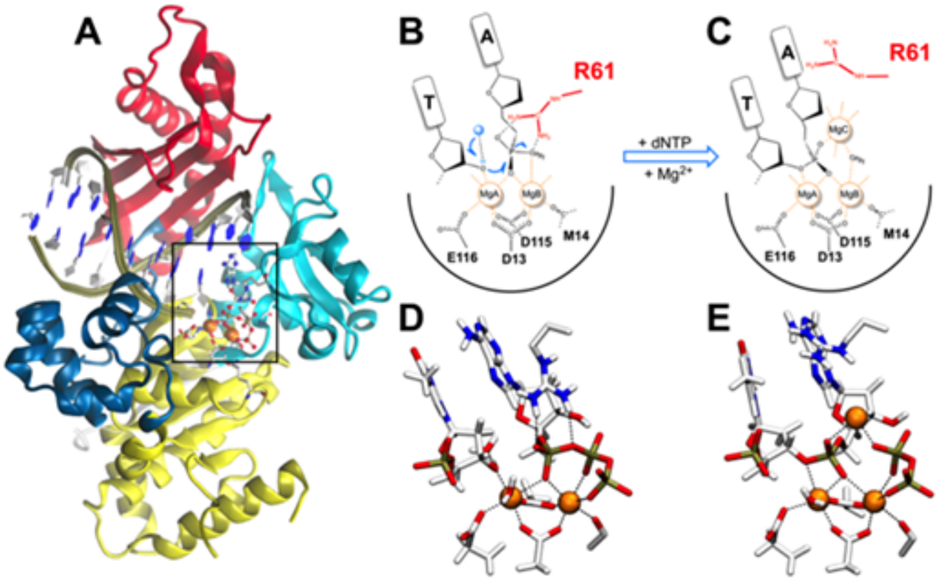


Fig. 3.1 Human polymerase-η structural and catalytic features. (**A**) Graphic of the ternary polymerase-η/dsDNA/dATP complex (1,2). Protein in graphic representation, dsDNA in ribbon representation; orange spheres indicate the two Mg2+ ions. The following domains are depicted: palm (yellow), thumb (blue), fingers (cyan) and little finger (red). On the right, scheme and close view of the catalytic site of Pol-η during DNA synthesis: (**B** and **D**) Reactant state with DNA and dNTP; (**C** and **E**) Product state with DNA and PPi. The different rotamer conformations of Arg61, from reactants to products, are highlighted.

Importantly, these recent structural data have revealed the significant flexibility of the conserved and catalytic Arg61[10,61,80] and, intriguingly, the binding of a third transient magnesium ion (MgC) at the catalytic site of Pol-η. In fact, Arg61 shows conformational transitions during catalysis, passing from the reactants to the products. While in the reactant state, Arg61 forms cation–π interactions with the incoming base and hydrogen bonds with the phosphate groups of the incoming dATP. Then, in the products, Arg61 loses its stacking interaction with the incoming base, forming with it a single H-bond. A third Arg61 conformation was also detected in the products, where Arg61 forms two H-bonds with the dATP and the apical oxygen of the template base, respectively (see Fig. 3.2-A).[8] A similar conformation of Arg61 was resolved in more recent crystal structures of Pol-η in complex with a dsDNA presenting a mispair condition, formed by a non-reactive dGTP paired with a templating thymine, in the product state.[71] In addition to the highly flexible Arg61, the third transient MgC has been observed only in the product state during Pol-η catalysis.[61] Taken together, these novel experimental results on Pol-η provide a comprehensive set of data for a thorough investigation of the mechanistic role of Arg61 and MgC to catalyze Pol-η function. In addition, other studies have recently revealed that DNA polymerases often contain multiple metal ions at the catalytic site, together with a highly flexible positively charged residue, like an arginine or a lysine.[1,8,23] How do these conserved structural features contribute to DNA replication?

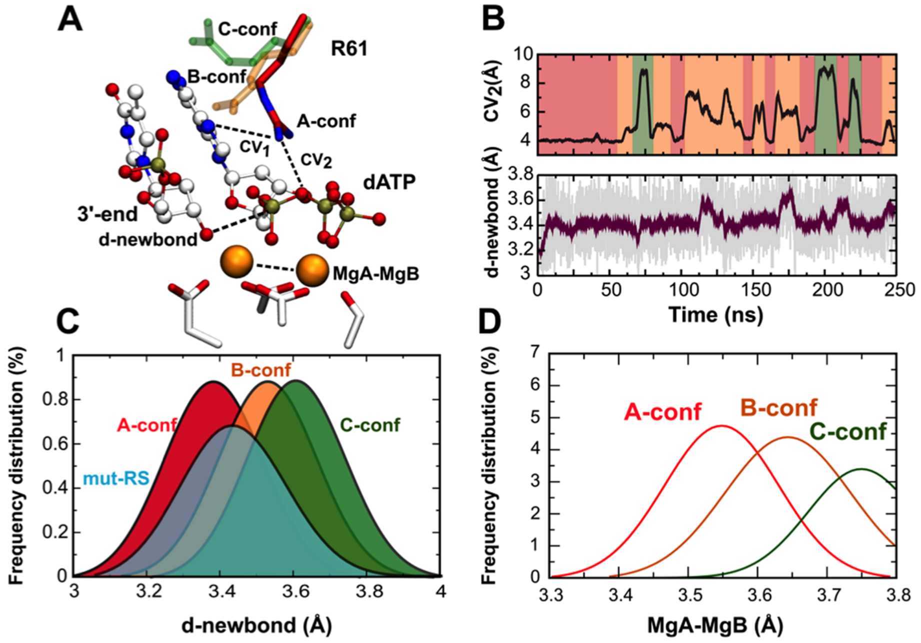


Fig. 3.2 Human polymerase-η structural and catalytic features. (**A**) Graphic of the ternary polymerase-η/dsDNA/dATP complex (1,2). Protein in graphic representation, dsDNA in ribbon representation; orange spheres indicate the two Mg^{2+} ions. The following domains are depicted: palm (yellow), thumb (blue), fingers (cyan) and little finger (red). On the right, scheme and close view of the catalytic site of Pol-η during DNA synthesis: (**B** and **D**) Reactant state with DNA and dNTP; (**C** and **E**) Product state with DNA and PPi. The different rotamer conformations of Arg61, from reactants to products, are highlighted.

Through a series of extensive classical molecular dynamics (MD) and free energy simulations, we unravel a highly cooperative mechanism, which is here described by an equilibrium ensemble of structures that connect the reactants to the products in Pol-η catalysis. This dynamic mechanism clarifies how Arg61 and multiple metal ions jointly catalyze DNA editing, and allows decoding the conformational transitions disclosed by the recent experimental data on Pol-η. We found that Arg61 initially adopts conformations that favor the recruitment of the incoming base and the proper formation of a pre-reactive complex. Then, in the products, Arg61 changes its conformation to act concertedly with a third transient metal ion to facilitate the exit of the leaving PPi after DNA elongation. Thanks to the conservation of a common catalytic structural motif, we discuss how this efficient and cooperative mechanism for DNA replication could likely be extended to other DNA polymerases.

3.2 Computational materials and methods

3.2.1 Structural models

We considered three different systems for the reactant and two for the product state[8] (1), namely: (i) *wild-type reactant state* (**wt-RS**), which is based on the recent time-resolved X-ray structures of the ternary complex (PDB ID: 4ECS) formed by human Pol-η in complex with double-strand DNA (dsDNA) and 2′-deoxyadenosine triphosphate (dATP). In this structure, the Arg61 side chain points toward the dATP α-phosphate, adopting the **A-conf** (85% occupancy in the X-ray); (ii) *Arg61Ala reactant state* (**mut-RS**), which is based on the same X-ray structure used for the **wt-RS** system, with the catalytic Arg61 mutated into an alanine residue, as studied experimentally;[61] (iii) *wild-type T:G mispaired reactant state* (**mp-RS**), which is modeled on the recent structure of Pol-η complexed with dsDNA and an incoming non-reactive dATP, which was substituted with a standard dGTP, to form a dGTP:T mispair (PDB ID: 3MR2).[61,71] In this system, Arg61 adopts a conformation defined as **C-conf** (50% occupancy in the X-ray), with a third H-bond that differentiates it from the **wt-RS** conformation (see Supplementary Fig. S3.1, **mp-RS**, Appendix Chapter 3); (iv) *wild-type product state* (**3M-PS**), based on the X-ray structure of the enzyme in complex with dsDNA and three magnesium ions (MgA, MgB and MgC). This model depicts the enzyme structure after completion of the nucleotidyl-transfer reaction and consequent products formation (PDB ID: 4ECW). The Arg61 side chain points toward the templating thymine, adopting the **C-conf** (60% occupancy in the X-ray)[8] (see Fig. 3.2). MgC coordinates three non-bridging oxygen atoms of the γ-phosphate, while three water molecules complete its octahedral coordination; (v) *$Mg_{A,B,}$ product state* (**2M-PS**), which is based on the same X-ray structure used to construct the model **3M-PS**, with MgC removed.

3.2.2 Molecular dynamics (MD) simulations

The all-atom AMBER/parm99SB-ILDN[81] force field was adopted for the Pol-η in complex with dsDNA, whereas dATP, dGTP and PPi were treated with the general amber force field.[82] The atomic charges were derived by fitting the electrostatic potential according to the Merz-Singh-Kollman scheme, also known as the RESP fitting procedure (see Supplementary Tab. S3.1, Appendix Chapter 3). The length of all covalent bonds, including hydrogen atoms, was set using the LINCS algorithm, allowing a time-integration step of 2 fs. All simulations were performed using Namd 2.9 code.[83] Long-range electrostatic interactions were calculated with the particle mesh Ewald method with a real space cutoff of 12 Å. Periodic boundary conditions in the three directions of Cartesian space were applied. Constant temperature (310 K) was imposed using Langevin dynamics[84] with a damping coefficient of 1 ps. A constant pressure of 1 atm was maintained with Langevin-Piston dynamics with a 200 fs decay period and a 50 fs time constant. The metal active site was treated with a flexible non-bonded approach based on the "atoms in molecules" partitioning scheme of the DFT-BLYP electronic density of the active site (see Supplementary Tab. S3.1).[85] We could thus consider the charge-transfer interactions between Mg^{2+} ions and their ligands, permitting possible structural rearrangements at the active site during the MD simulations. All the simulated systems were hydrated using TIP3P water molecules.[86] A total of 7 Mg^{2+} ions were added to each system to reach a final concentration of ~1 mM, while Na^{+} and Cl^{-} ions were added to neutralize the total charge. The size of the final systems was approximately 113 × 97 × 88 Å, with ~28 000 water molecules, resulting in a total number of ~70 000 atoms each.

We adopted the following simulation protocol: the systems were minimized using a steepest-descent algorithm and then slowly heated up to 310 K in 10 ns for a total of 2000 steps. The first 50 ns of production run are considered as the equilibration phase. Approximately ~1 μs of MD simulations were collected in the NPT ensemble for each of the five systems, resulting in a total of ~5 μs of dynamics. Coordinates of the systems were collected every 5 ps, for a total of ~200 000 frames for each run. Statistics were collected considering the equilibrated trajectories only, thus discarding the first ~100 ns of simulation for all the systems.

3.2.3 Free energy estimation

We used well-tempered metadynamics[87] to characterize Arg61 flexibility and estimate the free-energy landscape associated with its rearrangement. The free energy was determined as a function of two selected collective variable (CVs), which identify A-, B- and C-conf as observed during unbiased MD. The CVs are (see Fig. 3.2): (i) CV_1, which is the distance between the centers of mass of Arg61 heavy atoms (C_ζ, $N_{\eta1}$ and $N_{\eta2}$) and the heavy atoms of the incoming adenine, or the guanine in the dGTP:T pair detected in the mp-RS system; (ii) CV_2, which is the distance between the two centers of mass of Arg61 heavy atoms (C_ζ, $N_{\eta1}$ and $N_{\eta2}$) and α- and β-phosphate groups of the incoming dATP (P_α, P_β and the O atom between them). In particular, based on the crystallographic evidence and our classical MD simulations, when ~2.00 Å < CV_1 < ~4.40 Å and ~1.50 Å < CV_2 < ~4.50 Å, Arg61 is considered in A-conf, while different values of CVs refer to B- (~2.00 Å < CV_1 < ~5.50 Å and ~5.00 Å < CV_2 < ~7.00 Å) or C-conf (~3.00 Å < CV_1 < ~6.00 Å and ~7.00 Å < CV_2 < ~12.00 Å).

The fictitious temperature associated with the CVs was set to 930 K while the Gaussian function deposition rate was set to 1 ps. The initial hills height and width were set to 0.05 kcal mol^{-1} and 0.01 kcal mol^{-1}, respectively. The well-tempered simulations were carried out until their convergence (~100 ns of each of them), i.e. the progressive stabilization of the energetic minima on the free-energy surface (see Supplementary Fig. S3.2, Appendix Chapter 3). All other parameters correspond to those used for the plain MD simulations, described above.

Well-tempered metadynamics was also used to explore the PPi unbinding process in the presence of MgC using system (4).[62] In this case, we considered a single CV (CV_{PPi}) that describes the difference between the bound and unbound state of PPi during its departure. This CV is the distance between the centers of mass of the PPi atoms and the αC of the DED-motif (Supplementary Figure S3.2, Appendix Chapter 3). In this case, we applied a fictitious CV temperature of 4650 K. The Gaussian function deposition rate was set to1 ps while the initial hills height and width were respectively set to 0.05 kcal mol^{-1} and 0.01 kcal mol^{-1}.

3.3 Results

3.3.1 Pre-reactive configurations in Pol-η

Through two independent MD simulations (~500 ns) of the wild-type reactant **wt-RS** complex, formed by Pol-η and its substrates, dsDNA and dATP, we examined the evolution over time of key structural features of the catalytic pocket. Arg61 stochastically adopted different rotamer conformations along the trajectories (see Fig. 3.2-B). During the first ~60 ns of production run, Arg61 conserved its crystallographic **A-conf** pose (see Fig. 3.2-B). In this conformation, its guanidine group forms a bi-dentate H-bond interaction with the α/β-phosphate groups of dATP. Subsequently, Arg61 changed its side-chain orientation, adopting the **B-conf** pose, in which it forms its characteristic H-bond with the N7 atom of the incoming dATP. This conformational change was well described by the distance CV_2 (see Fig. 3.2-A), which increased from ~4.00 Å to ~5.00 (see Fig. 3.2-B). Furthermore, Arg61 occasionally assumed a **C-conf** pose, which is characterized by an additional H-bond formed by its guanidine group with the apical oxygen of the 5′-terminal thymine of the template strand (2). Each **B-** and **C-conf** lasted for a short time, ranging from ~10 to ~22 ns, before Arg61 returned to the native **A-conf** (see Fig. 3.2-B).

We then monitored the alignment of the substrates dsDNA and the incoming dATP to form the in-line conformation, which preludes the catalytic S_N2-like phosphoryl-transfer reaction in Pol-η. Notably, the 3′-end sugar pucker remains in a stable C3′-endo conformation for the entire simulations, in agreement with the X-ray data.[8] We also examined the *d-newbond* variation, i.e. the distance between the two reactive groups, the 3′-OH of the terminal base of the DNA, and the α-phosphorus of the incoming nucleotide (see Fig. 3.2-A). We found that *d-newbond* had a mean value of 3.35 ± 0.07 Å when Arg61 adopted **A-conf**, which agreed well with the X-ray value of *d-newbond* (3.30 Å). However, when Arg61 was in **B-** or **C-conf** during the simulations, *d-newbond* slightly increased to a longer stable value of 3.52 ± 0.07 Å and 3.65 ± 0.05 Å, respectively. This suggests a conformational state that is likely less prone to promote catalysis. Three Gaussians fit the frequency distribution of *d-newbond* when Arg61 assumes **A-conf**, **B-conf** and **C-conf**, pointing out the elongation of *d-newbond* when Arg61 moved away from **A-conf** (see Fig. 3.2-C). In particular, the greater the distance between the Arg61 guanidine group and the phosphate tail of the incoming base, the longer the length of *d-newbond* (**A-conf** > **B-conf** > **C-conf**), as confirmed by confidence interval analysis (see Supporting Text). Arg61 fluctuations seemed to affect *d-newbond* only, while the surrounding structural environment maintained its native geometry throughout the whole trajectory.

Concomitantly, the distance between MgA and MgB became shorter (3.51 ± 0.07 Å) when Arg61 assumed **A-conf**, in fine agreement with the X-ray structure of the pre-reactive state. This inter-metal distance became larger (3.65 ± 0.06 Å) when Arg61 switched from **A-** to **B-conf** and further increased (3.75 ± 0.06 Å) when Arg61 adopted **C-conf**. Three Gaussian curves indicate the frequency distribution of the inter-metal length (MgA-MgB) with respect to the three different Arg61 conformations (see Fig. 3.2-D), highlighting the plasticity of the catalytic pockets in aiding catalysis, as for those enzymes that use a bi(tri)-metal core to catalyze phosphoryl-transfer reactions.[1,76,85,88] See further analysis in Supplementary Fig. S3.3–6 and Supplementary Tab. S3.2, Appendix Chapter 3)

A mutant model system **mut-RS** (Arg61Ala) of the ternary Pol-η/dsDNA/dATP complex was also simulated for ~1 μs. The enzyme well maintained its overall structural framework and the catalytic metal ions conserved their typical octahedral coordination geometry. However, the only substrate showed a slight rearrangement already in the early stages of the simulation. After only ~50 ns of production run, we observed a conformational change of the adenine base of the incoming dATP, which was buckled with respect to the 2′-deoexyribose. Because of this buckling, the incoming dATP lost the H-bonds established with the complementary 5′-thymine (inter-base H-bonds). Indeed, the double H-bond interaction expected for a dATP:T pairing (according to Watson–Crick's base-pairing model) is present in only ~11.00% of the collected frames of the **mut-RS** simulations and ~20.00% of the **wt-RS** system frame. Hence, the lack of Arg61 compromises the base interaction pattern, which becomes unstable. As a consequence, the incoming dATP lost its native position. This motion is captured well by the dATP RMSD value, which increases from 0.33 ± 0.09 Å in the **wt-RS** to 0.89 ± 0.26 Å in **mut-RS** (see Supplementary Fig. S3.3, Appendix Chapter 3). This motion is also reflected by an increased fluctuation of *d-newbond*, which was 3.44 ± 0.14 Å.

The mispaired model **mp-RS** shows a non-Watson–Crick base-pairing dGTP:T in the dsDNA, with the formation of a ternary Pol-η/dsDNA/dGTP complex.[71] Notably, **C-conf** is here characterized by an H-bond pattern that exists in the **mp-RS** crystal only, due to the mispairing dGTP:T condition (see Supplementary Fig. S3.1, Appendix Chapter 3). We ran two simulations (~250 ns, each) using both **A-** and **C-conf** Arg61 conformations as a starting pose. Interestingly, in the **A-conf** system, Arg61 suddenly rotated around its φ dihedral angle (Cβ-Cγ-Cδ-Nϵ) readopting its native **C-conf**, which then remained stable. The fast **A**-to-**C-conf** switching is captured well by CV_2, which was ~3.50 Å (**A-conf**) in the first 5 ns of simulation. Then, CV_2 rapidly increased to a value of ~10.00 Å (**C-conf**) during the remaining part of the simulation. We also simulated a system in which Arg61 starts in **C-conf**. As expected, **C-conf** was stably maintained throughout the entire simulated timescale and no Arg61 conformational changes were observed (see Supplementary Fig. S3.6, Appendix Chapter 3). In fact, the three H-bonds characteristic of **C-conf** displayed high stability with an average length of 2.63 ± 0.19 Å. Here, *d-newbond* shows a value of 3.23 ± 0.07 Å, indicating a perfect alignment of the reactive groups. The limited simulated timescale of this system did not capture the large conformational changes in the catalytic pocket, which are displayed by X-ray structures of this system and which likely explain its lower catalytic efficiency (see "Discussion" section).

3.3.2 Energetics of Arg61 motions in pre-reactive states

We calculated the free energy surfaces (FESs) of the Arg61 A↔B/A↔C interconversions for the **wt-RS** and **mp-RS** systems (see Fig. 3.3), as a function of CV_1 and CV_2 (see "Computational materials and methods" section). The FES related to the **wt-RS** system was characterized by three low free-energy minima. The deepest minimum corresponded to the X-ray conformation, in which Arg61 was detected in **A-conf** (CV_1 = ~4.00 Å and CV_2 = ~2.00 Å). This reflects our unbiased MD simulations, in which Arg61 was most likely to be in this conformation. Nearby this absolute minimum, a second free-energy basin revealed a variant form of **A-conf** (CV_1 = ~4.00 Å, CV_2 = ~4.00 Å), which was frequently observed in **wt-RS** unbiased simulations. In this variant, only one of the bi-dentate H-bond interactions with the phosphate groups (characteristic of **A-conf**) was formed. Then, a less stable minimum was found for

Arg61 in **B-conf** (CV$_1$ = ~2.20 Å, CV$_2$ = ~7.00 Å), while a third relative minimum (CV$_1$ = ~3.80 Å and CV$_2$ = ~9.50 Å) identified Arg61 in **C-conf**. The overall relative stability of the three rotamers for Arg61 (**A-conf** > **B-conf** > **C-conf**) corroborated our findings from unbiased MD (see Supplementary Tab. S3.3, Appendix Chapter 3).

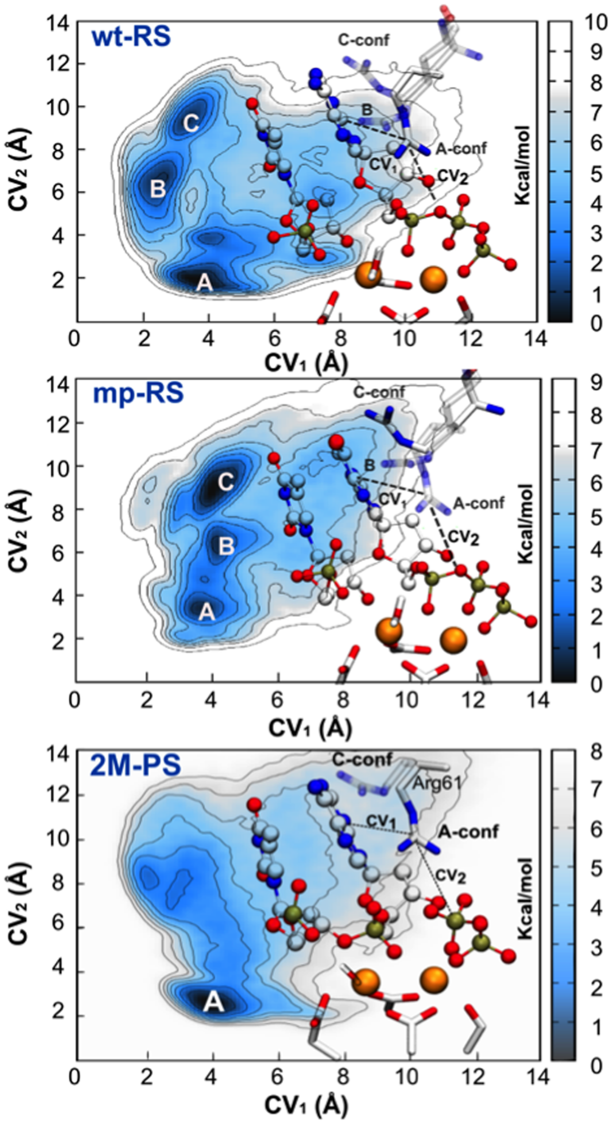


Fig. 3.3 Free energy simulations of Arg61 motions. (wt-RS) Free energy surface (FES) of wt-RS system showing three main energetic basins for different Arg61 conformations (A-, B- and C-conf). (mp-RS) FES of mp-RS system characterized by three minima. The deepest one corresponds to C-conf. Nomenclature is maintained as for wt-RS. CVs are also depicted (see 'Computational materials and methods' section). (2M-PS) Free energy landscape of 2M-PS system. CVs are also depicted (see 'Computational materials and methods' section). Isolines are every 1.00 kcal mol-1 for each FES.

The free-energy landscape for the **mp-RS** system was quite different, with the basin of Arg61 in **A-conf** (CV_1 = ~3.80 Å, CV_2 = ~3.00 Å), which was then no longer the absolute minimum of the FES (Figure 3). As expected, **C-conf** was now the deepest free-energy minimum (CV_1 = ~4.00 Å and CV_2 = ~9.50 Å). This explains the high population of this configuration during the unbiased MD simulations of **mp-RS**. Arg61 in **B-conf** (CV_1 = ~4.00 Å, CV_2 = ~7.00 Å) was found in a basin of ~3.50–4.00 kcal mol^{-1} higher than in **C-conf**. A variant form of the **A-conf** (CV_1 = ~4.00 Å, CV_2 = ~4.30 Å) was also found ~6.00 kcal mol^{-1} higher in free energy than **C-conf**. In this latter case, Arg61 H-bonds only one non-bridging oxygen atom of the Pα of the dGTP, rather than forming the favored bifurcated interaction as observed for the **wt-RS** system. As a result, these less stable Arg61 states are rarely detected in unbiased MD (see Supplementary Tab. S3.3, Appendix Chapter 3).

3.3.3 Post-reactive configurations in Pol-η

We collected ~1 μs of simulation time for two configurations of the product system, where the enzyme is characterized by Pol-η in complex with dsDNA, now extended by one nucleotide with respect to the reactants, and the free PPi leaving group.[8] Namely, we simulated **3M-PS** and **2M-PS**, which respectively have three and two Mg^{2+} ions in the catalytic pockets (see Fig. 3.1). Notably, in **3M-PS**, Arg61 steadily maintained **C-conf**, with no Arg61 A↔C/A↔B interconversions. Instead, in **2M-PS**, after removing MgC, Arg61 immediately changed its rotamer conformation from **C-** to **A-conf** (*t* = ~20 ns), while the overall protein, MgA and MgB ions maintained their general structure and coordination with respect to the X-ray structure. Hence, in **2M-PS**, Arg61 adopted different conformations along the MD runs, as in the pre-reactive state. In fact, Arg61 assumed all the known conformations with preference for the **C-conf**, which appeared at t = ~0 ns, ~280 ns, ~422 ns, ~475 ns, ~710 ns, ~940 ns (see Supplementary Fig. S3.7, Appendix Chapter 3).

Crucially, we observed a new transient Mg^{2+} ion that spontaneously reached, from the water bulk, the negative potential well near the catalytic pocket, filling up the vacancy originally created by the removal of MgC, between the PPi and the α-phosphate of the new 3′-end (see Fig. 3.4, and Supplementary Fig. S3.8, Appendix Chapter 3). This unprompted event rebuilt a structural framework that resembled the original **3M-PS** system (RMSD = 1.87 Å versus **3M-PS**). Then, further transient rearrangements of Arg61, DED-motif, MgB and PPi occurred during these MD runs, after a new third Mg^{2+} ion reached the catalytic pocket. In fact, for ~430 ns (following third Mg^{2+} ion binding) the geometry coordination of MgB was slightly altered (see Fig. 3.4). As a consequence, Asp13 and Glu118 increased their distance from MgB, passing from 2.00 ± 0.30 Å to 3.51 ± 0.20 Å, respectively. Concomitantly, the PPi leaving group interacted more closely with MgC, with an average distance of 2.10 ± 0.21 Å. This overall structural rearrangement allowed Arg61 to switch from **C-conf** to **A-conf**, forming bifurcated H-bond interactions with the newly formed adduct (MgB–MgC–PPi, see Fig. 3.4). This orchestrated motion, which involved MgB, MgC, PPi, Arg61, Arg55 and Tyr52, suggests a step-wise leaving mechanism of the PPi from the catalytic pocket, which is an essential part of the catalytic turnover for Pol-η's function. As a consequence, the distance between the centers of mass of the PPi and the DED-motif increased from ~3.50 Å to ~7.20 Å. However, after these ~430 ns of unbiased MD, this metastable intermediate was likely unable to overcome the energetic barrier to complete the leaving group release mechanism (see below). Consequently, the third metal left the cavity (*t* = ~800 ns) and the system fell back into the closest energetic minimum, restoring the original **2M-PS** structure, which was maintained well for the remainder of the simulations (see Fig. 3.4).

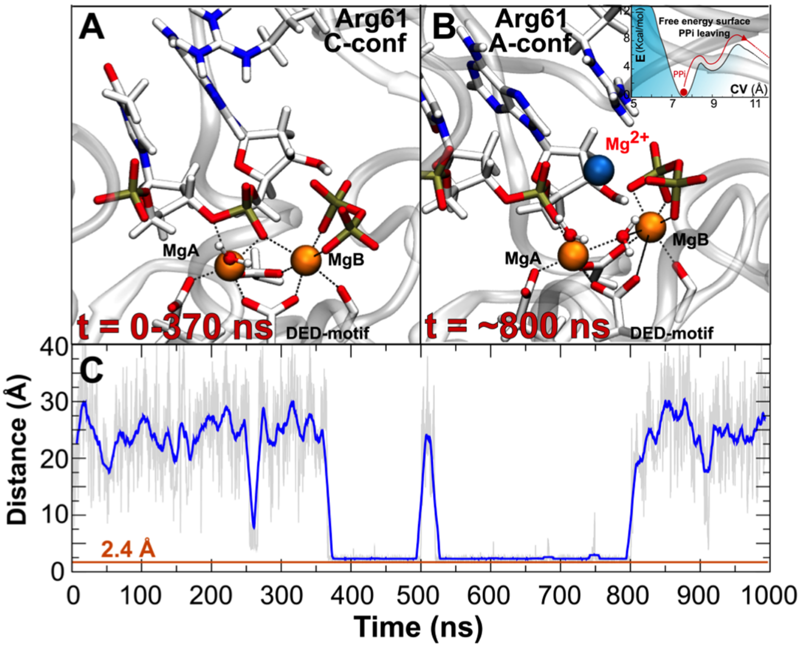


Fig. 3.4 Third Mg ion binding and release mechanisms. (**A**) Post-reactive system in which Arg61 adopts C-conf. (**B**) At *t* = ~530 ns, a third Mg^{2+} metal stably binds the catalytic pocket, forming the 3M-PS system, that matches the X-Ray structure 4ECW. This snapshot presages PPi departure, a mechanism investigated by using enhanced sampling methods. In the upper-right corner is reported an indicative free energy profile of the step-wise PPi leaving mechanism. Importantly, to favor PPi release, Arg61 switches from C- to A-conf. (**C**) MgC-PPi distance monitored along the simulated timescale. Blue line indicates the distance of the third Mg^{2+} from the pyrophosphate (PPi) center of mass. Orange line indicates the distance between third ion and center of mass of PPi detected in the crystallographic structure 4ECW.

3.3.4 Energetics of Arg61 motion in post-reactive configurations and PPi-releasing mechanism

In **2M-PS**, the FES showed one deep minimum in which Arg61 adopted the **A-conf** (CV_1 = ~4.00 Å and CV_2 = ~3.00 Å). However, the energy required to switch from **A-conf** to **C-conf** was only ~3.50 kcal mol^{-1}, which explains the spontaneous Arg61 A↔C interconversion detected in our unbiased MD simulations of **2M-PS** (see Fig. 3.3 and Supplementary Tab. S3.3 in Appendix Chapter 3).

We also investigated the full PPi-releasing mechanism, starting from the **3M-PS** system. We used a new CV defined as CV_{PPi} (i.e. the distance between the centers of mass of PPi and the αC of DED-motif and Met14 as a CV, see Supplementary Fig. S3.2). Our unbiased MD simulations showed a metastable conformation presaging a possible leaving mechanism of the PPi complexed with MgB and MgC. We thus considered PPi bound to the catalytic pocket when $CV_{PPi} \leq \sim 9.50$ Å, while the MgB–MgC–PPi was completely unbound when $CV_{PPi} \geq \sim 9.50$ Å.

Starting our simulations from the crystallographic PPi conformation still bound to the catalytic site (Figure 4B, $CV_{PPi} = \sim 7.50$ Å), a two-step mechanism allowed the release of the PPi, complexed with MgB and MgC, from the Pol-η catalytic site. First, the system reached a metastable intermediate, characterized by the partial unbound state of the MgB–MgC–PPi complex, as suggested in the unbiased MD. During this state, Arg61 passed from **C-conf** to **A-conf** and concertedly interacted (together with Arg55 and Tyr52) with the newly formed complex MgB–MgC–PPi, which underwent conformational changes that determined its detachment of ~3.70 Å, with respect to its native conformation. A free energy contribution of ~5 kcal mol^{-1} was required to overcome the first barrier of this two-step releasing mechanism. A second step then occurred, in which Arg61 rotated around its φ dihedral angle, readopting its native **C-conf**, to aid the final release of the product adduct MgB–MgC–PPi, which was completely disengaged from the DED-motif in Pol-η catalytic site. For this second step, a barrier of ~4 kcal mol^{-1} had to be overcome. The unbinding of the PPi from the catalytic site required the build-up of a given amount of biasing potential, which provided a representation of the energy landscape along the dissociation pathway (see Fig. 3.4).

3.4 Discussion

Recent structural data of the human DNA polymerase-η (Pol-η) have highlighted the significant structural flexibility of the conserved Arg61 and, intriguingly, the binding of a transient third ion resolved at the catalytic site. Here, extended MD and free energy simulations of several systems of Pol-η in different pre- and post-reactive conditions demonstrate how the third metal ion cooperate with different Arg61 configurations to facilitate DNA replication.

During the ∼1 μs of MD simulations of the wild-type reactant system **wt-RS**, we observed three different Arg61 conformations (**A-**, **B-** and **C-conf**), which agreed with the crystallographic evidence.[8,61,71] **A-conf** was the most sampled conformation, found in the 54.10% of the total frames, while **B-conf** and **C-conf** were found in 29.80 and 16.10% of the MD trajectory, respectively (**A-conf** occupancy of 85% in the 4ECS X-ray structure). Importantly, free energy simulations confirmed **A-conf** as the most stable Arg61 conformation in **wt-RS**, located in the deepest energetic minimum of the free-energy surface (see Fig. 3.3).

In addition, we found that *d-newbond*, which specifies the length of the forming bond during catalysis (Figure 2), was shorter (3.35 ± 0.07 Å) when Arg61 assumed **A-conf** and progressively increased when Arg61 adopted **B-** (3.52 ± 0.07 Å) or **C-conf** (3.65 ± 0.05 Å, see Confidence Interval analysis in Supplementary Text, Appendix Chapter 3). Concomitantly, *d-newbond* fluctuations correlated with the inter-metal distance (MgA-MgB). Indeed, only when Arg61 adopted **A-conf**, MgA-MgB length was stably maintained around 3.51 ± 0.07 Å, in fine agreement with the crystallographic evidence[8] and other several other studies,[88-93] highlighting the general flexibility of the catalytic metal ions, passing from a pre-reactive to a post-reactive configuration. Also, we found that an increase in *d-newbond* length is correlated with an increase in inter-metal distance, with respect to the different Arg61 conformations (see Fig. 3.2).

Interestingly, the Arg61 conformations along our MD simulations, in both pre- and post-reactive states, match well with two distinct sets of X-ray structures of Pol-η, resolved in complex with damaged DNA. Zhao *et al.*[92] have solved a ternary complex formed by Pol-η, a non-hydrolysable incoming nucleotide and a dsDNA containing a lesion caused by cisplatin.[94] In the crystallographic pre-reactive state (PDB ID: 4DL4), Arg61 adopts both **A-** and **C-conf** (see Supplementary Fig. S3.9, Appendix Chapter 3).

In this crystal, **A-conf** has an occupancy value of 60%, while **C-conf** has a lower occupancy of 40%. Here, *d-newbond* is 3.35 Å. On the other hand, in the post-insertion complex (PDB ID: 4DL6), Arg61 only adopts **C-conf**, with a longer *d-newbond*, corresponding to 3.80 Å. Our results also match well with the X-ray structures solved by Patra *et al.*[15], which report a ternary complex made by Pol-η, a non-hydrolysable incoming nucleotide and a dsDNA affected by the lesion 7,8-dihydro-8-oxo-2′-deoxyguanosine (8-oxoG).[95-97] Also here, when Arg61 adopts **A-conf**-like conformations (PDBid 4O3O), *d-newbond* has a value of 3.13 Å. On the other hand, in a second X-ray structure (PDB ID: 4O3Q), Arg61 assumes **C-conf**-like conformations and *d-newbond* increases to a value of 3.71 Å (see Supplementary Fig. S3.9, Appendix Chapter 3).

In the mutated system **mut-RS** (Arg61Ala), which was found experimentally to be from 2- to 6-fold less efficient in polymerization (in both normal and damaged DNA) than the wild-type form,[61,98] we found a less stable dATP conformation, which was buckled. This caused instability of the Watson–Crick's H-bond interaction pattern (Supplementary Fig. S3.5, Appendix Chapter 3). As a result, *d-newbond* fluctuations in **mut-RS** increase, which is reflected in a wider Gaussian distribution function (see Fig. 3.2). Overall, these data suggest that Arg61 is necessary to correctly guide the incoming base into the catalytic site (via **B-** and **C-conf**) and to promote (via **A-conf**) the phosphoryl-transfer reaction, catalyzing the efficient incorporation of the incoming nucleotide into the dsDNA during polymerization. Besides, these results help further rationalizing the outcome of recent MD simulations describing a complex network of local hydrogen-bond interactions, involving Arg61, within the catalytic site of Pol-η. These local interactions thermodynamically favor the selection and binding of specific deoxyribose nucleotides triphosphates in the presence of cyclobutane thymine−thymine dimers (TTDs, a UV-induced DNA damage). As such, they likely contribute to the fidelity and overall efficiency of Pol-η.[62,99]

In the mispaired system **mp-RS**,[71] we found that Arg61 was highly stable, never spontaneously leaving its starting **C-conf**. Even when Arg61 was manually placed into a starting **A-conf**, the key residue spontaneously reassumed **C-conf** in the very first ns of unbiased MD (see Supplementary Fig. S3.6, Appendix Chapter 3). This can be explained by the presence of an additional H-bond, which is only formed in **C-conf** with the dGTP incoming base (see Supplementary Fig. S3.1, Appendix Chapter 3). The preference of **C-conf** suggests a lower reactivity of this system. This agrees well with the kinetics experiments reporting a drop in Pol-η efficiency ($k_{\mathrm{cat}}/K_{\mathrm{M}}$), from 47.80 μM^{-1} min^{-1} for the canonical dATP:dT pairing to only 1.70 μM^{-1} min^{-1} for this specific dGTP misincorporation.[92] Along the same lines, a recent paper of Patra *et al.*[93] reports kinetics and structural data of Pol-η incorporating different nucleotides (dATP, dTTP, dCTP and dGTP), in presence of the 8-oxoG lesion. The dGTP:8-oxoG incorporation had a low efficiency value, equal to 0.016 ± 0.003 μM^{-1} s^{-1}. This can be compared with the value of 1.0 ± 0.16 μM^{-1} s^{-1} associated with the canonical dCTP:dG incorporation. Notably, Arg61 is in **A-conf** only in those X-ray structures obtained for the control systems (native base pairing), while Arg61 is in **C-conf** in those structures having non-native base pairings (see Supplementary Fig. S3.9, Appendix Chapter 3). Hence, a lower polymerization efficiency for misincorporation may be due to the lack of **A-conf** formation in pre-reactive states, as observed in our **mp-RS** dynamics. Free energy simulations of **mp-RS** also showed the triple-H-bonded **C-conf** as the deepest energetic basin, while **A-conf** and **B-conf** were found in two less stable minima on the FES (see Fig. 3.3, **mp-RS**). This likely prevents the formation of a proper pre-reactive state, as found in **wt-RS**.

During the MD simulations of the post-reactive two-metal-ion system **2M-PS**, we found that **C-conf** remained the most populated state, being present for ~62.00% of the simulation time. This suggests that the product formation, where the incoming base is covalently bound to the substrate primer strand, favors **C-conf** over **A-conf**. While the latter seems a specific conformation needed in the pre-reactive state, the preference for **C-conf** in the post-reactive state seems necessary to allow binding of the third ion, as observed experimentally.[8,13] In fact, during the unbiased MD simulations of **2M-PS**, a third Mg^{2+} moved spontaneously to the catalytic pocket at t = ~370 ns, reconstituting the three-metal-ion system **3M-PS** system (see Fig. 3.4). The unprompted entry of this transient third ion initially destabilized the pocket environment, with formation of a metastable MgB–MgC–PPi adduct which seemed to evolve toward the exit of the PPi leaving group. However, this metastable intermediate lasted for ~430 ns. After this time, the third ion spontaneously left the pocket and the leaving MgB–MgC–PPi adduct fell slowly back into its original position, restoring the canonical post-reactive coordination of the catalytic site of Pol-η.[8]

Intriguingly, this orchestrated structural motion suggests a putative leaving mechanism of the PPi group that implies a key role of MgC as an exit shuttle, together with MgB and the chelating Arg61, Arg55 and Tyr52. This possible leaving pathway was further suggested by our free energy simulations, which returned an estimated energy barrier of ~8 kcal mol^{-1} overcome via a two-step mechanism. First, in **3M-PS**, the overall system escaped from the crystallographic post-reactive state, containing three metal ions in the pocket and located at CV_{PPi} = ~7.50 Å. This first step slowly evolved into a metastable state, in which the MgB–MgC–PPi group was partially unbound, while Arg61 was in **A-conf**. The second step led to the final exit of the complexed PPi leaving group. Concomitantly, Arg61 rotated along its φ dihedral angle, accompanying the leaving group release (see Fig. 3.4).

Using the computed free energy values for the leaving PPi departure and the transition-state theory,[100,101] we could roughly estimate that the unbinding time of the PPi is around ~100 μs (for further details see Supplementary Text in Appendix Chapter 3). Experimental results indicate that the Y-family polymerases are low-processivity enzymes (with a rate constant of nucleotide incorporation that often is around ~1.00–3.00 s^{-1}).[61,93,98] Thus, the leaving PPi departure seems a rather fast step, compared to the time requested for the overall process.[15,93,102,103] For this reason, our results agree with the proposal that the leaving PPi departure is not the rate-limiting step of the overall process. The relaxation step of the whole ternary complex, after PPi release, seems more likely to be the rate-limiting step of the polymerization process catalyzed by human Pol-η, as already proposed for the structurally similar Y-family members Dpo4 and Pol-κ.

A structural superimposition between Pol-η, Pol-ι[104] and Dpo4,[75] that also are Y-family DNA polymerases,[13,16,28,105] highlights the key structural elements shared by these enzymes (see Fig. 3.5). Importantly, the conserved Lys77 in the crystallographic pre-reactive Pol-ι complex points down in an **A-conf-like** conformation, while Lys56 in the crystallographic post-reactive Dpo4 complex points up in a **C-conf-like** conformation. These different conformations of this positively charged residue agree with our MD simulation findings in Pol-η, where Arg61 adopted mostly **A-conf** in the reactants and mostly **C-conf** in the products. Notably, when Arg61 is mutated into a Lys residue in human Pol-η (see Supplementary Fig. S3.10, Appendix Chapter 3), the enzyme retains a slightly lower activity.[98] Besides, the presence of a positively charged residue in a similar position of the catalytic site is found in other Y-family polymerases, including *Escherichia coli* Pol-IV, *E. coli* Pol-V and human enzymes like Pol-κ and Rev1, which also catalyze a two-metal-aided (Mg^{2+}) phosphoryl-transfer reaction to process DNA.[32,104,106-110] This suggests its key role in recruiting and processing the (d)NTP/(d)NMP for DNA/RNA elongation/digestion in normal[111-113] and deranged conditions.[11,68,114,115] As for the third transient metal ion, this emerging and still puzzling aspect of two-metal-ion catalysis has been recently reported in several other nucleic-acid-processing enzymes[1] including, recently, DNA polymerase β,[23] suggesting a dynamic and functional rearrangement of a third additional metal ion in an extended neighborhood of the two-metal-ion catalytic pocket.

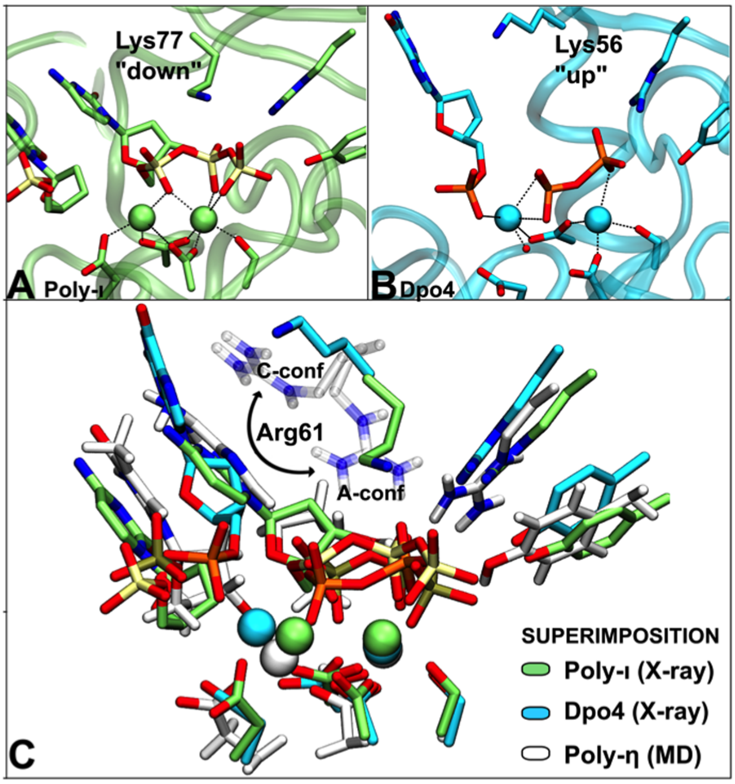


Fig. 3.5 Structural superimposition of different Y-family members. (**A**) X-ray structure (PDB ID: 2ALZ) of a pre-reactive configuration of the human Pol-ι enzyme. Lys77 points 'down' like Arg61 in A-Conf in Pol-η, establishing H-bond interactions with the phosphate groups of the incoming base. (**B**) X-ray structure (PDB ID: 2AGO) of a post-reactive configuration in the *Sulfolobus solfataricus* Dpo4 enzyme. Here, Lys56 adopts an 'up' conformation, similar to the C-conf detected in the products of Pol-η catalysis. (**C**) Superimposition of the three different catalytic sites (Pol-η, Pol-ι and Dpo4).

3.5 Conclusions

In conclusion, our results greatly enrich the interpretation of existing structural data on Pol-η and reveal an effective and cooperative mechanism for enzymatic repair of DNA operated by DNA-repairing polymerases. This is based on conformational transitions of a key positively charged residue coupled with a transient third metal ion at the catalytic site of Pol-η, which ultimately acts as an exit shuttle for the leaving PPi departure. Based on the conservation of these key structural features, this efficient mechanism detected in Pol-η may be shared by several DNA polymerases and other binuclear metalloenzymes for nucleic acid processing, with relevance for the de-novo design of enzymes for DNA processing or the structure-based discovery of small molecules for anti-cancer therapy that targets DNA polymerases in cancer cells.[17,27,30,116,117]

Supporting Information

Further details concerning MD setup, RESP charges, statistical analysis and additional results are available in the Appendix Chapter 3.

Acknowledgments

We thank Italian Association for Cancer Research (AIRC) [‘MFAG n. 14140’]. Funding for open access charge: Istituto Italiano di Tecnologia; Italian Association for Cancer Research (AIRC).

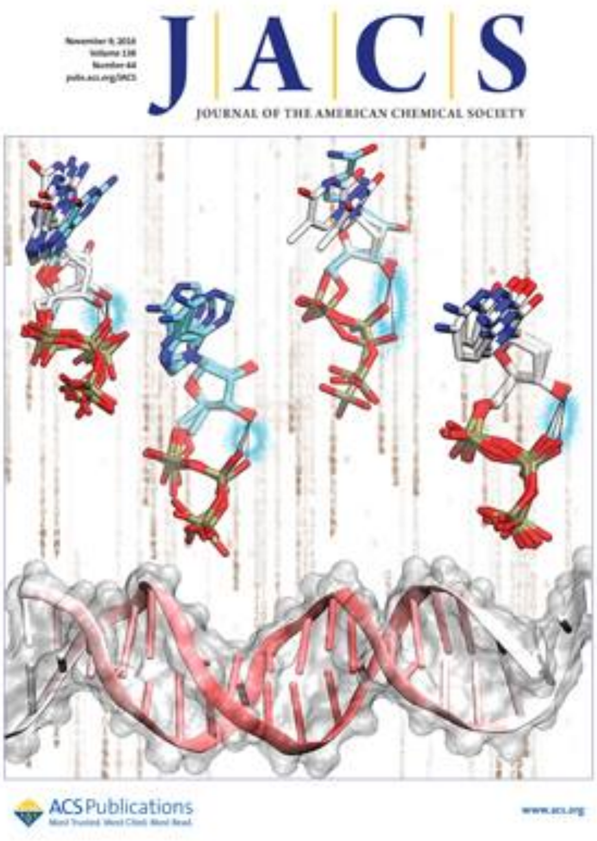
Chapter 4

A Self-Activated Mechanism for Nucleic Acid Polymerization Catalyzed by DNA/RNA Polymerases

Vito Genna, Pietro Vidossich, Emiliano Ippoliti, Paolo Carloni and Marco De Vivo

J. Am. Chem. Soc., **2016**, 138, 44, 14592-14598 (Cover Article)

Lockwood D., J. Am. Chem. Soc., **2016**, 138, 44, 14507-14508 (Spotlights)



Legend:

Ternary DNA/RNA polymerase (Pol) complexes show the formation of the previously unrecognized intramolecular H-bond formed by the nucleophilic 3′-OH and the β-phosphate of the incoming nucleotide. The illustration shows this crucial element used to propose and compute a novel self-activated mechanism (SAM) for nucleic acid polymerization in Pols. Thanks to Valentino Genna for post-processing graphical editing.

Abstract

The enzymatic polymerization of DNA and RNA is the basis for genetic inheritance for all living organisms. It is catalyzed by the DNA/RNA polymerase (Pol) superfamily. Here, bioinformatics analysis reveals that the incoming nucleotide substrate always forms an H-bond between its 3′-OH and β-phosphate moieties upon formation of the Michaelis complex. This previously unrecognized H-bond implies a novel self-activated mechanism (SAM), which synergistically connects the in situ nucleophile formation with subsequent nucleotide addition and, importantly, nucleic acid translocation. Thus, SAM allows an elegant and efficient closed-loop sequence of chemical and physical steps for Pol catalysis. This is markedly different from previous mechanistic hypotheses. Our proposed mechanism is corroborated via ab initio QM/MM simulations on a specific Pol, the human DNA polymerase-η, an enzyme involved in repairing damaged DNA. The structural conservation of DNA and RNA Pols supports the possible extension of SAM to Pol enzymes from the three domains of life.

4.1 Introduction

Nucleic acid polymerization is a key process for genetic inheritance across the three domains of life. This is performed by a set of DNA/RNA polymerases (Pols) that are often effective drug targets for treating cancer, viral and bacterial infections, and neurodegenerative diseases.[5118,119] Pols operate via the two-metal (Mg^{2+})-ion mechanism for incorporating an incoming nucleotide [(d)NTP] into the growing nucleic acid strand, via the typical S_N2-like phosphoryl-transfer reaction, with liberation of a pyrophosphate (PP_i) leaving group (see Fig. 4.1).[20]

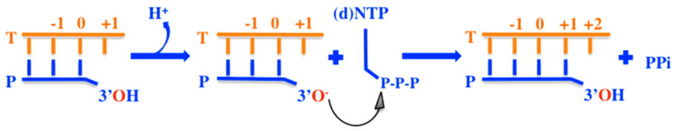


Fig. 4.1. Diagram of nucleic acid synthesis catalyzed by RNA/DNA polymerases. Nucleophile activation, nucleotide addition, and DNA translocation for nucleic acid polymerization, with liberation of a pyrophosphate (PPi) leaving group. Orange indicates the template strand (T) while blue indicates the primer strand (P).

The established two-metal-aided phosphoryl transfer reaction for nucleotide addition in Pols[1,77,120,121] is preceded by deprotonation of the 3′-hydroxyl (3′-OH) of the 3′-end deoxyribose. This generates the activated nucleophilic 3′-hydroxide ion. Importantly, the mechanism for nucleophile formation in Pols is yet unclear and debated.[1,121] In Pol's catalysis, the formation of the 3′-hydroxide ion is the very first chemical step to trigger a nucleophilic attack on the incoming nucleotide (see Fig. 4.1), which is bound to the enzyme thanks to a large conformational change for Watson–Crick nascent base pairs, as explained well for DNA polymerase-β catalysis.[122]

A first mechanism for nucleophile formation is via an Asp residue, which is part of the conserved DED motif that coordinates the two catalytic metal ions in Pols.[19] This residue can act as a general base for 3′-OH deprotonation, as shown by Warshel and collaborators for DNA polymerase of bacteriophage T7 (protein-activated mechanism).[121,123] Alternatively, the 3′-OH may be deprotonated via a transient bulk water molecule, which can then shuttle the migratory proton on the α-phosphate of the nucleotide, as first reported for catalysis in the lesion-bypass Dpo4 and Pol-κ enzymes [water-mediated and substrate-assisted (WMSA) mechanism].[77,78] Both of these mechanisms imply a stepwise catalytic process made by two formally independent chemical steps, i.e., nucleophile formation and subsequent NTP addition.

Here, bioinformatics analysis of all structures of ternary DNA/RNA Pols complexes (from all domains of life) reveals a previously unrecognized structural determinant that could play a key role in Pol catalysis and that, remarkably, is missing from all previous mechanistic proposals.[77,123] This crucial element is the intramolecular H-bond formed by the nucleophilic 3′-OH and the β-phosphate of the incoming nucleotide (distance d-PT in Fig. 4.2), which is consistently present across all the currently available structures of Pols adducts that include the (d)NTP (see “Results and Discussion” section).

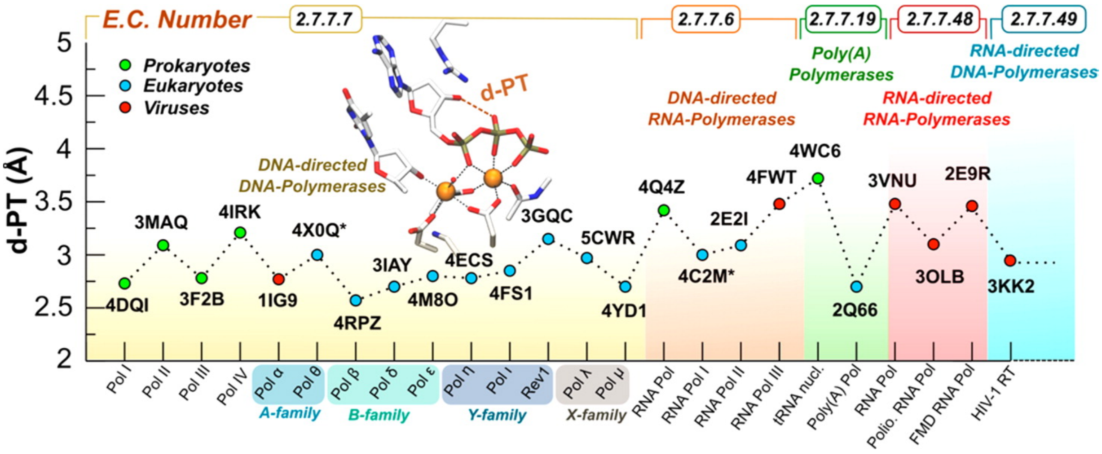


Fig. 4.2. Graph reporting the intramolecular H-bond d-PT in different polymerases. The length of d-PT is reported for structures of Pol families from each domain of life. The *X*-axis reports the protein name. The *Y*-axis reports d-PT (Å). Green dots identify X-ray structures (PDB IDs) of Pol from prokaryotes, cyan from eukaryotes, and red from viruses. The background color indicates the enzyme commission number (EC number provided above).

Importantly, we also found that such a short H-bond is favored only when the sugar pucker of the incoming nucleotide adopts its reactive C3′-endo conformation in the Michaelis complex, characterized by the intramolecular H-bond d-PT. Indeed, as reported by Schulten and co-workers,[124] NTP dispersed in solution adopts a more relaxed conformation that does not favor the formation of this H-bond, which therefore defines a productive state of DNA/RNA Pols when complexed with their substrates.[8,125] On the basis of these observations, we propose the following novel catalytic mechanism for nucleic acid polymerization in Pols. First, the key intramolecular H-bond in the incoming nucleotide prompts the in situ 3′-OH activation via its deprotonation in favor of the leaving PP_i (points B, C, Fig. 4.3).

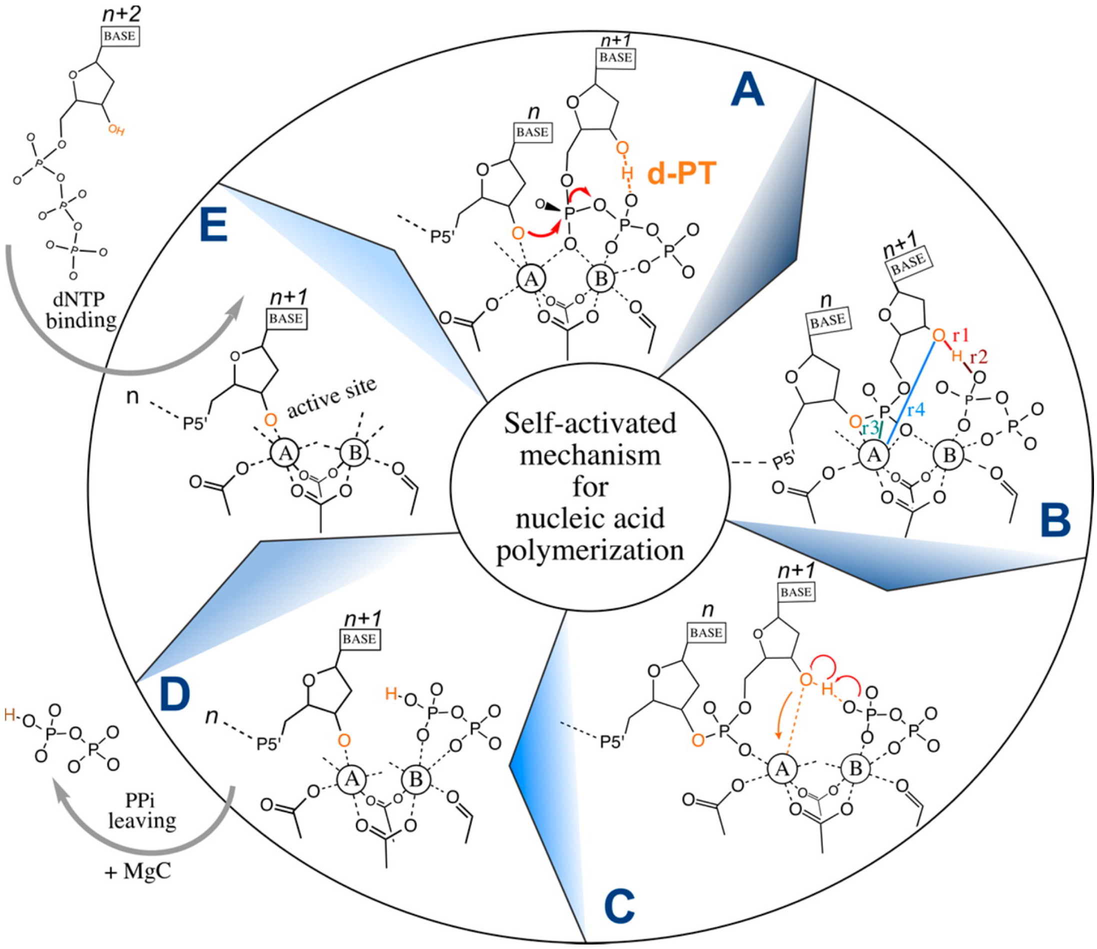


Fig. 4.3. Reaction scheme for the proposed self-activated mechanism (SAM) for nucleic acid polymerization. (A) Michaelis–Menten complex: This state leads to the two-metal-aided SN2-type phosphoryl transfer with liberation of pyrophosphate (PPi) leaving group. Notably, the nucleophilic oxygen is here already activated (deprotonated). (B) Products for nucleotide addition: Here, the incoming nucleotide was added to the primer strand. Colored lines indicate selected distances taken as collective variables (CV1 = r1 – r2 and CV2 = r3 – r4 for QM/MM metadynamics) to investigate SAM. (C) Nucleophile formation and nucleic acid translocation: the nucleophile 3′-OH is activated through its deprotonation in favor of the leaving PPi (PT1), while r4 is progressively shortened, indicating initial nucleic acid translocation. (D) PPi exit: at this point, the newly formed 3′-hydroxide group of the incoming nucleotide is coordinated on top of metal A, while the leaving PPi departs from the catalytic site, helped by the transient third metal ion. (E) dNTP binding and catalytic site closure: the enzyme is ready for the subsequent polymerization cycle upon binding of a new nucleotide, with closure of the catalytic cycle.

Then, the newly formed 3′-hydroxide ion in the incoming nucleotide slowly moves on top of MgA (points C, D, Fig. 4.3) during DNA translocation, assuming the typical coordination required for in-line nucleophilic attack and nucleotide addition, according to the two-metal-ion mechanism.[1,19] In this way, the catalytic cycle is closed and the enzyme is ready for the subsequent round of nucleic acid polymerization (points E, A, Fig. 4.3).

Thus, we describe a new mechanism characterized by a concerted closed-loop catalytic sequence of steps for nucleophile formation, nucleotide addition, and, importantly, nucleic acid translocation. These are synergistically interconnected chemical and physical steps that form a novel enzymatic mechanism for Pol catalysis. Hereafter, we refer to this mechanism as the "self-activated mechanism" (**SAM**) because it is initiated by a proton transfer for nucleophile formation that occurs within the incoming nucleotide for nucleic acid elongation.

4.2 Computational materials and methods

4.2.1. Structural model and Car-Parrinello QM/M simulations

Our ternary Pol-η/DNA/dNTP model system is based on the crystallographic structure of the enzyme structure after completion of the nucleotidyl-transfer reaction and consequent formation of products (PDB ID 4ECW).[8,126] This structural model was used here to verify the coupling between nucleophile formation and DNA translocation, as proposed in **SAM**. Toward this end, we performed ab initio CP simulations, in the QM/MM implementation,[57] coupled with metadynamics-based free-energy calculations[127] of Pol-η catalysis. As these are enhanced sampling simulations, they cannot provide information on the time scale of the events. The reactive region of the ternary complex was treated at the DFT/BLYP level and includes the Mg^{2+} coordination sphere (DED motif: D13, E115, D118, M14), part of the DNA dA:dT, dT_{-1} nucleotides, R61, pyrophosphate, and solvation water molecules (for a total of 183 QM atoms, Fig. 4.5).

The remaining part of the complex (∼70 000 atoms) was treated using the Amber force field. The valence electrons are described by a plane wave basis set up to a cutoff of 70 Ry. A $20 \times 20 \times 18$ Å^3 cell includes the QM part of the system. The interactions between valence electrons and ionic cores are described with norm-conserving Martins–Troullier pseudopotentials. CP QM/MM dynamics is carried out with a time step of 0.12 fs (for a total simulation time of ∼250 ps, including plain, steered, and metadynamics QM/MM simulations) and a fictitious electron mass of 500 au; constant temperature simulations are achieved by coupling the system with a Nose′–Hoover thermostat at 500 cm^{-1} frequency. The interactions between the MM and QM regions are coupled in a Hamiltonian scheme as discussed by Laio et al.[57] Notably, a rigorous Hamiltonian treatment of the electrostatic interaction between QM and MM regions is used, as in ref. 81. The approach has been shown to accurately describe a variety of metal-dependent enzymes[91,128-133] and, specifically, protein–DNA complexes.[1,76,88]

The CP QM/MM protocol includes an initial equilibration phase, followed by a short run where only the MM part is free to move, while the QM part is kept frozen. Notably, the starting configurations were retrieved from our recent microsecond-long classical MD study.[126] Then, the whole system is allowed to move and heat up to 300 K (∼2 ps).

Trajectories are then collected for analysis. Configurations from the equilibrated CP QM/MM simulations are used for free-energy calculations. Specifically, we used the extended Lagrangian metadynamics techniques in the context of first-principle simulations to reconstruct the free-energy landscape associated with nucleophile activation and DNA translocation. The free energy was determined as a function of two selected collective variables (CVs; see Fig. 4.5) that identify the main motions taken into consideration. CV1 is defined as the difference between the length of the breaking 3′-O–H bond (r1) and that of the forming H–O_{PPi} bond (r2). CV2 is the difference between the length of the breaking P^{α}–MgA (r3) and that of the 3′-O^{-}–MgA interaction (r4). The Gaussian function deposition rate was set to 24 fs. The initial hills height and width were set to 0.05 kcal mol^{-1} and 0.01 Å, respectively. A total of ~600 Gaussians were deposited from A to D, in two replica systems (~120 000 steps). The Lagrangian simulations were carried out until their convergence (~60 000 steps per replica), i.e. the progressive stabilization of the energetic minima on the free-energy surface (see Supplementary Fig. S4.7, Appendix Chapter 4). All other parameters correspond to those used for the plain QM/MM MD simulations described above. See the Appendix Chapter 4 for further details on the computational setup and calculations.

4.3 Results and Discussion

First and most importantly, we identified a previously unrecognized and conserved H-bond formed by the nucleophilic 3′-OH and the β-phosphate of the incoming nucleotide (distance d-PT in Fig. 4.2 and 4.4) in all the currently available structures of Pols ternary complexes, with values from ~2.50 to ~3.75 Å (see Fig. 4.4 and Supplementary Tab. S4.1, Appendix Chapter 4). On the basis of this experimental evidence, we propose a new catalytic mechanism for nucleic acid polymerization, which is characterized by a d-PT-prompted proton transfer for in situ 3′-OH activation (**SAM**, see Fig. 4.3). Here, we define **SAM** in human DNA polymerase-η (Pol-η) catalysis, aided by the wealth of structural and kinetics data on this important enzyme.[61,71,93] Pol-η is a trans-lesion Pol that catalyzes elongation of DNA affected by UV-induced cyclobutane–pyrimidine dimers (CPDs),[62,99] which are related to skin cancer onset.[33,92]

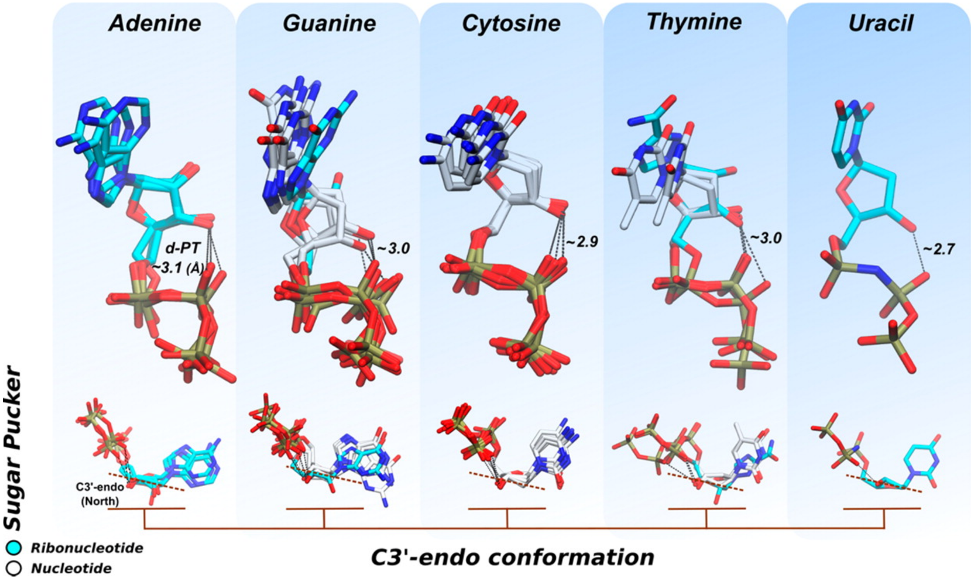


Fig. 4.4. Superimposition of (ribo)nucleotides co-crystallized in Pol's reactive ternary complexes. Structures extracted from different crystals (in Fig. 4.2 and Supplementary Tab. S4.1, Appendix Chapter 4) are superimposed following their species (A, C, G, T, U). The upper part indicates the conserved presence of the intramolecular H-bond (d-PT) in those (ribo)nucleotides complexed with Pol/DNA(RNA) binary complexes. The lower part shows the C3′-endo sugar pucker conformation always detected in those structures. Ribonucleotides (RNA) are cyan. Nucleotides (DNA) are white. Value reported for d-PT is the average value obtained for each type of (ribo)nucleotide.

Recent high-resolution time-resolved X-ray structures of the ternary Pol-η/DNA/dNTP complex have shown the incoming dNTP assuming its reactive C3′-endo sugar pucker conformation, which allows a short (2.78 Å) intramolecular H-bond formed by the nucleophilic 3′-OH and the β-phosphate of dNTP[8] (distance d-PT in Fig. 4.2). According to **SAM**, this H-bond d-PT, together with the initial DNA translocation, facilitates the deprotonation of the 3′-OH in favor of the β-phosphate (r1 and r2, Fig. 4.3 and 4.5) of the incoming dNTP (points B–C, Fig. 4.3). At this point of the catalytic cycle, the forming interaction between MgA and the approaching 3′-OH group is known to facilitate 3′-hydroxide formation by lowering the pK_a of the 3′-OH within the protein environment (typically ~7.5–10.5 instead of ~10.5–12.5).[122,134,135] Thus, the progressive decrease of the 3′-O$^-$-MgA distance during **SAM** (r4, Fig. 4.3 and 4.5) implies a significant electrostatic influence of the metal ion on the ionization state of the 3′-OH and nearby residues/groups,[136-139] as comprehensively explained by Warshel and collaborators for other nucleotidyltransferases undergoing significant conformational changes.[123,140,141] Thus, within **SAM**, the electrostatic attraction of the forming hydroxide ion with MgA helps DNA translocation. This was also demonstrated qualitatively by ab initio steered MD simulations and Car–Parrinello (CP) quantum mechanics/molecular mechanics (QM/MM) metadynamics, which consistently indicated that DNA translocation (i.e., shortening of the distance r4) is favored when in the presence of the activated 3′-O$^-$ group, compared to the case with the nucleophile 3′-OH still protonated (see Supplementary Fig. S4.1, Appendix Chapter 4). Indeed, the X-ray structure of Pol-η, in a state preceding nucleotide addition and DNA translocation (PDB ID 4ECS), has P$^\alpha$–MgA (r3, Fig. 4.3 and 4.5) and r4 distances equal to 3.42 and 7.05 Å, respectively.

Then, the postreactive structure of Pol-η (PDB ID 4ECW) shows r3 increased to 6.14 Å and r4 diminished to 2.29 Å, which reflect initial DNA translocation, with the complete translocation of the 3′-end after the breakage of the P$^\alpha$–MgA interaction. In this way, **SAM** leads to the (re)formation of an optimal 3′-O$^-$–MgA coordination, with the newly formed nucleophilic 3′-O$^-$ properly placed to perform the subsequent nucleophilic attack at the incoming nucleotide.[121] Thus, SAM infers a closed-loop catalytic cycle, in which the S$_N$2-type phosphoryl transfer for nucleotide incorporation in Pols ends by originating a new 3′-hydroxide group that, in turn, initiates the following catalytic addition of the next incoming nucleotide, after DNA translocation and PP_i departure (points A–E in 4.3).

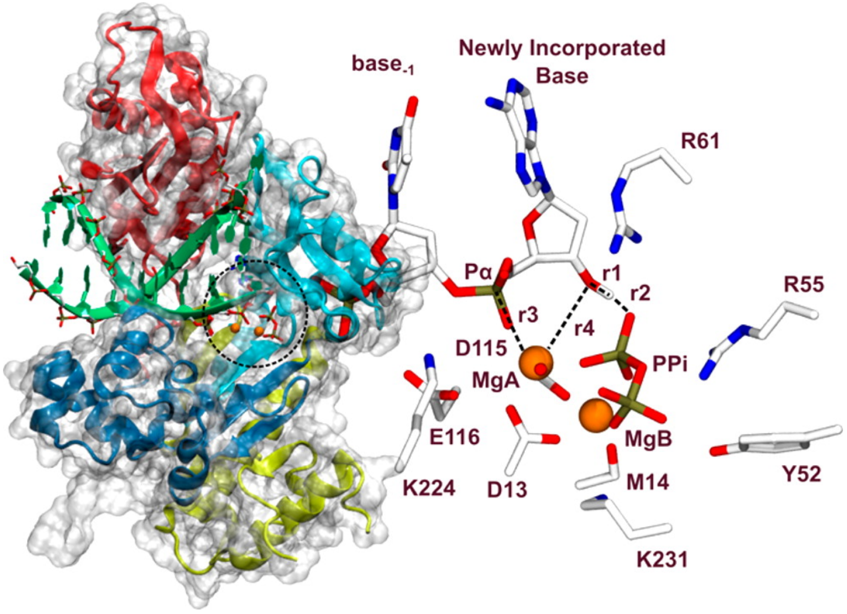


Fig. 4.5. Human DNA Pol-η structure after incorporation of the incoming base. (Left) Overview of the ternary Pol-η/DNA/(d)NTP complex. Each domain of Pol-η is a different color: palm, yellow; thumb, blue; fingers, cyan; and little finger, red. (Right) Close view of the catalytic site of Pol-η. The two Mg2+ ions are in orange, nitrogen is in blue, carbon is in white, oxygen is in red, and phosphorus is in maroon.

Remarkably, similar values and variation of r3 and r4 are found in X-ray structures of several other Pol reactive complexes, further suggesting a closed-loop catalytic sequence of both chemical and physical steps formed by nucleophile formation, nucleotide addition, and DNA translocation, as proposed in **SAM**. For example, bacteriophage N4 RNA-Pol is an enzyme recently studied by means of time-resolved X-ray crystallography to capture real-time intermediates in the pathway of transcription.[142] The series of crystallographic structures for bacteriophage N4 RNA-Pol shows RNA extension, from prereactive to postreactive states. In this case, the prereactive complex (PDB ID 4FF3) has r3 and r4 equal to 3.79 and 7.34 Å, respectively. These two distances correspond to 4.31 and 6.08 Å in the postreactive structure (PDB ID 4FF4), indicating initial nucleic acid translocation and formation of nucleophile–MgA coordination. These data further support the key role for Pol's catalysis of an intimate interconnection between the physical step for nucleic acid translocation and the chemical steps for nucleophile formation and nucleotide addition, as proposed in **SAM**.

Taken together, this structural evidence and extensive conservation between DNA and RNA Pols suggest an evolutionary convergence to preserve those specific structural features that are key to nucleic acid binding and processing in Pols. There are the conserved DED motif,[1,76] multiple catalytic Mg^{2+} ions,[88,143] a positively charged residue in the active site[126] and, ultimately, a short d-PT, which (according to **SAM**) is needed to trigger the 3′-OH deprotonation for nucleophile activation. Hence, **SAM** is remarkably different from previous mechanistic hypotheses of Pols catalysis. This is because **SAM** is characterized by a synergistic interplay between chemical (i.e., nucleophile formation and nucleotide addition) and physical (i.e., nucleic acid translocation) steps to form a closed-loop cycle for efficient Pols catalysis.[144,145]

QM/MM Simulations of Nucleophile Activation in Pol-η Catalysis

To further corroborate **SAM**, we next performed ab initio CP QM/MM simulations coupled with metadynamics-based free-energy calculations[127] of Pol-η's catalysis. This allowed us to determine the dynamics and semiquantitative energetics of **SAM** for nucleic acid extension in Pol-η. Here, we analyzed only the coupling between the chemical and physical steps for nucleophile formation and nucleic acid translocation (points B–D in Fig. 4.3), which precede the already well-characterized S_N2-like phosphoryl-transfer reaction for nucleotide addition[1,77,120,123] (point A in Fig. 4.3). Thus, we first investigated the proton-transfer along d-PT for in situ formation of the catalytically active 3′-hydroxide ion, using two selected collective variables (CV1 and CV2). CV1 is defined as the difference between the lengths of the breaking 3′-O–H (r1, Fig. 4.3. and 4.5) and forming H–O_{PPi} (r2) bonds; CV2 is the difference between the lengths of the P^{α}–MgA (r3) and the 3′-O^{-}–MgA (r4) coordination bonds. The free-energy surface (FES, Fig. 4.6), projected on those CVs, shows that our starting system was initially located in a metastable state B, retrieved by previous extensive MD simulations connecting pre- and postreactive states.[62,99,126] Thus, as expected, the system quickly fell from B into a large minimum D, where the 3′-hydroxide was fully formed, while the leaving PP_i was stably protonated (see Fig. 4.3).

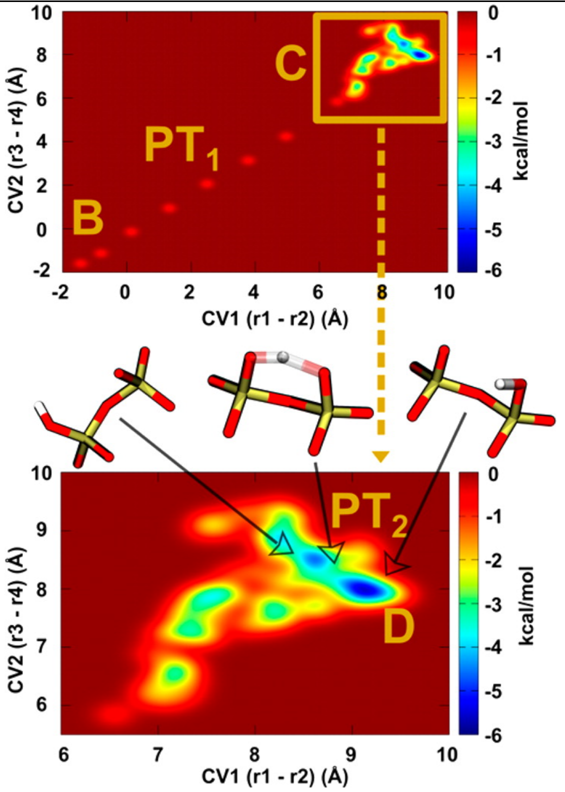


Fig. 4.6. Free-energy surface for **SAM** in human DNA Pol-η. B, PT_1, PT_2, C, and D identify saddle points for **SAM**-catalyzed nucleic acid polymerization in DNA Pol-η, moving from point B of the catalytic cycle to an ensemble of global minima at point D (see reaction scheme and points B and D in Fig. 4.3).

Importantly, two proton transfers occurred moving from B to D. First, the proton transfer for the self-activation of the nucleophile 3′-O^- occurred at PT_1. Then, the transferred proton was shuttled further away on the departing PP_i through a second proton transfer PT_2, before the systems fell into **D** (see Fig. 4.6). In detail, in B (CV1 ~ −4.0 Å and CV2 ~ −3.0 Å), the system was only ~1.2 kcal/mol more stable than its surrounding conformational space. However, a well-structured H-bond network centered on the catalytic Mg^{2+} ions stabilized the overall architecture of Pol-η's catalytic site. In B, r3 was 3.28 Å, reflecting a stable P^{α}–MgA coordination. The distance r4 was 5.01 Å, close to the value detected in the X-ray structure of the postreactive state conformation (PDB ID 4ECW, r4 = 7.05 Å). Also, the conserved surrounding residues R61, R55, Y52, and K231 formed a distinctive XRYK-motif centered on the PP_i. From B to PT_1, the system overcame a series of four small energetic barriers (~1 kcal/mol each, Fig. 4.6). Then, we observed the in-line 3′-O–H–

O_{PPi} proton transfer PT_1, with a barrier of ~2.0 kcal/mol, leading to the final 3′-hydroxide.

Notably, the protonation of the leaving PP_i was also observed in other similar enzymatic reactions, where the leaving PP_i served as the final proton acceptor for nucleophile formation.[77,78,134,146,147] Here, the 3′-OH deprotonation event is well-captured by r1 and r2, which gradually changed from 1.02 and 2.58 Å in B to 1.42 and 1.07 Å in PT_1, respectively. Interestingly, at this point, the variation of r3 and r4 (of 3.75 and 3.55 Å, respectively) reflects the shift of the newly generated $3'\text{-O}^-$, which slowly moved on top of MgA, while the phosphate group of the 3′-terminal base slid away (points B, C, Fig. 4.3). Altogether, this indicates an initial DNA translocation, which occurs concomitantly to nucleophile formation (see below). Also, during DNA translocation in SAM, the two catalytic metal ions increase their initial internuclear distance from 3.36 ± 0.14 Å in point A to about 4 Å in point B. Then, after DNA translocation, the two ions slowly return to their initial internuclear distance of ~3.5 Å, moving from C to D–E–A to stabilize the transition state along the phosphoryl transfer for nucleotide addition. Noteworthy, the cooperative motion of the two catalytic ions was reported for other nucleic acid-processing two-metal-ion enzymes.[1,76,120,148,149] Clearly, additional costly simulations of the overall catalytic cycle are needed to better establish the level of synchronicity and synergy of **SAM's** chemical and physical steps.

From PT_1, the system evolved toward PT_2 (CV1 ~ 6.5 Å and CV2 ~ 5.8 Å). This second intramolecular proton transfer PT_2 occurred from the β-group to the adjacent γ-group of the PP_i, with a barrier of ~2.0 kcal/mol. PT_2 is also shown by r1 and r2, which became ~10.5 and ~4.0 Å respectively, while r3 and r4 changed to ~5.8 and ~3.5 Å, further suggesting the initial DNA translocation. Precisely, the proton previously shuttled to O_{PPi} from 3′-OH in PT_1 was rotated by about ~270° with respect to its donor species. In this way, this proton pointed toward one of the nonbridging oxygen atoms of the γ-phosphate of PP_i. From here, it was then quickly shuttled (PT_2) on the adjacent phosphate of the PP_i, where it stably remained for the rest of the simulations. This protonation state of the PP_i was also found for T7 DNA polymerase catalysis,[77] further confirming the likely role of the PP_i as the ultimate acceptor of the shuttled proton generated by the 3′-OH deprotonation. Immediately after PT_2, the system rapidly fell into the deepest energetic minimum D of the FES (CV1 ~ 7.5 Å and CV2 ~ 6.0 Å), which is at ca. −6.0 kcal/mol (see Fig. 4.6). This energetic minimum was confirmed by additional ~25 ps of unbiased QM/MM simulations, during which the architecture of the metal-aided catalytic site, as well as the transferred proton on the PP_i γ-group, were maintained, matching well the crystallographic prereactive state of Pol-η (PDB ID 4ECS, RMSD ~ 3.0 Å; see Supplementary Fig. S4.2 in Appendix Chapter 4).

Notably, our calculations provide only a thermodynamics description of the process under investigation, while the overall relaxation step of the whole ternary complex, after PP_i release, is suggested to be the rate-limiting step of the polymerization process catalyzed by human Pol-η, as already proposed for the structurally similar Y-family members Dpo4 and Pol-κ.[93,103,109] The overall relaxation step of the whole ternary complex is therefore likely to remain the rate-limiting step of **SAM**, although this point remains to be clarified by further investigations. In addition, the recent time-resolved crystallographic structures of Pol-η have revealed a transient third ion bound at the catalytic site after nucleotide insertion.[22,23,26] This third ion is suggested to facilitate product formation during nucleotide addition and, as also proposed by our previous MD simulations, to serve as an exit shuttle for the leaving PP_i. In this respect, we preliminarily evaluated the effect of the third ion in **SAM**.

First, additional QM/MM simulations demonstrated that this transient third ion hampers nucleophile formation and DNA translocation, if bound to the pretranslocation complex, point B in Fig. 4.3 (see Supplementary Fig. S4.3 in Appendix Chapter 4). This explains the structural evidence that a third metal ion cannot be placed in the reactant enzyme–substrate complex, mainly because of steric clashes. On the other hand, further QM/MM simulations revealed also that a third ion bound at the catalytic site of Pol-η in the product state, i.e. after nucleophile formation and nucleotide addition, facilitates the exit of the PP_i leaving group, while preventing the reverse reaction of pyrophosphorolysis (see Supplementary Fig. S4.4 in Appendix Chapter 4).

These results further corroborate the evidence that the third metal can be transiently bound only at the product state during catalysis. Therefore, the key initial steps in **SAM** (i.e., nucleophile 3′-O^- formation and DNA translocation) do not require a transient metal ion that, again, was in fact experimentally found only in the products. This puzzling and nascent concept of a functional and cooperative dynamics of multiple catalytic metal ions for DNA and RNA processing undoubtedly merits further studies.[1,22]

Often, DNA polymerases contain a highly flexible positively charged residue, like an arginine or a lysine, which is conserved and located near the catalytic site. This residue is Arg61 in Pol-η.[61,71,80,126] We analyzed the role of this residue in **SAM**, and found that Arg61 stabilizes the negatively charged 3′-hydroxyl nucleophile, when it adopts what is referred to as the "A" conformation (A-conf). This conformation is characterized by bifurcated hydrogen bonds established with the leaving group PP_i. This likely prevents a back-proton migration from the protonated PP_i to the active nucleophile. Arg61 in the "C" conformation (C-conf), where it forms two H-bond interactions with the incoming base, generates an approximately 6.0 kcal/mol higher barrier for nucleophile formation and initial DNA translocation, compared to the system with Arg61 in A-conf (see Supplementary Fig. S4.5, Appendix Chapter 4). Thus, A-conf favors the nucleotide incorporation, while C-conf guides the incoming base into the catalytic site and assists PP_i departure toward the solvent-exposed part of the cavity, as previously reported (see Chapter 3).[126]

Overall, the present work does not rule out other possible mechanisms for the 3′-OH deprotonation in Pol-η (see Supplementary Fig. S4.6 in Appendix Chapter 4) and other previously reported mechanisms for Pol's catalysis.[77,123] Indeed, the WMSA mechanism remains a valid hypothesis for Pol-η's catalysis given the persistent presence, in the recent crystals, of a bulk water molecule properly located to act as a general base for nucleophile deprotonation.[8,22,77,78] However, we found that nucleophile formation via this bulk water molecule is energetically unfavored compared to SAM (see Supplementary Fig. S4.6 in Appendix Chapter 4). Indeed, other transient bulk waters, as well as surrounding residues, could in principle accept the proton from the nucleophile 3′-OH group.[121,123,150] However, when compared to these previously proposed mechanisms, we underline that only **SAM** does (i) account for the absolutely conserved intramolecular H-bond d-PT at the active site of DNA/RNA Pols, formed within the incoming nucleotide, and (ii) imply a highly efficient coupling of DNA translocation with nucleophile formation for enzymatic nucleic acid polymerization.

4.4 Conclusions

We propose a novel self-activated mechanism for efficient polymerase catalysis, which is based on the identification of an evolutionary convergence to preserve a key enzymatic structural element in all the available X-ray structures of DNA/RNA polymerases from all domains of life. This is a structurally conserved H-bond formed by the nucleophilic 3′-OH and the nonbridging oxygen of the β-phosphate in the incoming nucleotide, in the Michaelis complex only. **SAM** is characterized by the synergistic interplay of in situ nucleophile formation (via 3′-OH deprotonation), nucleotide addition, and, importantly, DNA translocation. Thus, **SAM** allows formation of a closed-loop catalytic cycle characterized by a concerted sequence of steps of an elegant and efficient nucleic acid polymerization, as shown here by our analyses of polymerase structures and by our simulations of DNA elongation catalyzed by Pol-η. Importantly, on the basis of the extensive structural conservation of RNA and DNA polymerases, we propose **SAM** to be transferable to a broad range of other nucleic acid-processing enzymes.

Supporting Information

Further details concerning CP QM/MM MD setup, metadynamics and additional results are available in the Appendix Chapter 4.

Acknowledgments

M.D.V. thanks the Italian Association for Cancer Research (AIRC) for financial support (MFAG no. 14140). The authors gratefully acknowledge the computing time granted by the John von Neumann Institute for Computing (NIC) and the JARA-HPC Vergabegremium, provided on the supercomputer JUROPA (Project id HGR25) and on the JARA-HPC Partition part of JURECA (Project id JIAS52) at Jülich Supercomputing Centre (JSC). We thank Grace Fox for proofreading and copyediting the manuscript.

Chapter 5

Newly-identified catalytic elements expand the two-metal-ion architecture of DNA and RNA processing enzymes.

Vito Genna, Matteo Colombo, Marco De Vivo and Marco Marcia

Manuscript Submitted for Publication

Abstract

Synthesis and scission of phosphodiester bonds in DNA and RNA regulate vital processes within the cell. Enzymes that catalyze these reactions operate mostly via the recognized two-metal-ion mechanism, where catalytic metals act in cooperation with surrounding acidic residues. Through systematic sequence and structural comparison, we identified key basic amino acids and monovalent cations optimally placed nearby the active site of two-metal-ion enzymes and ribozymes. Such elements interact with the reactants and orient the substrates into the active site, being indispensable for function according to mutational data. Our analysis suggests an unprecedented extension of the two-metal-ion architecture in DNA and RNA polymerases, nucleases like Cas9, and splicing ribozymes. In spite of different biopolymer scaffolds, size and biological function, these enzymes have preserved previously-unrecognized positively-charged elements at conserved structural positions to facilitate DNA and RNA processing.

5.1 Introduction

Enzymatic cleavage and formation of phosphodiester bonds in DNA and RNA is central to life and health. These reactions allow processing nucleic acids during DNA replication, DNA recombination, DNA repair, transcription, splicing, and defense from pathogens.[20] All these key chemical processes are controlled by vital cellular machineries including both protein and RNA enzymes, such as endo- and exonucleases, DNA and RNA polymerases, and ribozymes (i.e. group II intron and spliceosome) (see Tab. 5.1).

Independent of their biopolymer scaffold, these cellular machineries display a surprising degree of structural similarity,[20,151,152] suggesting convergence of their enzymatic reaction mechanism. Indeed, Steitz and Steitz first described the so-called two-metal-ion mechanism, which is shared among most nucleic acid-processing enzymes.[153] According to this general mechanism, divalent metal ions (typically Mg, Mn and Zn) are chelated by acidic groups of strictly-conserved catalytic residues that counterbalance the negatively-charged phosphodiester substrate.[18,19] Then, one metal (usually referred to as metal A, MgA) stabilizes the activated nucleophile, while the second metal (metal B, MgB) assists the release of the leaving group. Together, MgA and MgB stabilize the trigonal bipyramidal intermediate formed at the transition state.[19] Notably, some enzymes like the ββα-Me and HNH nucleases may operate with only one catalytic metal ion (MgA, which is structurally conserved), while a basic amino acid replaces the missing MgB for catalysis.[154] In all cases, a first-shell structural architecture centered on two conserved and positively-charged elements located in the catalytic site is crucial for efficient DNA and RNA processing.[1]

Over the years, new 3D structures have revealed additional important catalytic elements in specific classes of two-metal-ion enzymes. For example, in DNA polymerases a third transiently-bound divalent cation and a positively-charged, highly-flexible residue (Arg61 in human DNA polymerase-η) facilitate product release.[8,22,126] However, so far no catalytic element other than MgA-MgB and their coordinating acidic residues appeared conserved across different classes of two-metal-ion enzymes.
Recent crystal structures of self-splicing group II intron ribozymes revealed that, besides MgA-MgB and a set of conserved nucleotides, two potassium ions (K1 and K2) are essential for catalysis (see Fig. 5.1-A, Fig. 5.2 and Tab. 5.1).[34,155,156] Interestingly,

the group II intron active site is strikingly similar to that of type II restriction endonuclease BamHI, suggesting the existence of enzymatic amino-acid counterparts of K1 and K2, in spite of their fundamentally different biopolymer scaffold.[34] Such structural analogy between two-metal-ion protein and RNA enzymes would infer a more extended set of functional components for the enzymatic processing of DNA and RNA through the two-metal-ion mechanism. Intrigued by this hypothesis, here we used multiple sequence and structural alignments, molecular modeling and electrostatic potential maps to analyze DNA/RNA-processing enzymes, including splicing machineries like the group II intron and the spliceosome, nucleases like the contemporary Cas9 genome-editing tool, and DNA and RNA polymerases. Remarkably, in all these enzymes we identified two positively-charged residues with a similar spatial localization in the second coordination shell of the two-metal-aided catalytic site. These previously-unrecognized structural elements define a larger and more flexible two-metal-ion-centered enzymatic structure to ensure fidelity, substrate specificity and catalytic efficiency.

Tab. 5.1 Representative two-metal-ion enzymes that possess positively-charged elements in the second coordination shell of the active site. The complete list of enzymes analyzed is reported in Supplementary Table S5.1, Appendix Chapter 5.

Enzyme	PDB id	Enzymatic classification (EC)	K1	d1 (Å)[a]	d-ac (Å)[b]	mutants	functional defect	K2	d2 (Å)[a]	d-sub (Å)[c]	mutants	functional defect
Group II intron (*O. iheyensis*)	4FAR[34]	ribozyme	K1	3.83	2.60	K to Li, Na, Cs[34]	Distortion of active site, splicing defect	K2	7.93	2.78	K to Li, Na, Cs[34]	Distortion of active site, splicing defect
BamHI (*B. amyloliquefaciens*)	2BAM[157]	3.1.21.4	Lys61	5.87	via Tyr65	computational electrostatic model[158]	n.a.	Lys126	6.23	2.90	computational electrostatic model[158]	n.a.
Exo-λ (*Escherichia virus lambda*)	4WUZ[159]	3.1.11.3	Lys131	4.44	2.97	K131A[159]	No exonuclease activity	Arg28	9.55	2.96	R28A[159]	No exonuclease activity
Cas9 (*S. pyogenes*)	4CMQ/5F9R[160,161]	3.1.-.-	Lys974	6.05	3.08	K974A[162]	Altered substrate binding and kinetics	Lys968	8.82	3.79	K968A[162]	Altered substrate binding and kinetics
DNA Pol-η (*H. sapiens*)	4ECS[8]	2.7.7.7	Lys231	3.98	2.68	n.a.	n.a.	Lys224	4.71	3.43	n.a.	n.a.
RNA Pol-II (*S. cerevisiae*)	2E2H[163]	2.7.7.6	Lys752 (Rbp1) Arg1020 (Rbp2)	5.40 3.12	4.03 3.57	K752L[164] n.a.[e]	Lethal n.a.[e]	Lys979 (Rbp2) Lys987 (Rbp2)	8.86 7.13	2.58 3.34	K979R K987R[165]	Lethal Lethal

[a] closest distance between the ion / amino acid corresponding to K1 or K2 and the M_A-M_B center; [b] closest distance between the ion / amino acid corresponding to K1 and the acidic residues that chelate M_A-M_B; [c] closest distance between the ion / amino acid corresponding to K2 and the substrate; [d] to be published; [e] not available.

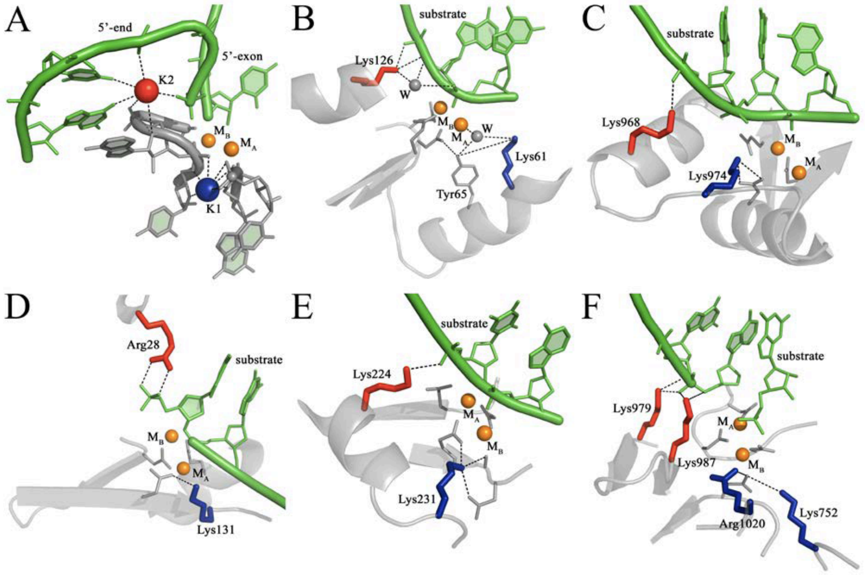


Fig. 5.1. Structural comparison of various two-metal-ion dependent enzymes. The catalytic sites of various two-metal-ion enzymes are depicted: A) group II intron (PDB ID: 4FAR); B) restriction endonuclease BamHI (2BAM); C) RuvC active site of SpyCas9 (combined from 4CMQ and 5F9R); D) exonuclease Exo-λ (4WUZ); E) DNA polymerase Pol-η (4ECS); F) RNA polymerase Pol-II (2E2H). All panels representing the catalytic site are drawn in the same reciprocal orientation. Catalytic divalent ions are depicted as orange spheres, acidic residues coordinating the divalent ions are represented as grey sticks, DNA/RNA substrates as pink sticks and K1- and K2-like elements are depicted as spheres (for ions) or sticks (for amino acids) in blue and red, respectively. Black dashed lines indicate interatomic distances in angstrom. All figures representing crystal or EM structures have been drawn using PyMOL.

5.2 Methods

5.2.1 Sequence and structural alignments

Representative sequences of each enzyme have been selected in BLAST[166] and multiple-sequence alignments have been performed in ClustalOmega.[167] Structure-based sequence alignments of RNA-dependent RNA polymerases have been performed using T-COFFEE,[168] as described previously.

Structural alignments have been performed manually in Coot[169] using the di-nuclear metal center MgA-MgB, the substrates, and/or the coordinating acidic residues in the first resolution shell of MgA-MgB as a guide. The figures depicting the structures were drawn using PyMOL.[170]

5.2.2 Calculation of electrostatic potential maps

Calculations of the electrostatic molecular surfaces were performed by solving a non-linear Poisson-Boltzmann equation with APBS[171] package. For all the systems we used a multilevel grid approach with a fixed grid length of 129 points in each spatial direction obtaining a grid spacing of ≤ 0.365 Å. To generate the electrostatic maps, all biomolecules were treated with a low dielectric medium ($\varepsilon = 2$)[172] and the surrounding solvent as a high dielectric continuum ($\varepsilon = 78.54$). The ionic strength was set to 150 mM corresponding to a Debye length of 8 Å, while temperature was set to 300 K. AMBER package[82] was used to add hydrogen atoms to crystallographic models and the AMBER ff99SB-ILDN[81] force field was used to derive atomic radii and charges. In order to diminish the sensitivity of computations to the grid setup, we have used cubic B-spline discretization method to map the charges of our systems into the grid points and the nine-point harmonic averaging approach to the surface-based dielectric and ion-accessibility coefficients. Finally, we applied Dirichlet boundary conditions by using multiple Debye–Huckel functionality.

5.3 Results

5.3.1 Positively-charged residues may define geometry and electrostatic potential f the active site of two-metal-ion enzyme

In the group II intron, potassium ion K1 anchors the residues that coordinate MgA-MgB and potassium ion K2 stabilizes the 5'-splice junction before catalysis and the scissile phosphate after catalysis.[34,155,156] K1 and K2 produce two discrete regions of positive electrostatic potential (+7.4 kT/e) at 7-9 Å from the active site, which interrupt the negative potential created by first-shell coordinators of MgA-MgB (Fig. 5.3). Replacement of K1 with smaller (i.e. lithium and sodium) or bigger (i.e. cesium) ions prevents MgA-MgB binding thus affecting catalysis.[34] In summary, the K1 and K2 ions rigidify the architecture of the MgA-MgB binding pocket, mediate functional conformational changes during the splicing cycle, and stabilize reactants in the pre- and post-catalytic states.[34]

Stiffening of the active site, modulation of its electrostatic environment, and orientation of the reactants relative to the MgA-MgB center are needed by two-metal-ion enzymes to ensure specificity in substrate recognition and to augment the fidelity of the reaction.[173174175176136]

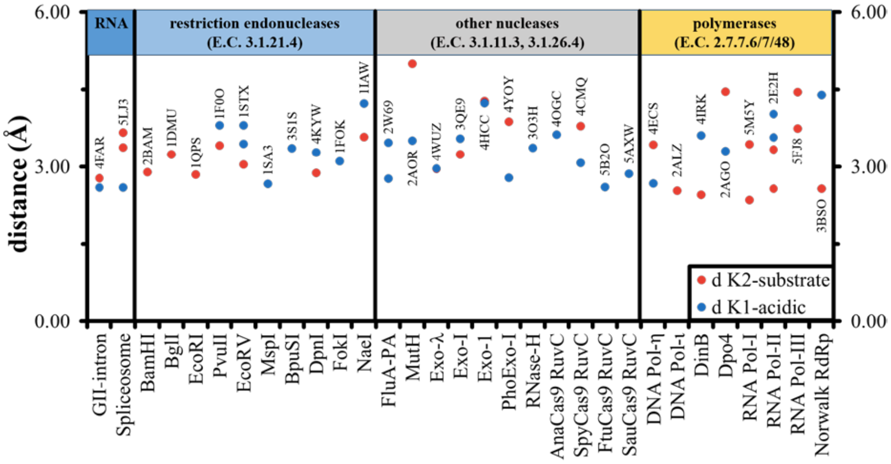


Fig. 5.2 Distances in angstrom of K1- and K2-like elements from the acidic residues that coordinate M_A-M_B and from the substrate, respectively. The names of the enzymes are on the x-axis, corresponding PDB codes are indicated on top of every data point. Enzymes are arranged by classes. Enzymatic Classification (E.C.) numbers are indicated on the top of the graph (related to Tab. 5.1 and Supplementary Tab. S5.1).

Therefore, we reasoned that counterparts of K1 that rigidify the first coordination shell of MgA-MgB and counterparts of K2 that help orienting the substrates into the active site may exist in other families of two-metal-ion enzymes. We thus systematically compared 3D structures of two-metal-ion enzymes to identify putative K1 and K2-like residues in their active sites. Relevant PDB identification numbers, sequence alignments, geometric parameters, and functional data related to the analyzed residues are compiled in Tab. 5.1, Fig. 5.2, Supplementary Tab. 5.1, Supplementary Fig. S5.1, and Supplementary File 1 in Appendix Chapter 5.

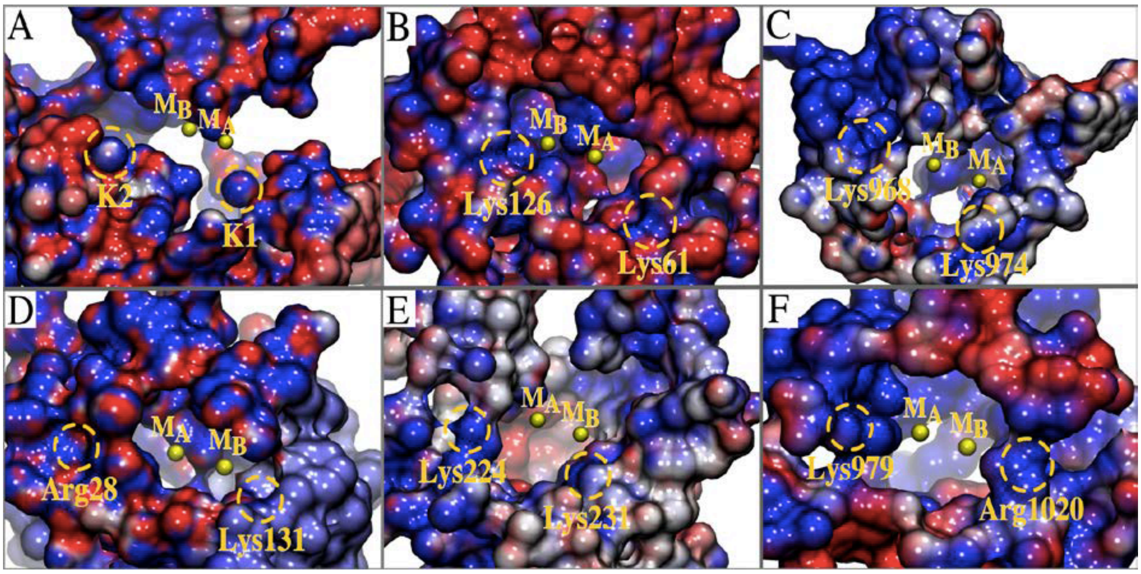


Fig. 5.3 Distribution of the electrostatic potential on the molecular surface surrounding the two-metal-ion active site (radii 15 Å) for: A) Group II intron (PDB ID: 4FAR); B) restriction endonuclease BamHI (PDB ID: 2BAM); C) RuvC active site of SpyCas9 (modeled from PDB ID: 4CMQ and 5F9R); D) Exonuclease-λ (modeled from PDB ID: 4WUZ and 3SM4); E) DNA Polymerase-η (PDB ID: 4ECS); F) RNA Polymerase Pol-II (2E2H). The electrostatic potential ranges from -10 kT/e (darkest red) to +10 kT/e (darkest blue) for all panels.

5.3.2 Functional monovalent cations of the group II intron match functional basic amino acids of endo- and exonucleases, including Cas9

We first superposed the coordinates of *Oceanobacillus iheyensis* group II intron with those of substrate-bound BamHI. Two BamHI lysines match the group II intron K1 and K2 ions (Fig. 1B). Specifically, Lys61 engages in a hydrogen-bond network with catalytic Asp94 via Tyr65,[34] resembling K1, and Lys126 forms a direct contact with the substrate DNA, analogously to K2 (see Tab. 5.1). Lys61 and Lys126 generate an electrostatic potential distribution around the active site similar to that observed for K1 and K2 ions in the group II intron (~8.5 kT/e, Fig. 5.3). Notably, Lys61 and Lys126 are strictly conserved in different bacterial species (Supplementary File 1, Appendix Chapter 5) and are important for function, because they aid cognate DNA binding via large amplitude motions[177] and favorable Coulombic interactions.[158]

Interestingly, a similar structural architecture with positively-charged residues arranged in the second coordination shell of M_A-M_B can also be observed: 1) In other orthodox PD-(D/E)XK restriction endonucleases,[178] such as EcoRI, which produces 5'-overhangs, BglI, which produces 3'-overhangs, and PvuII, which produces blunt ends (Supplementary File 1, Supplementary Fig. S5.1, Supplementary Tab. 5.1 and 5.2, Appendix Chapter 5). In other subfamilies of bacterial restriction endonucleases, such as class IIP endonuclease MspI, class IIG endonuclease BpuSI, class IIM endonuclease DpnI, class IIS endonuclease FokI, and class IIE endonuclease NaeI (see Supplementary Tab. 5.1, Appendix Chapter 5). Such second-shell positively-charged residues are highly-conserved for all these enzymes (Supplementary File 1 in Appendix Chapter 5) and are essential for DNA binding and catalysis,[173,174] playing key functional and structural roles[175] (Supplementary Tab. S5.1 in Appendix Chapter 5). Other bacterial endonucleases, such as MutH, which is involved in DNA repair,[178] or viral endonucleases, such as the PA subunit of influenza virus RNA polymerases (FluA, FluB), which is involved in cap-snatching,[179] also present evolutionarily-conserved, functionally-important residues at K1- and K2-like positions nearby their two-metal-ion active sites (Supplementary File 1, Supplementary Fig. S5.1, Supplementary Tab. S5.1, Appendix Chapter 5).

Second, we identified K1- and K2-like residues in the RuvC domain of Cas9, an RNA-guided DNA bacterial endonuclease associated with the CRISPR type II adaptive immunity system[180] and nowadays largely exploited for genome editing.[181] A complete picture of the RuvC active site of Cas9 can be obtained by superposition of a *Streptococcus pyogenes* Cas9 (SpyCas9) structure that displays the MgA-MgB center but no substrates (PDB ID: 4CMQ) with a SpyCas9 structure that displays the substrate DNA but no divalent metals (PDB ID: 5F9R).[160] Such superposition suggests that Lys974 and Lys968 occupy a very similar structural environment as the K1 and the K2 ions of the group II intron, respectively (see Fig. 5.1-C). Lys974 and Lys968 also produce two regions of positive electrostatic potential on the surface of SpyCas9 (+9.8 kT/e and +13.0 kT/e, respectively) interrupting the negative potential (-9.5 kT/e) that surrounds the catalytic metals. Interestingly, Lys974 and Lys968 have precise counterparts in the Cas9 active site of *Actinomyces naeslundii* (AnaCas9), *Francisella tularensis* (FtuCas9) and *Staphylococcus aureus* (SauCas9), and in RNase-H, which is evolutionarily related to the RuvC domain of Cas9[76,88,160,180] (Supplementary Fig. S5.1, Appendix Chapter 5).

Finally, K1- and K2-like residues are also persistently conserved among exonucleases. *Escherichia* virus lambda exonuclease (Exo-λ), which catalyzes 5’ to 3’ exonucleolytic cleavage in dsDNA,[178] presents Lys131 in a position analogous to the K1 ion and Arg28 in a position equivalent to K2 (see Fig. S5.1-D). Both residues are strictly conserved (see Appendix Chapter 5), have been implicated in catalysis[159] (see Tab. 5.1) and have parallel counterparts in LHK-exonuclease and RecE.[182] Counterparts of Lys131 and Arg28 in human exonuclease-λ (Exo-λ), which catalyzes 3’ to 5’ exonucleolytic cleavage in dsDNA, are Lys85 and Arg92, respectively. Intriguingly, Arg92 forms a bifurcated H-bond interaction with both scissile and 5’-end phosphate groups, as observed for K2 ions in group-II introns. Finally, similar structural determinants are also present in prokaryotic exonucleases, such as Exo-λ of the mesophilic host *Escherichia coli* (conserved His181 and Arg165 in K1- and K2-like positions, respectively) and Exo-λ of the extremophilic host *Pyrococcus horikoshii* (PhoExo-I, Arg142 and Lys136 in K1- and K2-like positions, respectively).

5.3.3 K1/K2-like residues are also strategically located nearby the catalytic site of DNA and RNA polymerases.

Having assessed that K1- and K2-like residues are recurrent in enzymes that catalyze scission of phosphodiester bonds, we then analyzed DNA and RNA polymerases.
We first considered the human DNA Polymerase-η (Pol-η), a Y-family polymerase involved in DNA repair.[8,15,55,63,126] Here, Lys231, which anchors MgA-MgB, is in a position analogous to the K1 ion and Lys224, which orients the phosphate backbone in the proximity of the cleavage site, is in a position analogous to K2 (see Fig. S5.1-E). Both Lys231 and Lys224 are strictly conserved (see Appendix Chapter 5) and are essential for substrate binding and stabilization[67] (see Tab. S5.1). For instance, Lys231 in the pre-reactive complex contribute to properly position the incoming substrate aiding catalysis, while in the post-reactive state, it counterbalances the negative charge of departing pyrophosphate.[126] The surface electrostatic map of Pol-η confirms that Lys231 and Lys224 produce discrete regions of positive electrostatic potential at 8-9 Å from the catalytic metals, closely mimicking the electrostatic effect of K1- and K2-like residues in endo- and exonucleases. Moreover, Lys231 and Lys224 have precise structural and functional counterparts in other human Y-family DNA polymerases, such as DNA polymerase-ι (Pol-ι) and DNA polymerase-κ (Pol-κ), in yeast DNA polymerases, such as Pol-η from *S. cerevisiae*, and in mesophilic and thermophilic bacterial DNA polymerases, such as *E. coli* DinB and *S. solfataricus* Dbh and DPO4 (see Supplementary Tab. S5.1, Fig. 5.2).[108]

Second, we analyzed DNA-directed RNA polymerases (DdRp). In *Saccharomyces cerevisiae* RNA polymerase II (Pol-II), which synthetizes most RNA transcripts in eukaryotic cells,[183] Lys752 (Rbp1 subunit) and/or Arg1020 (Rbp2) occupy the K1 position and Lys979 and/or Lys987 (Rbp2) occupy the K2 position (see Fig. 5.1-F). These amino acids are strictly conserved across Pol-II from yeast to humans, and they are also conserved in RNA polymerase I (Pol-I) and III (Pol-III)[184] (see Appendix Chapter 5). A similar configuration to eukaryotic DdRps is also maintained in viral RNA-directed RNA polymerases (RdRps). Norwalk virus RdRp presents Lys374 in a position equivalent to K1 and Arg392 in a position analogous to K2 (see Supplementary Fig. S5.1, Appendix Chapter 5). Structure-based sequence alignments suggest that Lys374 and Arg392 are highly-conserved in RdRps from other *Caliciviridae* and from other single-stranded positive RNA viruses (viral group IV, according to the Baltimore classification),[185] i.e. *Picornaviridae* and *Flaviviridae* (see Appendix Chapter 5). In other viral families, the presence of such residues is less evident from sequence alignments, but structural superpositions help identifying putative K1- and K2-like counterparts. For instance, in the RdRps from *Orthomixoviridae* and *Bunyaviridae* the role of K1 and K2 may be played by functional residues consistently located in ultraconserved polymerase motifs D, F, and G, which are common to all single-stranded negative RNA viruses (viral group V) and to the reverse transcriptase of retroviruses (viral group VI, Supplementary Tab. S5.1).[186,187]

5.3.4 Criteria to predict the structural position of K1- and K2-counterparts in nucleic acid-processing enzymes and known exceptions.

Our work shows that K1- and K2-like residues are a common catalytic feature of nucleases and polymerases. While some K1- and K2-like residues play idiosyncratic roles in specific enzyme classes (i.e. K1-like Lys144 of BglI and Lys131 of Exo-λ activate the reaction nucleophile;[182] K2-like Arg61 of Pol-η facilitates pyrophosphate departure after catalysis),[126] all K1- and K2-like residues obey the following four general rules, which we found valid across all classes of two-metal-ion enzymes considered here:

1) Conservation in evolution: K1- and K2-like residues are highly conserved for each enzymatic class (Supplementary File 1);
2) Implication in function: Mutation of K1- and K2-like residues cause catalytic defects (Table 1 and Supplementary Table 1);
3) Modulation of surface electrostatic potential: K1- and K2-like residues produce discrete peaks of positive electrostatic potential that discontinues the negative potential surrounding MgA-MgB (see Fig. 5.3);
4) Occupancy of preserved structural positions: The K2-like residue forms ionic or hydrogen-bonding interactions with the reaction substrate and/or product (see Fig. 5.2, Tab. 5.1, and Supplementary Tab. S5.1). Within the typical active site geometry of two-metal-ion enzymes,[19] the K2-like residue is part of the

substrate-binding site 3 (see Fig. 5.4). The K1-like residue forms ionic or hydrogen-bonding interactions with residues of the first coordination shell of MgA-MgB (see Fig. 5.2, Tab. 5.1, and Supplementary Tab. S5.1). As such, the K1-like residue persistently occupies a previously-unrecognized and functionally-important position of the active site of two-metal-ion enzymes, hereafter named “site 4”. Site 4 is structurally juxtaposed to sites 1 and 2, which flank the scissile phosphate on the 5’- and the 3’-sides, respectively, and it is typically located closer to the 3’-end side of the substrate and thus closer to site 2 than to site 1. Relative to the substrate backbone, site 4 is consistently located opposite to the nitrogenous base moieties (see Fig. 5.4).

Based on these criteria, our work allows classifying K1- and K2-like residues in potentially any class of two-metal-ion enzymes. Considering the vastness and heterogeneity of two-metal-ion enzymes, our work also allows identifying exceptional enzymatic classes that do not possess K1- and/or K2-like residues. For instance, class IIF endonucleases present K2 counterparts (i.e. Arg102 in NgoMIV, Arg218 in SfiI, and Arg152 in SgrAI, respectively), but no K1 counterparts, the latter possibly being replaced by conserved lysines that directly coordinate MgA-MgB (i.e. Lys187 in NgoMIV, Lys102 in SfiI, and Lys242 in SgrAI) or by other non-basic residues (see Supplementary Tab. S5.1). DNA polymerase β (Pol-β) also represents an exception, because in this enzyme the MgA-MgB-binding acidic residues are not coordinated by residues near the putative K1-site (Arg183 or Arg149), but rather by the K2-like residue (Arg254, see Supplementary Tab. S5.1).

5.4 Discussion

Here, we have presented a comparison of RNA and protein enzymes that follow a two-metal-ion mechanism of catalysis, including self-splicing group II introns, nucleases and DNA/RNA polymerases. By multiple sequence and structural alignments, molecular modeling and electrostatic potential maps, we have identified two previously-unrecognized positively-charged elements (basic amino acids or monovalent cations) persistently located in the immediate vicinity of the two-metal-aided active site. Preserved spatial localization of these second-shell residues and catalytic defects caused by their mutation (see Tab. 5.1, Supplementary Tab. S5.1, Appendix Chapter 5) suggest that the somewhat obvious presence of such positively-charged residues proximal to the polyanionic phosphodiester backbone of nucleic acids is actually coupled to a functional role. Indeed, these residues do not just contribute to generate an appropriate electrostatic environment, but they specifically help in the chemical step of catalysis by rigidifying the coordination center of metals MgA-MgB and by orienting the substrates correctly into the active site.

Moreover, by interacting directly with the reactants, these residues also participate in the physical steps of the catalytic cycle mediating functional conformational changes, such as the transition from the first to the second step of splicing in group II introns or translocation and nucleotide addition in polymerases.[55,71,188] In specific classes of enzymes, K1- and K2-like residues also play other roles during catalysis. For instance, Lys144 of BglI and Lys131 of Exo-λ, which are analogous to K1, activate the reaction nucleophile, while Arg61 of Pol-η, which is analogous to K2, facilitates pyrophosphate departure after catalysis. Thus, these second-shell residues extend the known architecture of two-metal-ion enzymes beyond those residues that coordinate MgA-MgB and the substrate directly during catalysis (Fig. 5.4).

Built upon the comparison of high-resolution structures (< 3 Å) that display all reactants and metal ions, our analysis also offers novel insights into two-metal-ion enzymes for which only medium-resolution structures have been determined so far. For instance, 3.6-5.8 Å cryo-electron microscopy structures have been obtained recently for the spliceosome, which catalyzes nuclear pre-mRNA splicing in eukaryotes and which is evolutionarily linked to the group II intron. The precise correspondence of active site

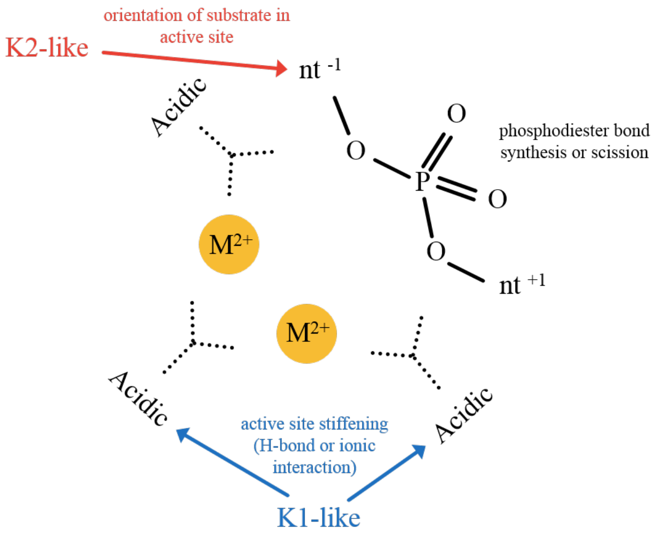

orientation of substrate in active site
K2-like
nt -1
Acidic
O
O
P
O
phosphodiester bond synthesis or scission
O
nt +1
M2+
M2+
Acidic
Acidic
active site stiffening (H-bond or ionic interaction)
K1-like

Fig. 5.4. Structural model recapitulating positions and roles of K1- and K2-like positive charges in the active site of two-metal-ion enzymes. In the active site of two-metal-ion dependent enzymes, strictly-conserved acidic residues ("acidic") coordinate two divalent ions (Mg^{2+}), which catalyze synthesis or scission of phosphodiester bonds in the DNA/RNA substrate ("nt^{-1}" and "nt^{+1}" indicate nucleotides on the 5'-(-1) and 3'-side (+1) of the reaction site, respectively). In all enzymes discussed in this work, such active site is completed by the presence of two positively-charged elements that correspond to the group II intron K1 and K2 ions. These positively-charged elements (basic amino acids or ions) contribute to rigidify the active site for optimal M^{2+} binding and to orient the substrate for catalysis.

structure and reaction chemistry with group II intron suggests that K1 and K2-like elements are preserved in the spliceosome, too.[189,190] Here, a putative K1 cavity essentially identical to the one of group II intron exists between U6-snRNA residues G52, A59, G60, and U80 (see Fig. 5.2 and Supplementary Tab. S5.1). Although not modeled in the current structures, a potassium ion could optimally fit this putative K1 site, which would explain the spliceosomal dependence on potassium for splicing and debranching.[191,192] In addition, Lys611 (PROCN domain of the Prp8 subunit,[193] see Appendix Chapter 5), which stabilizes the curvature of the intron backbone (U2 and U4) proximal to the splicing junction in the C complex, and/or conserved Arg614 (Prp8), which makes contacts with the exon in the C* complex (see Fig. 5.2 and Supplementary Tab. S5.1), seem to constitute optimal spliceosomal counterparts of K2. These residues produce regions of positive electrostatic potential at 7-9 Å from the active site (+6.7 kT/e), similar to the K2 ion in group-II intron (see Supplementary Fig. 56.3).

5.5 Conclusions

In summary, understanding geometric and electrostatic requirements of two-metal-ion dependent enzymes is important for biomedicine and chemistry. Thus, mutagenesis and engineering studies on K1- and K2-like residues may have prospective applicability, for example in designing enzymes with novel substrate selectivity or higher substrate specificity, which would have immediate biotechnological impact, i.e. in the optimization of CRISPR-Cas9 genome editing machineries. Furthermore, design of small molecules that bind to K1- and K2-like sites may lead to identify new antimicrobial or antitumoral drugs with improved specificity against two-metal-ion enzymes.[194] Finally, reproducing the chemical architecture of the K1-K2-M_A-M_B phosphorus-oxide cluster formed by the group II intron may lead to develop nucleic-acid-processing biocatalysts, such as DNAzymes or RNAzymes with important applications in nanoengineering.[195]

Supporting Information

Further details are available in the Appendix Chapter 5.

Acknowledgments

We thank all members of the Marcia and De Vivo labs for helpful discussion. We also thank W. Galej (EMBL Grenoble) for sharing the coordinates of the spliceosomal C* complex from *S. cerevisiae* (PDB ID: 5MPS) before release. Work in the Marcia lab is partly funded by the Agence Nationale de la Recherche (ANR-15-CE11-0003-01) and by the Agence Nationale de Recherche sur le Sida et les hépatites virales (ANRS) (ECTZ18552) and uses the platforms of the Grenoble Instruct Center (ISBG: UMS 3518 CNRS-CEA-UJF-EMBL) with support from FRISBI (ANR-10-INSB-05-02) and GRAL (ANR-10-LABX-49-01) within the Grenoble Partnership for Structural Biology (PSB). MC is funded by the EI3POD (EMBL / EU Marie Curie Actions Cofund) postdoctoral programme. MDV thanks the Italian Association for Cancer Research (AIRC) for financial support ('MFAG n. 14140' and "IG 18883").

Appendix Chapter 3

Cooperative motion of a key positively charged residue and metal ions for DNA replication catalyzed by human DNA Polymerase-η

Vito Genna,[1] Roberto Gaspari,[1,2] Matteo Dal Peraro[3,4] and Marco De Vivo[1,5*]

[1] Laboratory of Molecular Modeling & Drug Discovery,
Istituto Italiano di Tecnologia, Via Morego 30, 16163, Genoa, Italy

[2] CONCEPT Lab.
Istituto Italiano di Tecnologia, Via Morego 30, 16163, Genoa, Italy
[3] Institute of Bioengineering, School of Life Sciences,

École Polytechnique Fédérale de Lausanne - EPFL, Lausanne, Switzerland
[4] Swiss Institute of Bioinformatics - SIB, Lausanne, Switzerland

[5]IAS-5 / INM-9 Computational Biomedicine Forschungszentrum Jülich
Wilhelm-Johnen-Straße 52428 Jülich, Germany

*Corresponding author:
Marco De Vivo
Email: marco.devivo@iit.it

Supporting Information

Table of contents

Supplementary text

Supplementary figures

Supplementary tables

Well-Tempered Metadynamics

In well-tempered metadynamics, a history-dependent repulsive potential, $V_{bias}(s,t)$ (where the quantity s is the vector of our CVs), is built up by adding a biasing potential term, in the form of a small Gaussian "hill", along the CVs after some MD steps:

$$V_{bias}(s,t) = \qquad [1]$$

where δs and ω are the width and height of the Gaussian hills, which are centered at the values of the CVs that have already been visited by the system and deposited at a time interval τ_G. As the hills are build up along the CVs, the system is forced to escape local *minima* exploring therefore higher energy regions of the free energy surface, $F_G(s) = -V_{bias}(s, t \to \infty)$.

Notably, the error associated with this approach is demonstrated be related on the metadynamics parameters, $w, \delta s$ and τ_G, Its origin lies in the false assertion that the estimate of the free energy obtained from the biasing potential tends towards the real and unbiased value. Indeed, once the biasing potential fully compensates for $F(s)$, the addition of extra Gaussian hills serves to introduce error, so that $-V_{bias}(s,t)$ fluctuates around the correct value of $F(s)$, with the magnitude of the fluctuations depending on the size of the hills added. Barducci et al. have proposed an approach called "well-tempered" and "smoothly converging" form of the algorithm, herein used, by which the height of the Gaussian hills added are modified as follows:

$$w = \ \omega e^{-[V_{bias}(s,t)/\Delta T]\tau_G} \qquad [2]$$

where ω is the bias deposition rate with units of energy per unit of time, $V_{bias}(s,t)$ is the estimate of the free energy at the current CVs positions and the current time step (as also described in ref Barducci et al.), and ΔT is a tunable temperature-like parameter that regulates how rapidly w decreases.

In this scheme the final value of the biasing potential is a scaled approximation to $F(s)$:

$$F(s) = -\frac{T + \Delta \mathrm{T}}{\Delta T} V_{bias}(s, t \to \infty) \qquad [3]$$

Through this approach, at the beginning of metadynamics run, the biasing potential is zero and therefore $w = \omega$, which can be quite a large value, allowing the wells to be filled quickly and leading to a fast exploration of phase space. As the wells begin to fill up, ω is scaled and progressively smaller perturbations are made to V_{bias}, allowing it to smoothly converge to a correct approximation of $F(s)$. The value of ω and ΔT are chosen to achieve the efficiency.

Confidence Interval analysis

To determine whether the *d-newbond* distributions are significantly different with respect to the different Arg61 conformations (i.e. **A-**, **B-** and **C-conf**) we performed a confidence interval analysis. The null hypothesis is expressed by $\boldsymbol{\mu}_{\mathrm{X}}=\boldsymbol{\mu}_{\mathrm{Y},}$ where μ represents the mean of the distribution, while X and Y are two different conformations among **A**, **B** and **C**.

In order to test the null hypothesis, we computed the distributions of the *d-newbond* differences, namely $\mathrm{d}_{\mathrm{X}}\text{-}_{\mathrm{Y},}$ for all possible cases and evaluated the mean μ and the associated standard error σ_E. Values are reported in Fig. S3-D, together with the confidence interval, C.I., at which the null hypothesis can be rejected. The populations of **A-**, **B-** and **C-conf** were respectively 10586, 8640, 7950 and were obtained by sampling the MD trajectory every 10 ps, in the various conformations (see Fig. S3, panel D).

Transition State Theory

Transition State Theory (TS-theory) in the Eyring-Polanyi (Laidler,K. and King,C., 1983. *J. Phys. Chem.*, 15, 2657*)* formulation reads:

$$k_{a\to b} = \frac{kT}{h} e^{\frac{\Delta G^*_{a\to b}}{kT}} \qquad (1)$$

where $k_{a\to b}$ generally represents the rate of interconversion from state a to state b, $\Delta G^*_{a\to b}$ is the associated free energy barrier (~8.00 kcal mol^{-1} in our case), kT is the Boltzmann factor (~0.60 kcal mol^{-1} at room temperature), and h is the Planck constant.

In our case state a and state b correspond to the MgB-MgC-PPi adduct bound and unbound to the active site, respectively. In eq. (1) we can set $\Delta G^*_{a\to b}$ to the value estimated by metadynamics simulations (see Result section in the manuscript). TS-theory provides a strongly simplified picture of the PPi unbinding kinetic and serves here as a tool to compare the order of magnitude of the PPi leaving rate with the whole enzymatic turnover rate.

Our estimation of $k_{a\to b}$ is $10^4\ s^{-1}$ to be compared with the experimentally measured PPi leaving constants, k_{cat} = ~25 s^{-1}, in Dpo4, a member of Y-family polymerase (Beckman et al., **2008**, *J. Bio. Chem.*, 283, 36711-36723). Since kcat depends of many enzymatic steps, besides PPi leaving, (see enzymatic turnover from reagents to products), our data are indicate PPi unbinding process as a non-rate limiting step in DNA polymerization catalyzed by Pol-η, in agreement with the findings proposed by Beckman et al. (**2008**, *J. Biol. Chem.*, 283, 36711-36723) and Zhao and collogues (**2014**, *FEBS J.*, 281, 4394-4410) for Y-family members Dpo4 and Pol-ϰ, respectively.

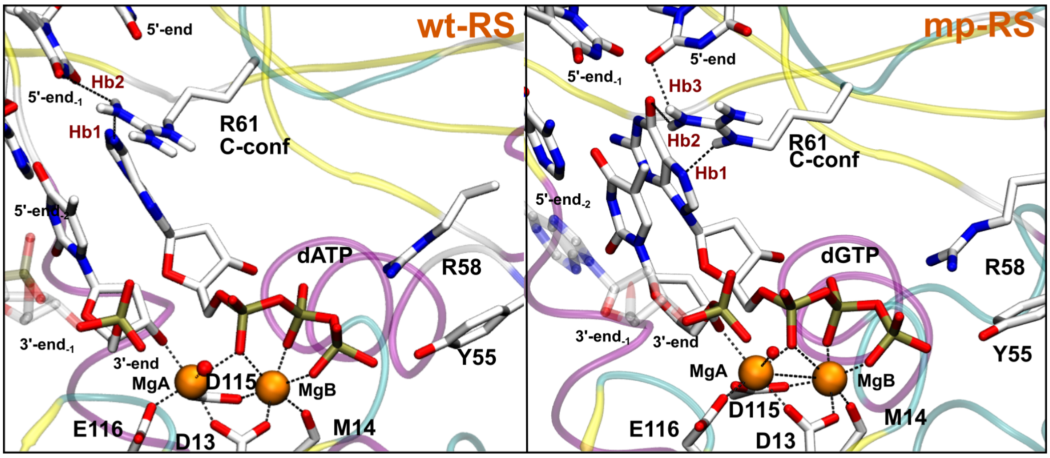


Figure S3.1. Different prereactive states of Poly-η. Left: Wild-type **wt-RS** system showing Arg61 in C-conf, where it establishes two hydrogen bonds (Hb1, Hb2) with the N7 atom of the incoming adenine and the apical oxygen of the templating thymine, respectively. **Right**: Model system containing a dGTP:T mispair, **mp-RS**. Here, Arg61 establishes three hydrogen bonds (Hb1, Hb2 and Hb3): Hb1 with the N7 atom of the incoming adenine; Hb2 with the O atom of the guanine; and, Hb3 with the O2 oxygen of terminal thymine. White is used for carbon, red for oxygen, blue for nitrogen, tan for phosphorus and orange for the two magnesium ions.

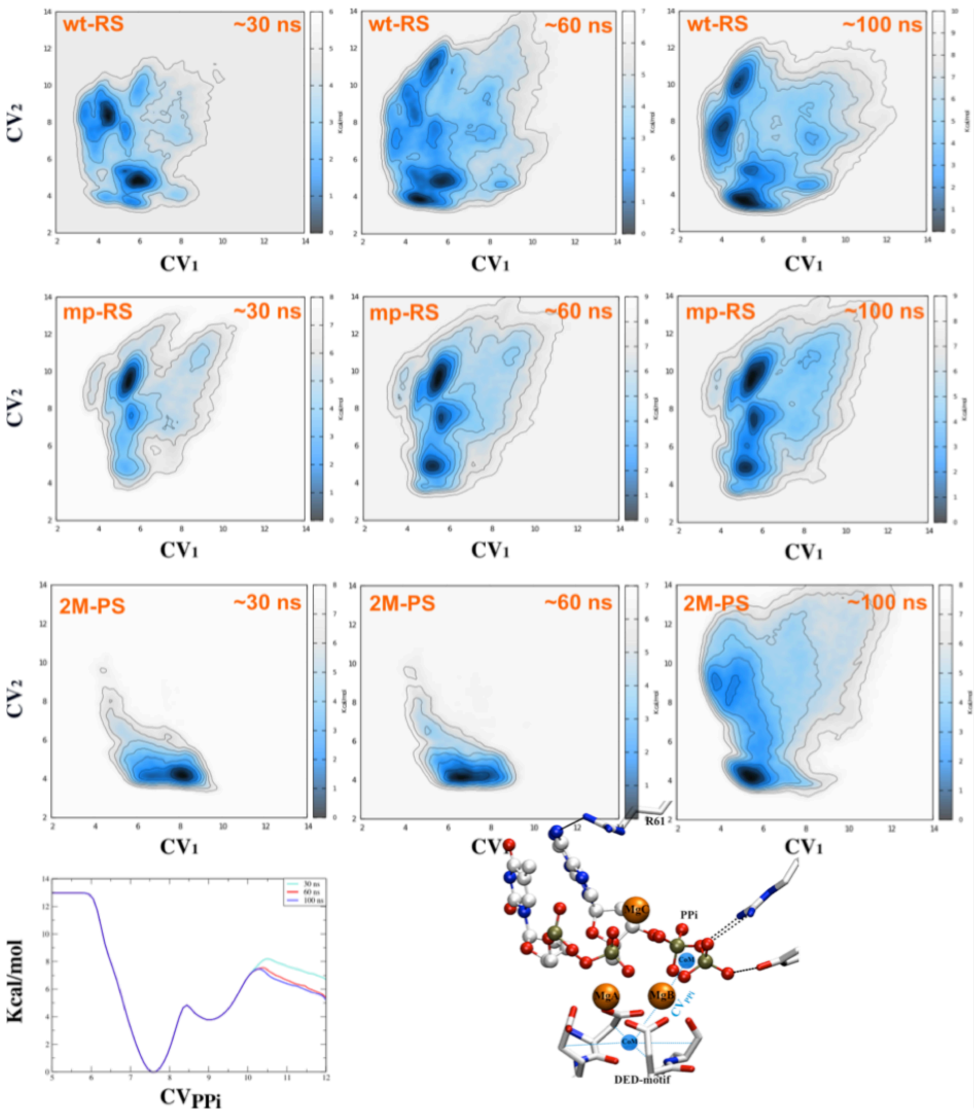


Figure S3.2. Convergence of the free energy simulations. Convergence was checked considering the location of the minima as a function of time. From ~60 ns to ~100 ns, no significant changes were detected. On this basis, we considered converged our well-tempered metadynamics simulations. The ensemble of conformational states of Arg61 found on the reconstructed free energy surfaces reproduced well the overall Arg61 rotamers detected throughout the unbiased MD simulations, improving the overall conformation sampling associated with the motion of this key residue. The postreactive site representation (bottom), indicates the collective variable adopted (CV_{PPi}) during PPi-leaving metadynamics, where the blue spheres indicate the two centers of mass used to define CV_{PPi}. Protein residues are represented in licorice while nucleic acid residues in CPK style. White is used for carbon, red for oxygen, blue for nitrogen, tan for phosphorus and orange for magnesium ions.

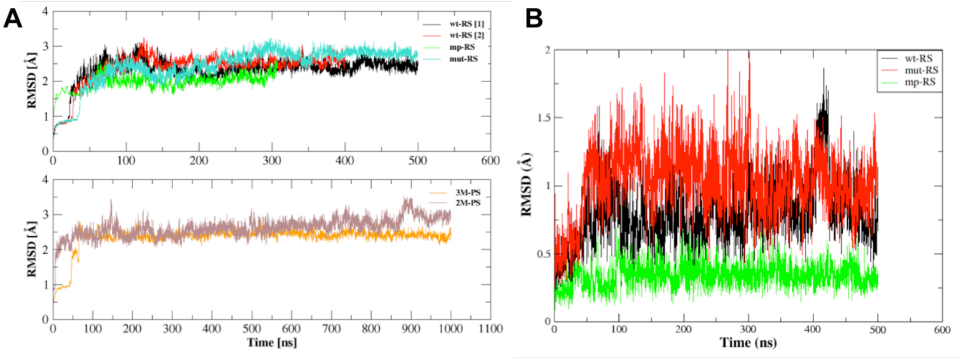
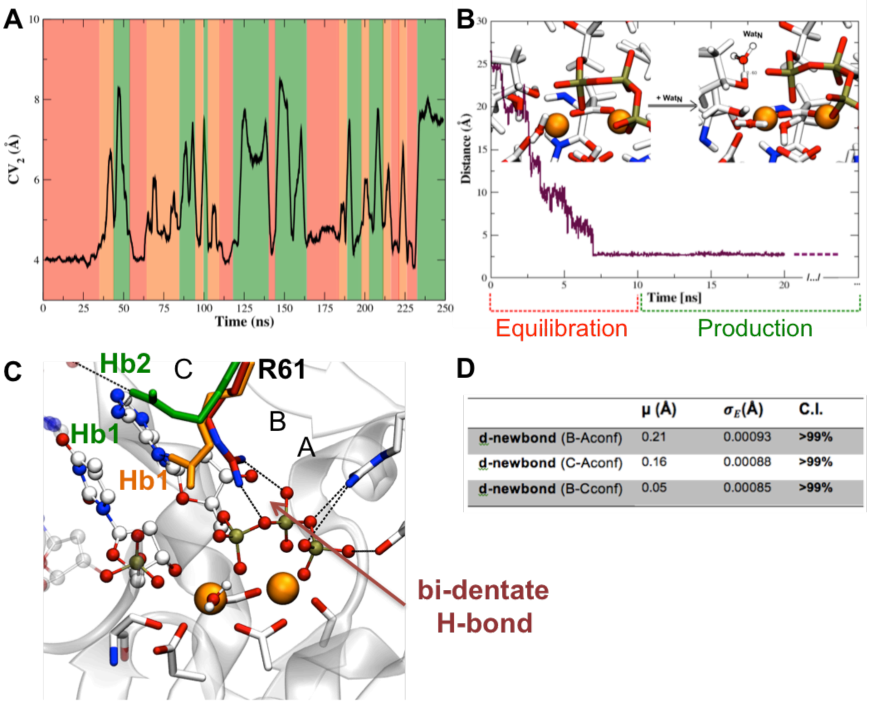


Figure S3.3. Root Mean Square Deviation (RMSD) of Pol-η and incoming nucleotide, backbone heavy atoms. (A, upper), Prereactive state: wt-RS system is in black, a second replica of wt-RS system is in red, the mp-RS system is in green, and mut-RS system (Arg61Ala mutant) is in cyan. The ternary Pol-η/dsDNA/dATP(dGTP) complex showed high stability in each prereactive state. The RMSD value for protein backbone is: 2.53 ± 0.27 Å for wt-RS, 2.73 ± 0.25 Å for mut-RS and 2.12 ± 0.24 Å for mp-RS. **(A, bottom)**, Postreactive state: in orange is indicated the 3M-PS system while in brown the 2M-PS system. The ternary Pol-η/dsDNA/PPi complex showed high stability in both systems. The backbone RMSD of the whole ternary complex with respect to the X-ray structure is 2.72 ± 0.33 Å and 2.53 ± 0.52 Å for the 3M-PS and 2M-PS, respectively. The DED-motif, which coordinated both MgA and MgB in the catalytic pocket, also exhibited high stability with an RMSD value of only 0.27 ± 0.07 Å, preserving the octahedral coordination for MgA and MgB, as well as MgC in 3M-PS, which conserved the polyhedral coordination geometry well. **(B)** Root Mean Square Deviation (RMSD) of the dATP heavy atoms in prereactive states. Black color indicates the wt-RS system; red indicates the mut-RS system, while green indicates dGTP RMSD value in the mp-RS system.



	μ (Å)	σ_E(Å)	C.I.
d-newbond (B-Aconf)	0.21	0.00093	**>99%**
d-newbond (C-Aconf)	0.16	0.00088	**>99%**
d-newbond (B-Cconf)	0.05	0.00085	**>99%**

Figure S3.4 Arg61 swinging in wt-RS. (A) Black line shows the different values of CV_2 along the simulated time scale. Red background marks CV_2 values corresponding to **A-conf**, orange to **B-conf** while green indicates **C-conf**. Here, the A↔B interconversion occurred seven times (t = ~30, ~65, ~110, ~180, ~210, ~220, ~230 ns) while nine A↔B interconversions were measured for the first replica system (at t = ~60, ~90, ~100, ~140, ~150, ~160, ~170, ~180, ~240 ns, see Fig. 2A in the manuscript). A↔C conformational switch, where **C-conf** was just transiently formed, happened six times in this second replica (t = ~55, ~115, ~140, ~145, ~165, ~230 ns) and four times during the simulated timescale of the first replica (t= ~190, ~210, ~215, ~225 ns, see Fig. 2A in the manuscript). **(B)** Notably, during the first few ns of simulation time, one water molecule (Wat$_N$) reached the catalytic site from the bulk, H-bonding the nucleophile 3'-OH of the incoming nucleotide. Once formed, the Wat$_N$-3'-OH H-bond exhibited great stability, maintaining an average length of 2.10 ± 0.08 Å throughout the collected trajectories. In this position, Wat$_N$ is properly placed to act as a general base for the activation of the nucleophilic 3'-OH of the primer end and to initiate the catalytic phosphoryl-transfer reaction, as recently proposed by Nakamura et al. (ref.1 in manuscript) **(C)** Different Arg61 conformations detected in our MD simulations of wt-RS system**.** Dashed lines indicate H-bond interactions. Hydrogen atoms were removed for clarity. Arg61 is shown assuming A-conf (red), B-conf (orange), and C-conf (green). Typical H-bond are reported for each Arg61 conformation. **(D)** Confidence interval data of *d-newbond* in wt-RS, considering the 3 conformations (A, B and C, see text in SI for details on confidence interval calculations).

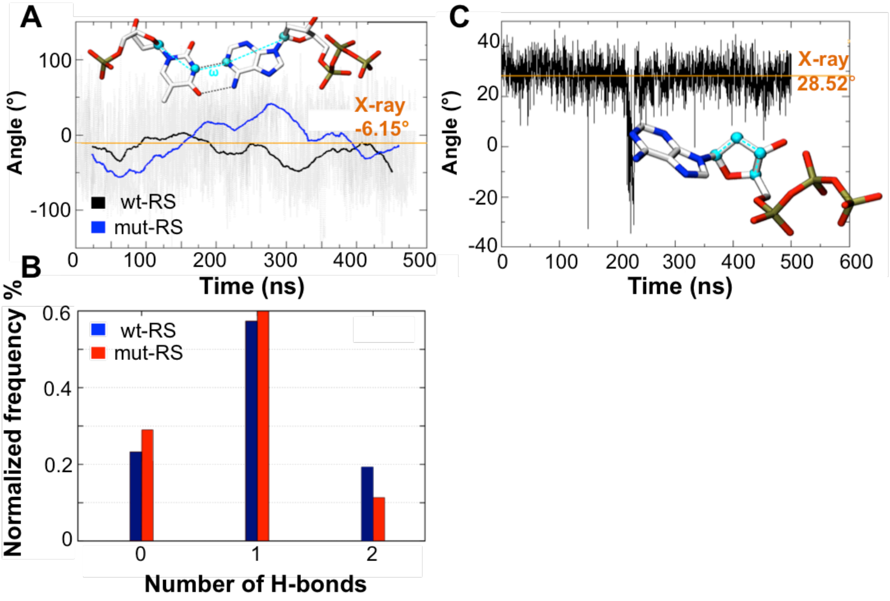


Figure S3.5. Further characterization of wt-RS vs. mut-RS. (A) value of the ω pseudo dihedral measured along the MD simulations of wt-RS and mut-RS systems. Angle's value quantifies the buckling of the terminal base-pair (dATP:T). Atoms forming ω pseudo dihedral (C1-N1-N3-C1) are represented in CPK and colored in cyan. Black line indicates fluctuations of the ω in wt-RS system while blue line describes those in mut-RS system. Orange line describes ω value detected in X-ray structure (PDBid 4ECS). As showed in the graph, in mut-RS system the terminal base pair get buckled, with a ω rotation of ~28.00° (with respect to the -6.15° found in the X-ray structure). Moreover, in mut-RS, the angles N-H-O and N-H-N and the distance N-O, N-N (which are characteristic of Watson and Crick's structure) increase to an average value of ~40.00° for N-H-O and N-H-N angles, and ~4.20 Å for N-O and N-N distances. This is in comparison to the X-ray, where they are ~20.00° and ~3.00 Å, respectively. **(B)** Frequency distribution of the Watson and Crick H-bond of the terminal base pair (dATP:T) in prereactive states. H-bond network between the incoming nucleotide and its templating base in wt-RS (Blue) and mut-RS (Red) systems. The graph shows a pronounced instability of the W&C interaction framework in mut-RS system. In fact, only the ~11% of the collected frames report the presence of two H-bond interactions between dATP:T. (**C**) Sugar pucker conformation adopted by dATP ribose in wt-RS. Atoms forming the analyzed pseudo dihedral are colored in cyan. Orange line shows the value of the angle detected in the crystal structure (PBDid 4ECS).

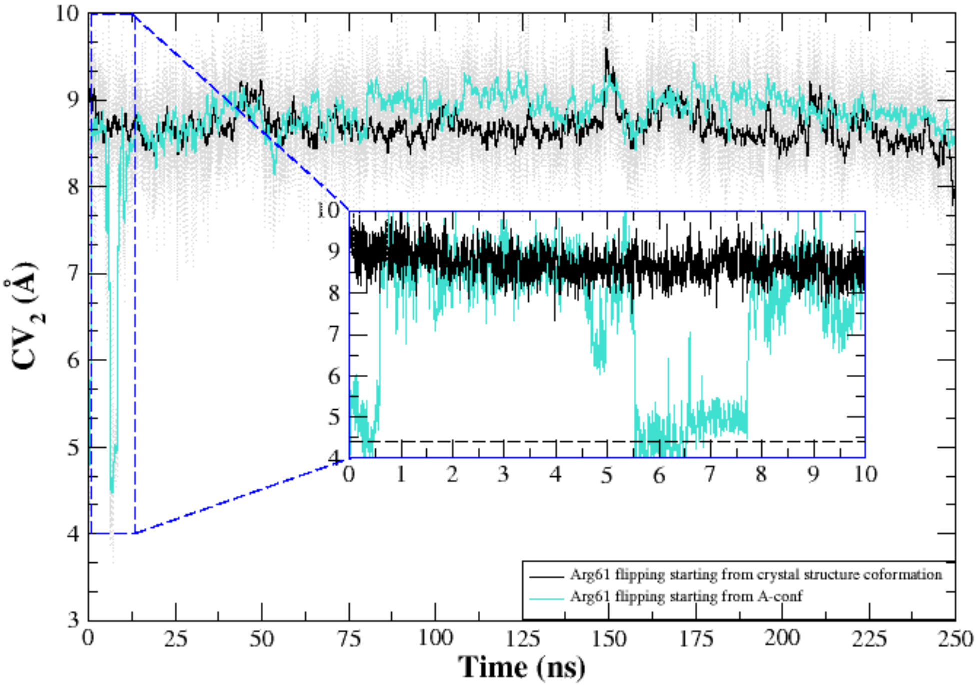
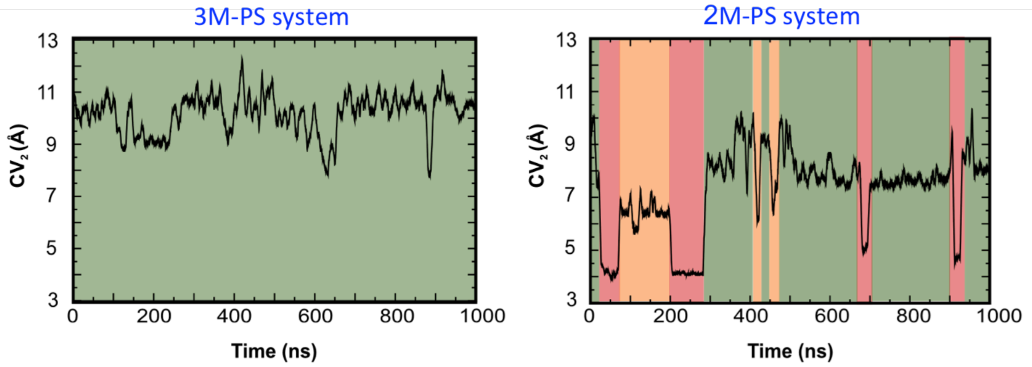


Figure S3.6. Arg61 swinging in mp-RS system. Black indicates CV_2 values in the mp-RS MD simulations in which Arg61 assumes its crystallographic pose as a staring conformation (C-conf, forming Hb1, Hb2 and Hb3). Cyan line represents CV_2 values in mp-RS in which Arg61 adopts A-conf as a starting pose. The graph shows the fast conformational A↔C interconversion even when A-conf is the starting pose.

Postreactive state of DNA Pol-η



Arg61 swinging in 3M-PS vs 2M-PS

Figure S3.7. Arg61 swinging in 3M-PS vs. 2M-PS. 3M-PS: Pol-η in complex with dsDNA, PPi and three Mg^{2+} ions. **Right**, 2M-PS: Pol-η in complex with dsDNA, PPi and two Mg^{2+} ions. Black line indicates value of CV$_2$ (see manuscript) along the simulated timescale. Red, orange and green represent those CV$_2$ values which defining Arg61 **A- B-** or **C-conf**, respectively.

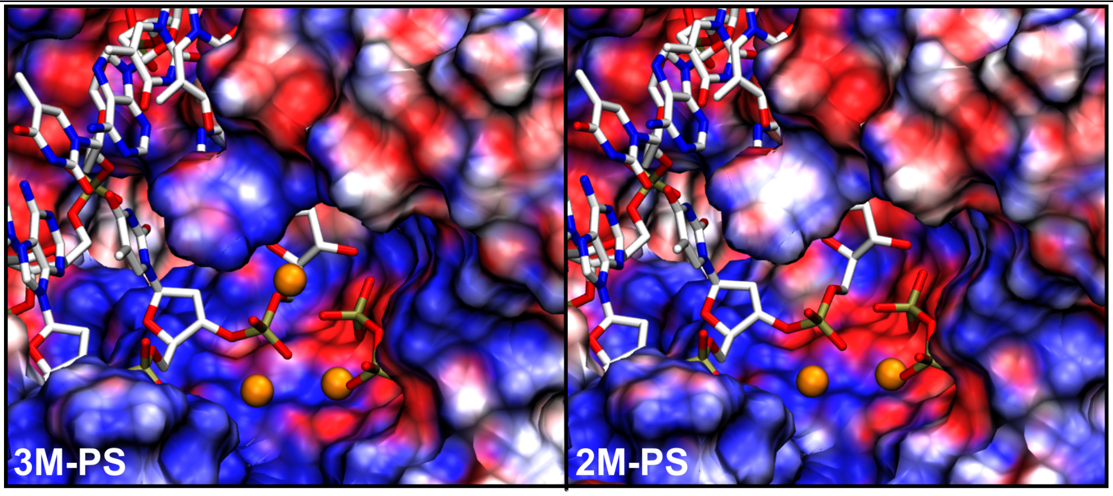


Figure S3.8. Electrostatic surface of Pol-η active site in postreactive state. Left, 3M-PS: Pol-η in complex with dsDNA, PPi and three Mg^{2+} ions. **Right**, 2M-PS: Pol-η in complex with dsDNA, PPi and two Mg^{2+} ions. The lack of the third metal, in 2M-PS, determines the formation of a more negatively charged environment which serves as driving force to recruit the third metal from the bulk solution.

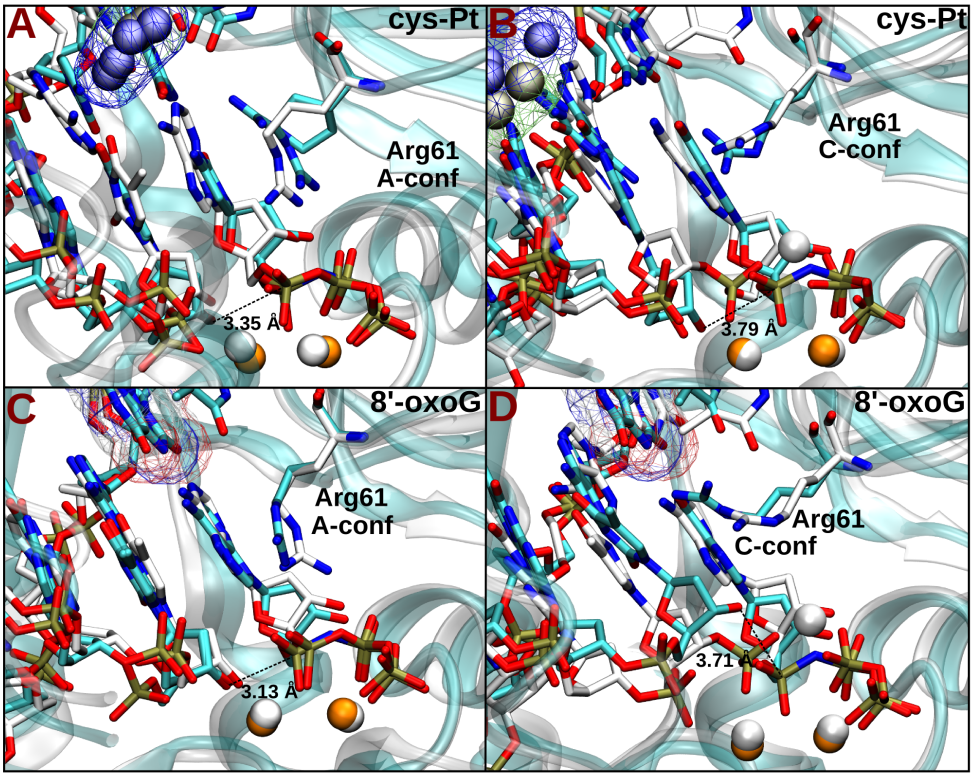


Figure S3.9. Structural superimpositions of Pol-η in complex with damaged or undamaged dsDNA. White carbons and metals indicate snapshot coming from MD simulations while cyan identify crystallographic structures; *d-newbond* is indicated by a black dashed line. (**A**) X-ray structure (PDBid 4DL4) of a prereactive configuration of Pol-η enzyme in complex with DNA containing a cis-Pt adduct (van der Waals representation). Crystallographic **A-conf** (cyan) well match **A-conf** described by our MD simulation (white). (**B**) X-ray structure (PDBid 4DL6) of a post-insertion configuration of Pol-η enzyme in complex with DNA containing a cis-Pt adduct (van der Waals representation). Crystallographic **C-conf** (cyan) well matches **A-conf** described by our MD simulation (white); here *d-newbond* adopts higher value with respect system reported in A. (**C**) X-ray structure (PDBid 4O3O) of a pre-reactive configuration of Pol-η enzyme in complex with DNA containing a 8-hydroxyguanine (8-oxoG) lesion (wireframe representation). Crystallographic **A-conf** (cyan) well matches **A-conf** described by our MD simulation (white). (**D**) X-ray structure (PDBid 4O3Q) of a pre-reactive configuration of Pol-η enzyme in complex with DNA containing a 8-hydroxyguanine (8-oxoG) lesion Crystallographic **C-conf** (cyan) well match **C-conf** described by our MD simulation (white).

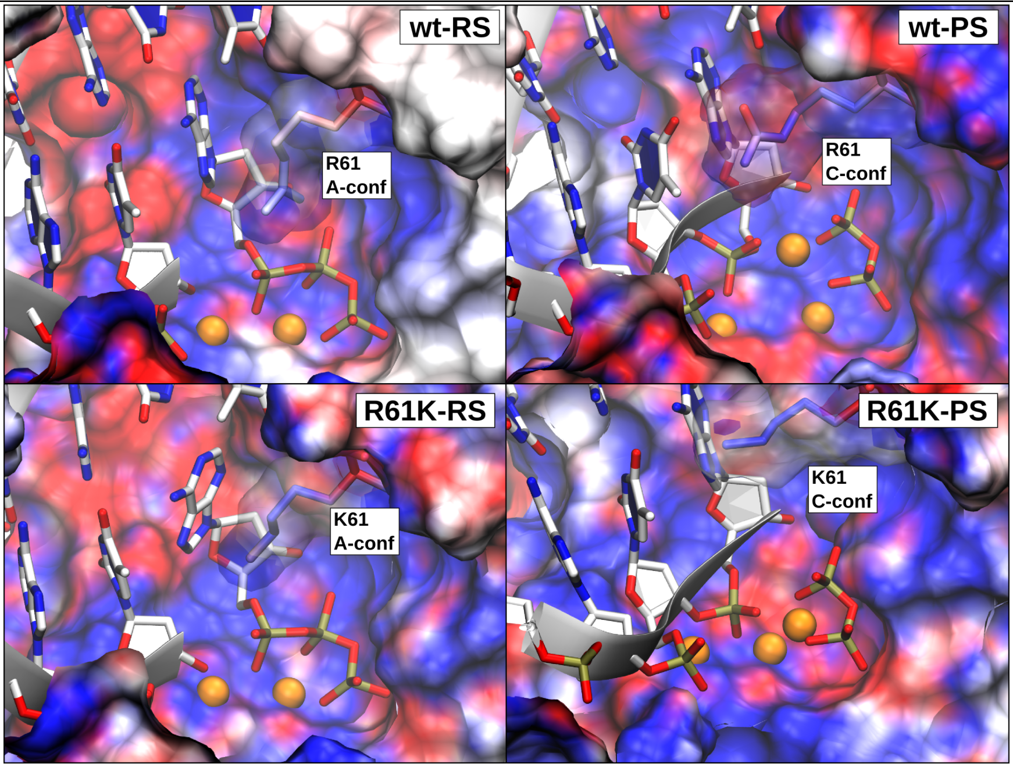
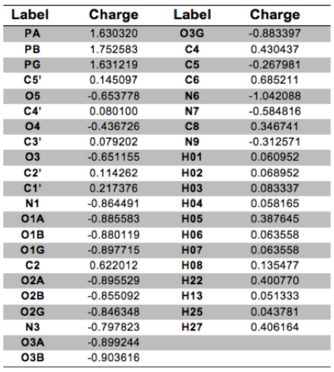
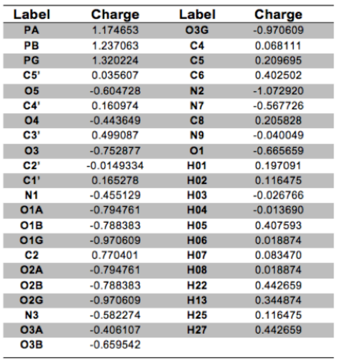


Figure S3.10. Electrostatic distribution of Pol-η active site in wild type and R61K mutant system. wt-RS, wild type pre-reactive state with Arg61 adopting **A-conf**. **wt-PS**, wild type post-reactive state with Arg61 adopting **C-conf**. **R61K-RS**, pre-reactive state of a system in which Arg61 was mutated to lysine. **R61K-PS**, active site electrostatic distribution of a post-reactive system affected by the mutation R61K. Nucleic acid is drawn in licorice and ribbon style. Mg^{2+} ions are depicted as orange spheres. Incoming nucleotide and PPi are represented in licorice. The lack of the third metal, in 2M-PS, determines the formation of a more negatively charged environment which serves as driving force to recruit the third metal from the bulk solution, as also reported in Fig. S8. Hydrogen atoms were removed for clarity.

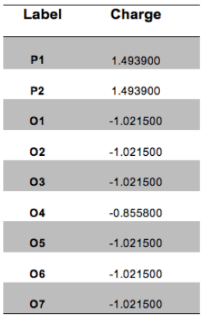
System/Distance	MgA-MgB	MgA-lig	MgB-lig



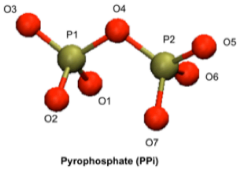
Label	Charge	Label	Charge
PA	1.630320	O3G	-0.883397
PB	1.752583	C4	0.430437
PG	1.631219	C5	-0.267981
C5'	0.145097	C6	0.685211
O5	-0.653778	N6	-1.042088
C4'	0.080100	N7	-0.584816
O4	-0.436726	C8	0.346741
C3'	0.079202	N9	-0.312571
O3	-0.651155	H01	0.060952
C2'	0.114262	H02	0.068952
C1'	0.217376	H03	0.083337
N1	-0.864491	H04	0.058165
O1A	-0.885583	H05	0.387645
O1B	-0.880119	H06	0.063558
O1G	-0.897715	H07	0.063558
C2	0.622012	H08	0.135477
O2A	-0.895529	H22	0.400770
O2B	-0.855092	H13	0.051333
O2G	-0.846348	H25	0.043781
N3	-0.797823	H27	0.406164
O3A	-0.899244		
O3B	-0.903616		

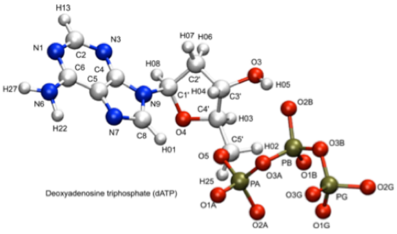


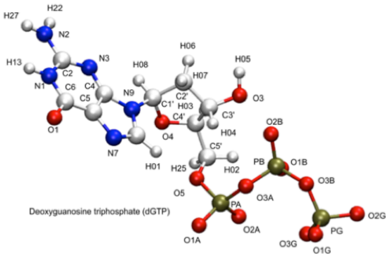
Label	Charge	Label	Charge
PA	1.174653	O3G	-0.970609
PB	1.237063	C4	0.068111
PG	1.320224	C5	0.209695
C5'	0.035607	C6	0.402502
O5	-0.604728	N2	-1.072920
C4'	0.160974	N7	-0.567726
O4	-0.443649	C8	0.205828
C3'	0.499087	N9	-0.040049
O3	-0.752877	O1	-0.665659
C2'	-0.0149334	H01	0.197091
C1'	0.165278	H02	0.116475
N1	-0.455129	H03	-0.026766
O1A	-0.794761	H04	-0.013690
O1B	-0.788383	H05	0.407593
O1G	-0.970609	H06	0.018874
C2	0.770401	H07	0.083470
O2A	-0.794761	H08	0.018874
O2B	-0.788383	H22	0.442659
O2G	-0.970609	H13	0.344874
N3	-0.582274	H25	0.116475
O3A	-0.406107	H27	0.442659
O3B	-0.659542		

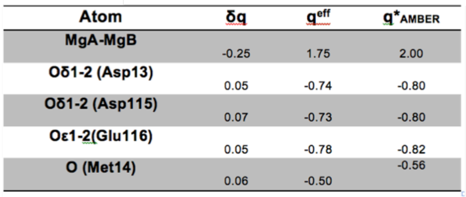


Label	Charge
P1	1.493900
P2	1.493900
O1	-1.021500
O2	-1.021500
O3	-1.021500
O4	-0.855800
O5	-1.021500
O6	-1.021500
O7	-1.021500









Atom	δq	q^{eff}	q^*_{AMBER}
MgA-MgB	-0.25	1.75	2.00
Oδ1-2 (Asp13)	0.05	-0.74	-0.80
Oδ1-2 (Asp115)	0.07	-0.73	-0.80
Oε1-2(Glu116)	0.05	-0.78	-0.82
O (Met14)	0.06	-0.50	-0.56

Table S3.1. Calculated RESP charges. Upper left, dATP. **Upper center**, dGTP. **Upper right**, PPi leaving group. The QM electrostatic potential (ESP), required during the RESP fitting procedure, was calculated at the HF/6-31G* level of theory by using Gaussian 09. **Bottom**, Corrected charges for metal ions and ligands with respect to AMBER RESP charges.

wt-RS	3.57 ± 0.09 Å	1.93 ± 0.05 Å	1.90 ± 0.05 Å
mut-RS	3.59 ± 0.09 Å	1.93 ± 0.05 Å	1.89 ± 0.04 Å
mp-RS	3.57 ± 0.08 Å	1.91 ± 0.05 Å	1.92 ± 0.07 Å
2M-PS	3.68 ± 0.17 Å	1.90 ± 0.05 Å	1.90 ± 0.05 Å
3M-PS	3.65 ± 0.11 Å	1.89 ± 0.04 Å	1.94 ± 0.05 Å

System	bi-dentate	Hb1	Hb1	Hb2	Hb3
	A-conf	B-Conf	C-Conf	C-Conf	C-Conf
wt-RS	2.48 ± 0.16 Å	2.51 ± 0.24 Å	2.23 ± 0.19 Å	3.44 ± 0.24 Å	
mp-RS			2.62 ± 0.19 Å	2.58 ± 0.22 Å	2.68 ± 0.17 Å
3M-PS			2.64 ± 0.30 Å	2.48 ± 0.22 Å	
2M-PS	2.52 ± 0.18 Å	2.57 ± 0.21 Å	2.64 ± 0.30 Å	2.48 ± 0.22 Å	

Table S3.2. Key distances considered for each of the simulated system. Average value was calculated for MgA-lig and MgB-lig. In **wt-RS** the average distance between MgA and the carboxylate groups was 1.93 ± 0.05 Å. The MgA-$O_{3'\text{-}OH}$ distance was 2.12 ± 0.05 Å, whereas the MgA-$O_{\alpha\text{-phosphate}}$ and MgA-O_{H2O} distances were 2.11 ± 0.04 Å and 2.02 ± 0.05 Å, respectively. MgB coordinated the carboxylate groups of Asp13 and Asp115, the backbone oxygen atom of Met14, and the non-bridging oxygen atoms of the α-, β-, and γ-phosphate groups of the incoming nucleotide. The average distance between MgB and the carboxylate groups (Asp13 and Asp115) was 1.91 ± 0.05 Å, while that to the oxygen atoms of the phosphate groups (α- β- and γ-phosphate groups) was 2.10 ± 0.05 Å. In **mut-RS** both metals maintained their native coordination as observed in wt-RS systems. Precisely, the averaged MgA-carboxylates length was 1.93 ± 0.05 Å. The MgA-$O_{3'\text{-}OH}$ distance was 2.14 ± 0.05 Å, whereas the MgA-$O_{\alpha\text{-phosphate}}$ and MgA-O_{H2O} distances were 2.10 ± 0.05 Å and 2.00 ± 0.05 Å, respectively. In **mp-RS** the metal geometry coordination is stable maintained as in the wt-RS system. Indeed, metal-carboxylates values are very similar to those reported for wt-RS system. Both postreactive state systems share the same value for metal-ligands interaction lengths. Indeed, for both of **3M-PS** and **2M-PS**, the averaged MgA-carboxylates length is 1.89 ± 0.04 Å while that of MgB-carboxylates is 1.92 ± 0.05 Å. Instead, the inter-metal length is slightly increased with respect to the prereactive state. Here, the averaged MgA-MgB distance is 3.67 ± 0.14 Å. The table below specifies the lengths of the H-bonds established by Arg61 via its different conformation along the collected trajectories. Orange background indicates H-bonds lengths detected for A-conf, blue indicate H-bond lengths for B-conf while tan indicate H-bond lengths corresponding to C-conf.

	wt-RS	mut-RS	mp-RS	2M-PS
	Kcal mol^{-1}	*Kcal mol^{-1}*	*Kcal mol^{-1}*	*Kcal mol^{-1}*

Arg61 A-conf	-9.50 ± 1	---	-9.00 ± 1	-7.00 ± 1
Arg61 B-conf	-7.00 ± 1	---	-5.50 ± 1	-4.00 ± 1
Arg61 C-conf	-6.00 ± 1	---	-5.50 ± 1	-4.00 ± 1

Table S3.3. Energetics of R61 conformations. Table reporting the energy values computed for each Arg61 conformation in the simulated systems.

Appendix Chapter 4

A self-activated mechanism for nucleic acid polymerization catalyzed by DNA/RNA polymerases

Vito Genna[1,2,], Pietro Vidossich[2], Emiliano Ippoliti[2], Paolo Carloni[2*] and Marco De Vivo[1,2*]

[1] Laboratory of Molecular Modeling & Drug Discovery, Istituto Italiano di Tecnologia, Via Morego 30, 16163, Genoa, Italy

[2] IAS-5 / INM-9 Computational Biomedicine and JARA-HPC, Forschungszentrum Jülich, Wilhelm-Johnen-Straße 52428 Jülich, Germany

*Corresponding Authors

Prof. Paolo Carloni
E-mail: p.carloni@fz-juelich.de

Dr. Marco De Vivo
E-mail: marco.devivo@iit.it

Supporting Information

Table of contents

Supplementary Text

Supplementary Figures

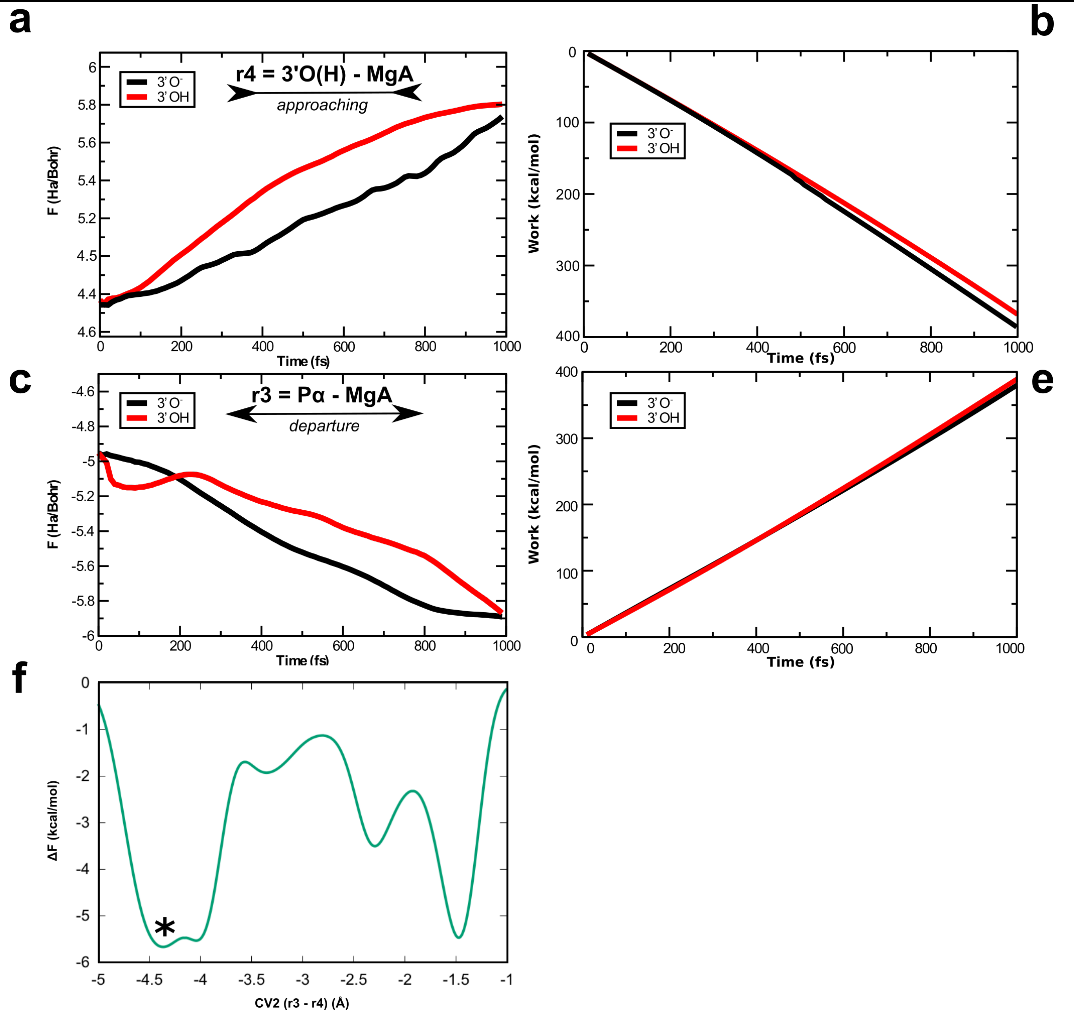
Supplementary tables

Car-Parrinello Molecular Dynamics (QM/MM MD). Car-Parrinello MD is an established *ab initio* method that does not rely on empirically determined potentials[196]. Instead, internuclear forces are determined on-the-fly from electronic structure calculations using the Kohn-Sham formulation of the Density Functional Theory (DFT)[197]. Such approach, despite being much more computationally demanding than classical and semiempirical methods, has been demonstrated to be robust and informative when used to study chemical reactivity, including systems based on metals and their complexes. Due to its high computational cost, a full QM treatment of the whole system (about ~70,000 atoms) is not tractable. Thus, we rely on a hybrid QM/MM potential[57] to investigate SAM for Pol-η catalysis. According to this approach, the QM description is reserved to the enzyme's active site (including the two catalytic metals and a relevant portion of the DNA substrate and incoming nucleotide), while the remaining part of the system is treated at classical level. In detail, the reactive region of the complex – 183 atoms in total, including MgA and MgB and the M17, R61, D113, D115, E116 residues, five water molecules and key atoms of the nucleic acid involved in the reaction – was treated at the DFT level using the BLYP functional[48,198]. This level of theory has been shown to describe well a variety of enzymatic reactions, with a fairly good (semi-qualitative) estimation of the energetics for catalysis[76,127,132,199].

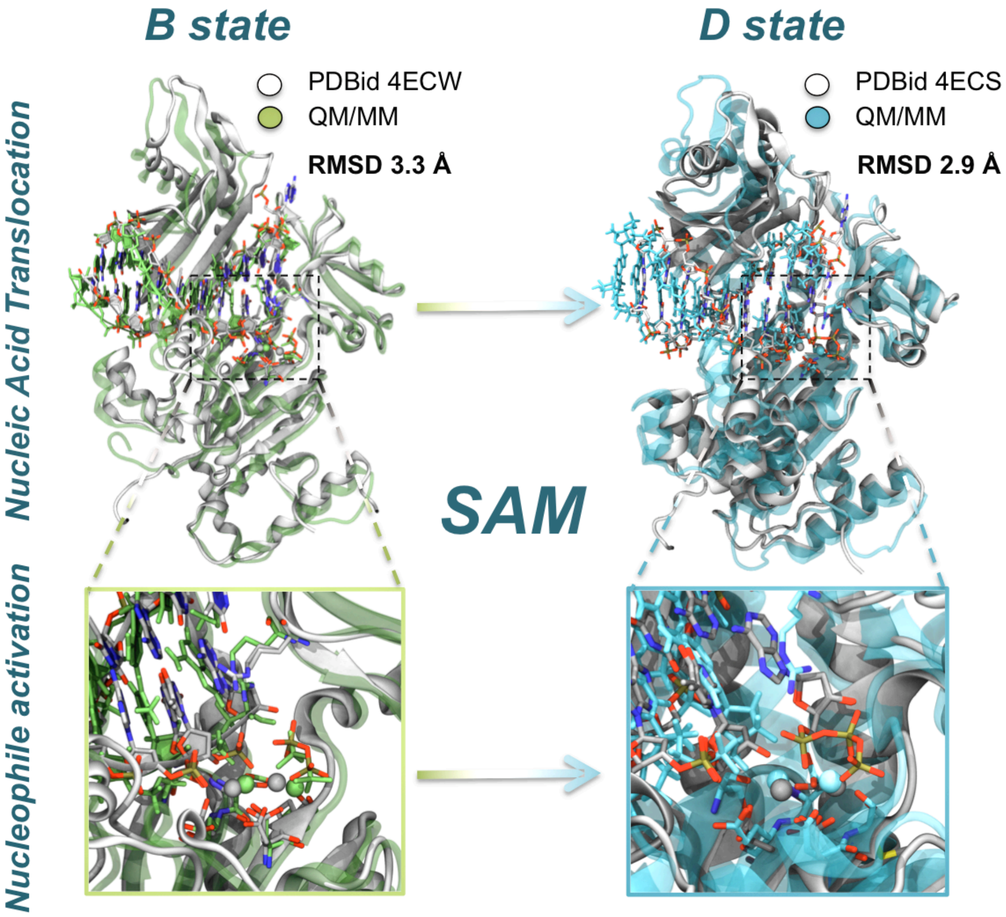
The remaining part of the system, including all the other water molecules and counterions, was treated at the AMBER parm99 force field level[200]. In more details, the wave functions are expanded in a plane wave basis set up to 70 Ry in a QM cell of proper dimensions, taking as reference our previous studies[76,201,202]. Only the valence electrons are treated explicitly (in the case of Mg atoms, electrons in the n = 5 shell are also included in the valence), while the core electrons will be described using norm-conserving pseudopotentials of the Martins–Troullier type[203]. An adapted monovalent carbon pseudopotential is employed to saturate the dangling bonds in between the QM and MM regions. Isolated system conditions[204] is imposed in the QM part by employing the Martyna-Tuckerman scheme[205]. Notably, this approach has been shown to accurately describe a variety of enzymatic systems and protein/DNA complexes[199,206,207]. All simulations were performed with the code CPMD (www.cpmd.org) version 4.1.1, developed at IBM Research Division, Zürich and the Max-Planck Institute, Stuttgart.

This program features an efficient QM/MM interface developed by Prof. Röthlisberger and co-workers[57] based on the Gromos MD engine.[208]

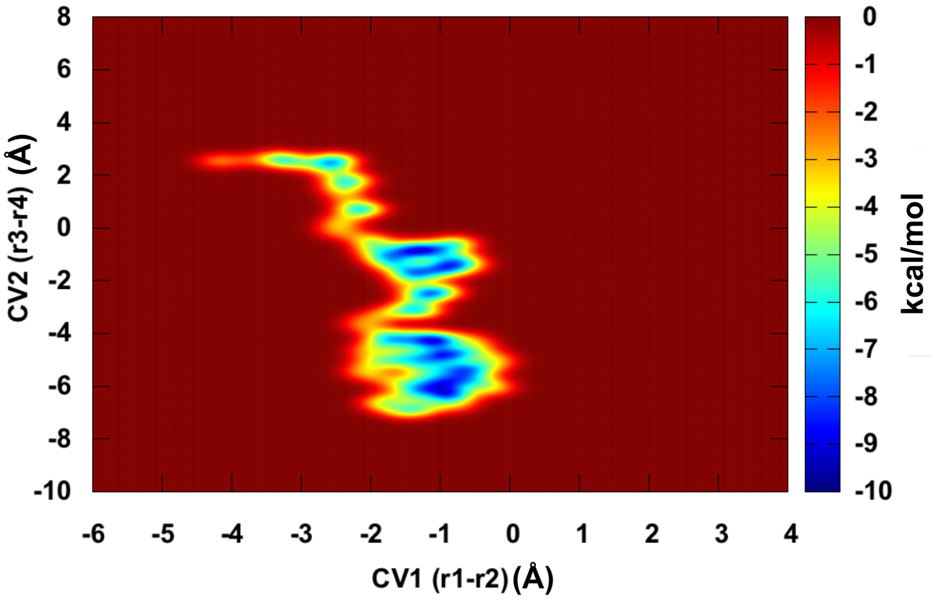
Role of Metal ions in SAM. The 2M-containing catalytic pocket actively assists this self-assisted and orchestrated mechanism. In fact, we observed fluctuations in the MgA-MgB distance and variation in the MgA coordination geometry along the different steps of the reaction. In **A**, metals are 3.35 Å apart with MgA showing a coordination vacancy in its apical position. This vacancy is then fulfilled by the shifting of the 3'O(H) group, with final formation of the octahedral coordination centered on MgA, in **B**. Finally, in **B**, the inter-metal distance adopts a value of ~4.00 Å, which agrees well with ternary complexes of postreactive states of different polymerases.[1,26,68]



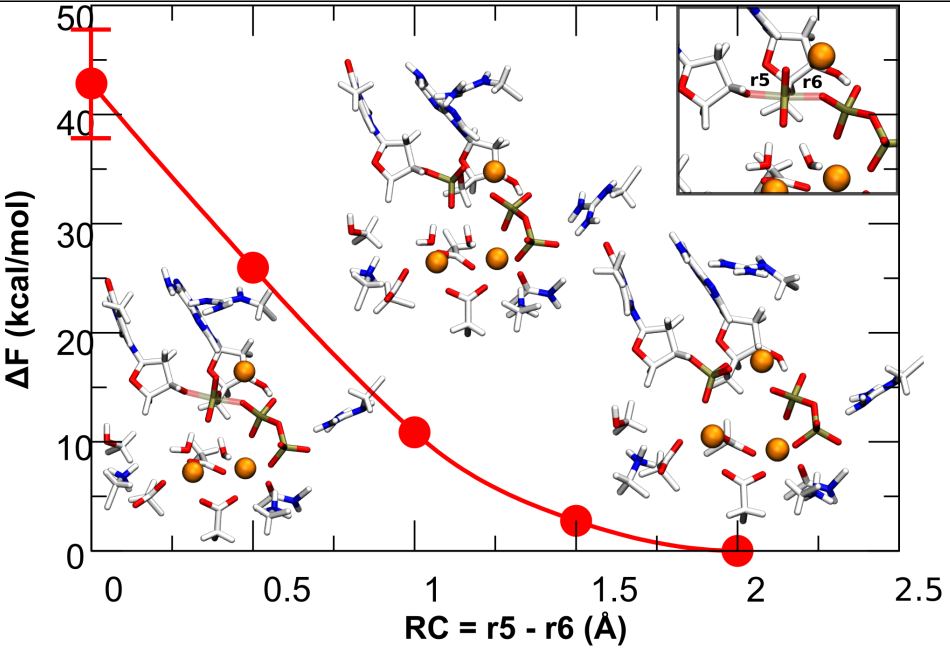
Supplementary Figure. 4.1 A and **C,** Force profiles derived from *ab initio* CP QM/MM steered MD simulations (SMD) upon r3 and r4 (CV2) components to determine DNA translocation. The plot shows the resulting mean values by averaging the force profiles from different SMD run for 3’OH and five for 3’O$^-$. **B** and **E,** external work calculated upon force profiles. **F,** Free energy surface for SAM in human DNA Pol-η considering CV2 (i.e. r3 – r4) only. The ensemble of global minimum refers to the starting point (see reaction scheme and point B in Fig. 5 in the manuscript). The decoupling between CV1 and CV2 make more stable the staring point suggesting the interdependency between the chemical step (CV1, i.e. proton-transfer) and the physical step (CV2, i.e. nucleic acid translocation). The free energy landscape was investigated by employing the same computational protocol applied for SAM with no transient metal ion bound to the active site (hills height and width were set to 0.05 kcal mol^{-1} and 0.01 Å, respectively). A total of 250 gaussian hills were deposited.



Supplementary Figure 4.2. Before and after SAM. Superimposition between X-ray structures (in white) and snapshots extracted from our QM/MM metadynamics (see Fig. 6 in Chapter 4). **B-state** represents the starting model system, in green, superimposed with the X-ray structure (PDBid 4ECW[8], in white) of Pol-η's ternary complex in the post-reactive state. Bottom: close view of the ternary complex showing protonated 3'OH. **D-state** represents the superimposition between a snapshot extracted from our QM/MM simulations and the X-ray structure (PDBid 4ECS[8] in cyan; see also Fig. 4.6 in Chapter 4). Here, the DNA strand is translocated of one base. Bottom: close view of the active site ready to accommodate a new incoming NTP.

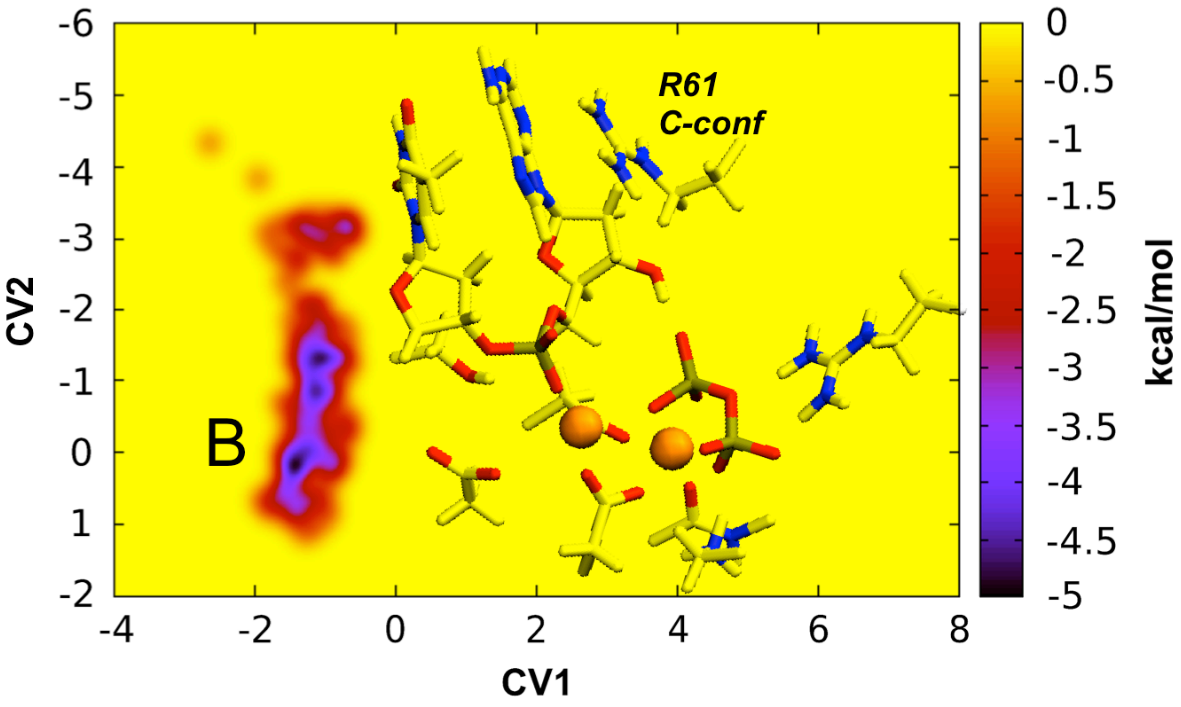


Supplementary Figure 4.3. Free energy surface for SAM in human DNA Pol-η with third Mg^{2+} ion bound to the pre-translocation complex. The ensemble of global minimum refers to the starting point (see reaction scheme and point B in Fig. 5 in Chapter 4). Neither proton-transfer nor DNA translocation was observed with MgC bound in this conformation further corroborating structural evidence that a third metal ion cannot be placed in the reactant enzyme-substrate complex, mainly because of steric clashes.[22] The free energy landscape was investigated by employing the same computational protocol applied for SAM with no transient metal ion bound to the active site (hills height and width were set to 0.05 kcal mol^{-1} and 0.01 Å, respectively). Also the same collective variables were used (CV1 and CV2, see Fig. 4.3 in Chapter 4) for the proton transfer under consideration, defined as the difference between the lengths of the 3'O—H (r_a) and the H—O_{Wat} bonds (r_b). A total of 250 gaussian hills were deposited.

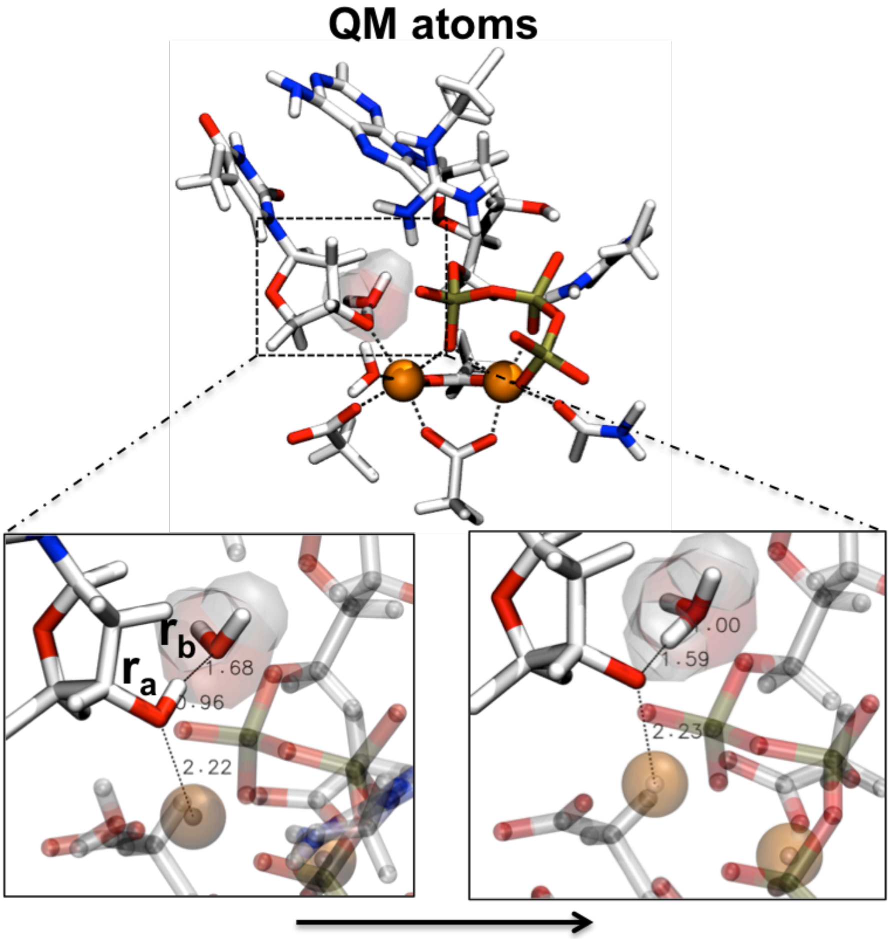


Supplementary Figure 4.4. Free energy profiles and structural properties of the back-reaction (i.e. pyrophosphorolysis) in in presence of transient MgC. Each reaction pathway is studied using Car-Parrinello (CP) QM/MM dynamics: the reactive region of the complex (~200 atoms) is treated at the quantum level (DFT-BLYP) and includes the same atoms reported in the methods section of the manuscript plus MgC. The remaining part of the system is treated using classical force field. The valence electrons are described by plane wave basis set up to a cutoff of 70 Ry. QM/MM CP dynamics are carried out with a time step of 0.12 fs (for a total of run time of ~50 ps)

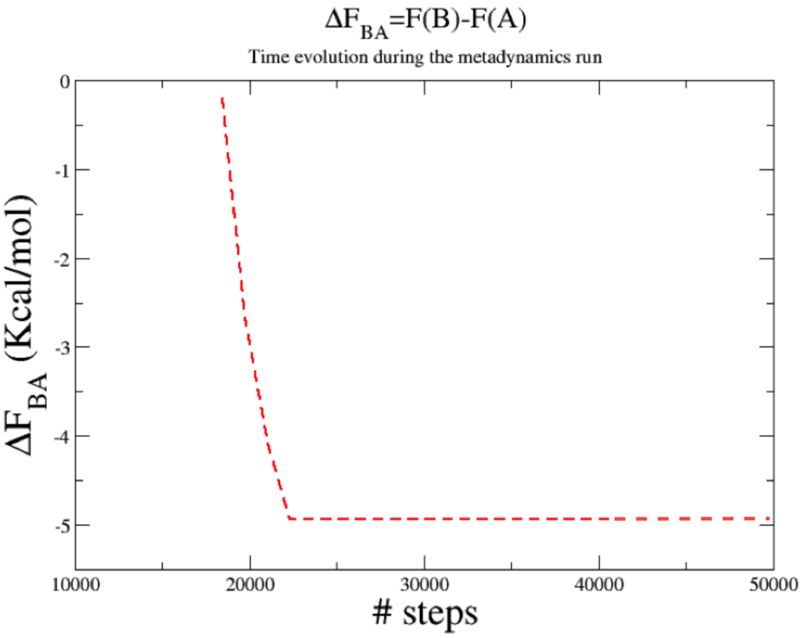
QM/MM protocol includes an initial equilibration of the configuration produced by MD simulations, followed by run where only the MM part is free to move, while the QM-part is kept frozen. Then, the whole system is allowed to move and heat up to 300K (~2 ps); after that trajectories are collected for analysis. Configurations from the equilibrated QM/MM simulations are used for free energy calculations. The pyrophosphorolysis reaction is described with a reaction coordinate (RC) defined as the difference between the length of the breaking bond (3'O—P^{α}, r5) and that of the forming bond (P^{α}—O_{PPi}). This RC is well suited for S_N2-like reactions. "Blue-Moon" ensemble simulations are carried out adiabatically constraining the RC, while leaving all other degrees of freedom free to evolve. The free energy surface (FES) of the reaction is obtained by thermodynamic integration. The pathway from reactants to products is divided in 4 steps, with a resolution of 0.5 Å. Each step is simulated for at least 5 ps, or until the force on the constraint is equilibrated (i.e., the running averages over 1 ps windows varies less than 5%). The MD-averaged force acting on the 1-D RC is plotted versus the length of the RC. The free energy profile is then obtained by integration of the force profile, leading to the description of the FES. The error associated to the critical points of FES is calculated by propagating the error on forces at every step, using the propagation of error formula for linear functions. The free energy values should be considered approximate due to the still limited sampling accessible to first principles DFT calculations, the choice of a 1-D RC, and the limitations of current GGA XC functionals. The catalytic pathways are then characterized in terms of the variation of critical bond lengths averaged over the equilibrated trajectory of each simulation step. The catalytic pathways are then characterized in terms of variation of critical bond lengths averaged over the equilibrated trajectory of each simulation step.



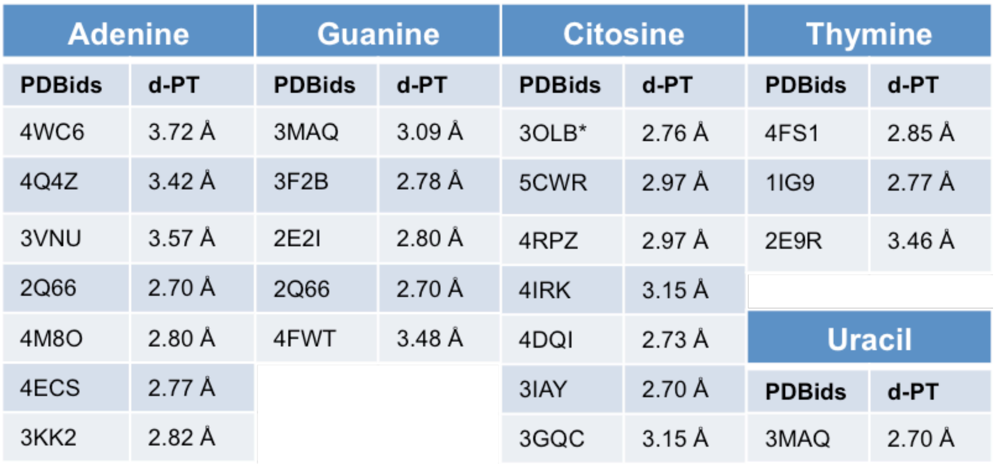
Supplementary Figure 4.5. Free energy surface for SAM in human DNA Pol-η with R61 adopting C-conf.[126] **B** identifies an ensemble of global minimum that refers to the starting point (see reaction scheme and point B in Fig. 3 in Chapter 4). Neither proton-transfer nor DNA translocation were observed with R61 in this conformation.



Supplementary Figure 4.6. We investigated nucleophile deprotonation in favor of a water molecule (Wat_N) conveniently located nearby the 3’OH group, as previously suggested by Nakamura et al.[8] Notably, in this mechanism, the 3’OH group is already properly placed (PDBid 4ECS)[8] to perform the in-line nucleophilic attack on the incoming nucleotide after DNA translocation, which therefore has no role here. Thus, these simulations were started from a model system representing the reactant state of Pol-η as in PDBid 4ECS. The reaction pathway was investigated via the same computational protocol applied for SAM (hills height and width were set to 0.05 kcal mol^{-1} and 0.01 Å, respectively) but using a single collective variable (CV_{Wat}) for the proton transfer under consideration, defined as the difference between the lengths of the 3’O—H (r_a) and the H—O_{Wat} bonds (r_b). A total of 250 gaussian hills (in ~50 000 steps) were deposited. In this case, the deepest minimum on the FES was the initial system conformation, where r_a = 0.96 Å and r_b = 1.68 Å. Afterward, the system evolved to reach the transition state for nucleophile activation, where r_a = r_b = 1.82 Å, which returned a free-energy barrier of ~7.2 kcal/mol. Subsequently, when CV_{Wat} = ~0.6 Å, the 3'-hydroxide and the hydronium ion were stably formed. As a result, nucleophile activation via Wat_N seems unfavored compared to SAM for Pol-η catalysis.



Supplementary Figure 4.7. Free energy difference between state **B** and state **A** (see Fig. 6 in the manuscript for the definition of the states) during the progress of the metadynamics run. Convergence is reached after ~22,000 steps. As you can see, ΔF_{BA}, which is the relevant quantity for understanding if the 3'OH is more acidic than a PPi group bound to Mg^{2+}, appears well converged, at least within the metadynamics runs performed in our study.



Adenine		Guanine		Citosine		Thymine	
PDBids	d-PT	PDBids	d-PT	PDBids	d-PT	PDBids	d-PT
4WC6	3.72 Å	3MAQ	3.09 Å	3OLB*	2.76 Å	4FS1	2.85 Å
4Q4Z	3.42 Å	3F2B	2.78 Å	5CWR	2.97 Å	1IG9	2.77 Å
3VNU	3.57 Å	2E2I	2.80 Å	4RPZ	2.97 Å	2E9R	3.46 Å
2Q66	2.70 Å	2Q66	2.70 Å	4IRK	3.15 Å		
4M8O	2.80 Å	4FWT	3.48 Å	4DQI	2.73 Å	Uracil	
4ECS	2.77 Å			3IAY	2.70 Å	PDBids	d-PT
3KK2	2.82 Å			3GQC	3.15 Å	3MAQ	2.70 Å

Supplementary Table 4.1. Value of the intramolecular H-bond (d-PT) measured in the nucleotide triphosphate of different ternary complexes of Polymerases, across each domain of life.

References

(1) Car, R.; Parrinello, M. *Phy. Rev. Lett.* 1985, *55*, 2471.

(2) Kohn, W.; Sham, L. J. *Phy. Rev. Lett.* 1965, *140*, A1133

(3) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Chem. Phys.* 2002, *116*, 6941.

(4) Becke, A. D. *Phys. Rev. A.* 1988, *38*, 3098.

(5) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B.* 1988, *37*, 785.

(6) De Vivo, M.; Dal Peraro, M.; Klein, M. L. *J. Am. Chem. Soc.* 2008, *130*, 10955.

(7) Dal Peraro, M.; Ruggerone, P.; Raugei, S.; Gervasio, F. L.; Carloni, P. *Curr. Opin. Struct. Biol.* 2007, *17*, 149.

(8) Iannuzzi, M.; Laio, A.; Parrinello, M. *Phys. Rev. Lett.* 2003, *90*, 238302.

(9) De Vivo, M.; Ensing, B.; Klein, M. L. *J. Am. Chem. Soc.* 2005, *127*, 11226.

(10) Wang, J.; Cieplak, P.; Kollman, P.A *J. Comput. Chem.* 2000, *21*, 1049.

(11) De Vivo, M.; *Front Biosci (Landmark Ed).* 2011, *16*, 1619.

(12) De Vivo, M.; Ensing, B.; Dal Peraro, M.; Gomez, G. A.; Christianson, D. W.; Klein, M. L. *J. Am. Chem. Soc.* 2007, *129*, 387.

(13) Troullier, N.; Martins, J. L. *Phy. Rev. B.* 1991, *43*, 1993.

(14) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *J. Chem. Phys.* 2005, *122*, 14113.

(15) Martyna, G.J.; Tuckerman, M.E. *J. Chem. Phys.* 1999, *110*, 2810.

(16) De Vivo, M.; Cavalli, A.; Carloni, P.; Recanatini, M. *Chemistry.* 2007, *13*, 8437.

(17) Cavalli, A.; De Vivo, M.; Recanatini, M.; *Chem. Comm.* 2003, 1308.

(18) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. Vdf Hochschulverlag AG an der ETH Zürich, Zürich, Switzerland, 1996, 1.

(19) Vaisman, A.; Ling, H.; Woodgate, R.; Yang, W. *Embo J.* 2005, *24*, 2957.

(20) Vyas, R.; Reed, A. J.; Tokarsky, E. J.; Suo, Z. C. *J. Am. Chem. Soc.* 2015, *137*, 5225.

(21) Palermo, G.; Cavalli, A.; Klein, M. L.; Alfonso-Prieto, M.; Dal Peraro, M.; De Vivo, M. *Acc. Chem. Res.* 2015, *48*, 220.

(22) Nakamura, T.; Zhao, Y.; Yamagata, Y.; Hua, Y. J.; Yang, W. *Nature.* 2012, *487*, 196.

(23) Gao, Y.; Yang, W. *Science.* 2016, *352*, 1334.

(24) Genna, V.; Gaspari, R.; Dal Peraro, M.; De Vivo, M.; *Nucleic Acids Res.* 2016, *44*, 2827.

Appendix Chapter 5

Newly-identified catalytic elements expand the two-metal-ion architecture of DNA and RNA processing enzymes

Vito Genna[1,#], Matteo Colombo[2,#], Marco De Vivo[1,3,*], Marco Marcia[2,*]

[*]To whom correspondence should be addressed.

E-mail: marco.devivo@iit.it;
mmarcia@embl.fr

The supplemental information file contains:

- Online methods
- Supplementary References
- Supplemental Figures 1 to 3
- Supplemental Table 1
- Supplemental File 1

Supporting Information

Online methods

Sequence and structural alignments.

Representative sequences of each enzyme have been selected in BLAST[166] and multiple-sequence alignments have been performed in ClustalOmega[167]. Structure-based sequence alignments of RNA-dependent RNA polymerases have been performed using T-COFFEE[168], as described previously[209].

Structural alignments have been performed manually in Coot[169] using the di-nuclear metal center M_A-M_B, the substrates, and/or the coordinating acidic residues in the first resolution shell of M_A-M_B as a guide. The figures depicting the structures were drawn using PyMOL[170].

Calculations of electrostatic potential maps.

Calculations of the electrostatic molecular surfaces were performed by solving a nonlinear Poisson-Boltzman equation with APBS[171] as described previously[210]. For all the systems we used a multilevel grid approach with a fixed grid length of 129 points in each spatial direction obtaining a grid spacing of $\leq$ 0.365 Å. To generate the electrostatic maps, all biomolecules were treated with a low dielectric medium ($\varepsilon = 2$)[172] and the surrounding solvent as a high dielectric continuum ($\varepsilon = 78.54$). The ionic strength was set to 150 mM corresponding to a Debye length of 8 Å, while temperature was set to 300 K. AMBER package[82] was used to add hydrogen atoms to crystallographic models and the AMBER ff99SB-ILDN force field[211] was used to derive atomic radii and charges. In order to diminish the sensitivity of computations to the grid setup, we have used cubic B-spline discretization method to map the charges of our systems into the grid points and the nine-point harmonic averaging approach to the surface-based dielectric and ion-accessibility coefficients. Finally, we applied Dirichlet boundary conditions by using multiple Debye–Huckel functionality.

Supplementary references

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).
2. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539 (2011).
3. Notredame, C., Higgins, D.G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-17 (2000).
4. Marcia, M., Ermler, U., Peng, G. & Michel, H. A new structure-based classification of sulfide:quinone oxidoreductases. *Proteins* **78**, 1073-83 (2010).
5. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**, 2126-32 (2004).
6. Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.3r1. (2010).
7. Baker, N.A., Sept, D., Joseph, S., Holst, M.J. & McCammon, J.A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* **98**, 10037-41 (2001).
8. Kazantsev, A.V., Krivenko, A.A. & Pace, N.R. Mapping metal-binding sites in the catalytic domain of bacterial RNase P RNA. *Rna* **15**, 266-76 (2009).

9. Misra, V.K. & Draper, D.E. A thermodynamic framework for Mg2+ binding to RNA. *Proc Natl Acad Sci U S A* **98**, 12456-61 (2001).

10. Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A. & Case, D.A. Development and testing of a general amber force field. *J Comput Chem* **25**, 1157-74 (2004).

11. Lindorff-Larsen, K. et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950-8 (2010).

12. Galej, W.P. et al. Cryo-EM structure of the spliceosome immediately after branching. *Nature*, doi: 10.1038/nature19316 (2016).

13. Newman, M. et al. Crystal structure of restriction endonuclease BglI bound to its interrupted DNA recognition sequence. *EMBO J* **17**, 5466-76 (1998).

14. Flores, H., Osuna, J., Heitman, J. & Soberon, X. Saturation mutagenesis of His114 of EcoRI reveals relaxed-specificity mutants. *Gene* **157**, 295-301 (1995).

15. Sen, S. & Nilsson, L. Structure, interaction, dynamics and solvent effects on the DNA-EcoRI complex in aqueous solution from molecular dynamics simulation. *Biophys J* **77**, 1782-800 (1999).

16. Horton, J.R. & Cheng, X. PvuII endonuclease contains two calcium ions in active sites. *J Mol Biol* **300**, 1049-56 (2000).

17. Nastri, H.G., Evans, P.D., Walker, I.H. & Riggs, P.D. Catalytic and DNA binding properties of PvuII restriction endonuclease mutants. *J Biol Chem* **272**, 25761-7 (1997).

18. Horton, N.C. & Perona, J.J. DNA cleavage by EcoRV endonuclease: two metal ions in three metal ion binding sites. *Biochemistry* **43**, 6841-57 (2004).

19. Horton, N.C., Connolly, B.A. & Perona, J.J. Inhibition of EcoRV endonuclease by deoxyribo-3 '-S-phosphorothiolates: A high-resolution X-ray crystallographic study. *Journal of the American Chemical Society* **122**, 3314-3324 (2000).

20. Horton, N.C. et al. Electrostatic contributions to site specific DNA cleavage by EcoRV endonuclease. *Biochemistry* **41**, 10754-63 (2002).

21. Wenz, C., Jeltsch, A. & Pingoud, A. Probing the indirect readout of the restriction enzyme EcoRV. Mutational analysis of contacts to the DNA backbone. *J Biol Chem* **271**, 5565-73 (1996).

22. Xu, Q.S., Kucera, R.B., Roberts, R.J. & Guo, H.C. An asymmetric complex of restriction endonuclease MspI on its palindromic DNA recognition site. *Structure* **12**, 1741-7 (2004).

23. Shen, B.W. et al. Characterization and crystal structure of the type IIG restriction endonuclease RM.BpuSI. *Nucleic Acids Res* **39**, 8223-36 (2011).

24. Mierzejewska, K. et al. Structural basis of the methylation specificity of R.DpnI. *Nucleic Acids Res* **42**, 8745-54 (2014).

25. Wah, D.A., Hirsch, J.A., Dorner, L.F., Schildkraut, I. & Aggarwal, A.K. Structure of the multimodular endonuclease FokI bound to DNA. *Nature* **388**, 97-100 (1997).

26. Bitinaite, J., Wah, D.A., Aggarwal, A.K. & Schildkraut, I. FokI dimerization is required for DNA cleavage. *Proc Natl Acad Sci U S A* **95**, 10570-5 (1998).

27. Huai, Q., Colandene, J.D., Topal, M.D. & Ke, H. Structure of NaeI-DNA complex reveals dual-mode DNA recognition and complete dimer rearrangement. *Nat Struct Biol* **8**, 665-9 (2001).
28. Deibert, M., Grazulis, S., Sasnauskas, G., Siksnys, V. & Huber, R. Structure of the tetrameric restriction endonuclease NgoMIV in complex with cleaved DNA. *Nat Struct Biol* **7**, 792-9 (2000).
29. Vanamee, E.S. et al. A view of consecutive binding events from structures of tetrameric endonuclease SfiI bound to DNA. *EMBO J* **24**, 4198-208 (2005).
30. Dunten, P.W. et al. The structure of SgrAI bound to DNA; recognition of an 8 base pair target. *Nucleic Acids Res* **36**, 5405-16 (2008).
31. Dias, A. et al. The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. *Nature* **458**, 914-8 (2009).
32. Crepin, T. et al. Mutational and metal binding analysis of the endonuclease domain of the influenza virus polymerase PA subunit. *J Virol* **84**, 9096-104 (2010).
33. Reich, S. et al. Structural insight into cap-snatching and RNA synthesis by influenza polymerase. *Nature* **516**, 361-6 (2014).
34. Lee, J.Y. et al. MutH complexed with hemi- and unmethylated DNAs: coupling base recognition and DNA cleavage. *Mol Cell* **20**, 155-66 (2005).
35. Junop, M.S., Yang, W., Funchain, P., Clendenin, W. & Miller, J.H. In vitro and in vivo studies of MutS, MutL and MutH mutants: correlation of mismatch repair and DNA recombination. *DNA Repair (Amst)* **2**, 387-405 (2003).
36. Orans, J. et al. Structures of human exonuclease 1 DNA complexes suggest a unified mechanism for nuclease family. *Cell* **145**, 212-23 (2011).
37. Korada, S.K.C. et al. Crystal structures of Escherichia coli exonuclease I in complex with single-stranded DNA provide insights into the mechanism of processive digestion. *Nucleic Acids Research* **41**, 5887-5897 (2013).
38. Miyazono, K. et al. Structural basis for substrate recognition and processive cleavage mechanisms of the trimeric exonuclease PhoExo I. *Nucleic Acids Res* **43**, 7122-36 (2015).
39. Rychlik, M.P. et al. Crystal structures of RNase H2 in complex with nucleic acid reveal the mechanism of RNA-DNA junction recognition and cleavage. *Mol Cell* **40**, 658-70 (2010).
40. Figiel, M. et al. The structural and biochemical characterization of human RNase H2 complex reveals the molecular basis for substrate recognition and Aicardi-Goutieres syndrome defects. *J Biol Chem* **286**, 10540-50 (2011).
41. Jinek, M. et al. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**, 1247997 (2014).
42. Hirano, H. et al. Structure and Engineering of Francisella novicida Cas9. *Cell* **164**, 950-61 (2016).
43. Nishimasu, H. et al. Crystal Structure of Staphylococcus aureus Cas9. *Cell* **162**, 1113-26 (2015).

44. Sawaya, M.R., Prasad, R., Wilson, S.H., Kraut, J. & Pelletier, H. Crystal structures of human DNA polymerase beta complexed with gapped and nicked DNA: evidence for an induced fit mechanism. *Biochemistry* **36**, 11205-15 (1997).

45. Kraynov, V.S., Showalter, A.K., Liu, J., Zhong, X. & Tsai, M.D. DNA polymerase beta: contributions of template-positioning and dNTP triphosphate-binding residues to catalysis and fidelity. *Biochemistry* **39**, 16008-15 (2000).

46. Xiang, Y., Oelschlaeger, P., Florian, J., Goodman, M.F. & Warshel, A. Simulating the effect of DNA polymerase mutations on transition-state energetics and fidelity: evaluating amino acid group contribution and allosteric coupling for ionized residues in human pol beta. *Biochemistry* **45**, 7036-48 (2006).

47. Nair, D.T., Johnson, R.E., Prakash, L., Prakash, S. & Aggarwal, A.K. Human DNA polymerase iota incorporates dCTP opposite template G via a G.C + Hoogsteen base pair. *Structure* **13**, 1569-77 (2005).

48. Uljon, S.N. et al. Crystal structure of the catalytic core of human DNA polymerase kappa. *Structure* **12**, 1395-404 (2004).

49. Silverstein, T.D. et al. Structural basis for the suppression of skin cancers by DNA polymerase eta. *Nature* **465**, 1039-43 (2010).

50. Johnson, R.E., Trincao, J., Aggarwal, A.K., Prakash, S. & Prakash, L. Deoxynucleotide triphosphate binding mode conserved in Y family DNA polymerases. *Mol Cell Biol* **23**, 3008-12 (2003).

51. Sharma, A., Kottur, J., Narayanan, N. & Nair, D.T. A strategically located serine residue is critical for the mutator activity of DNA polymerase IV from Escherichia coli. *Nucleic Acids Res* **41**, 5104-14 (2013).

52. Vaisman, A., Ling, H., Woodgate, R. & Yang, W. Fidelity of Dpo4: effect of metal ions, nucleotide selection and pyrophosphorolysis. *EMBO J* **24**, 2957-67 (2005).

53. Zhou, B.L., Pata, J.D. & Steitz, T.A. Crystal structure of a DinB lesion bypass DNA polymerase catalytic fragment reveals a classic polymerase catalytic domain. *Mol Cell* **8**, 427-37 (2001).

54. Tafur, L. et al. Molecular Structures of Transcribing RNA Polymerase I. *Mol Cell* **64**, 1135-1143 (2016).

55. Hoffmann, N.A. et al. Molecular structures of unbound and transcribing RNA polymerase III. *Nature* **528**, 231-6 (2015).

56. Zamyatkin, D.F. et al. Structural insights into mechanisms of catalysis and inhibition in Norwalk virus polymerase. *Journal of Biological Chemistry* **283**, 7705-7712 (2008).

57. Ng, K.K. et al. Crystal structures of active and inactive conformations of a caliciviral RNA-dependent RNA polymerase. *J Biol Chem* **277**, 1381-7 (2002).

58. Hansen, J.L., Long, A.M. & Schultz, S.C. Structure of the RNA-dependent RNA polymerase of poliovirus. *Structure* **5**, 1109-22 (1997).

59. Ferrer-Orta, C. et al. Structure of foot-and-mouth disease virus RNA-dependent RNA polymerase and its complex with a template-primer RNA. *J Biol Chem* **279**, 47212-21 (2004).

60. Ago, H. et al. Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Structure* **7**, 1417-26 (1999).

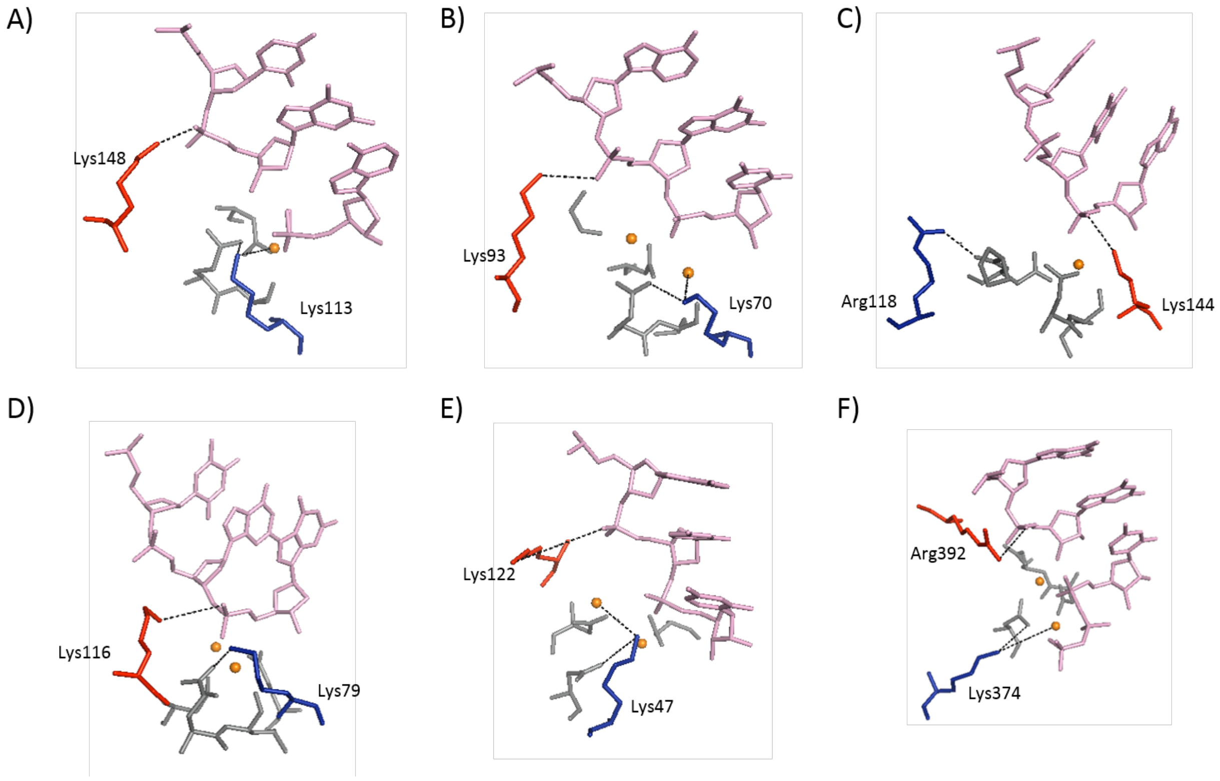
61. Yap, T.L. et al. Crystal structure of the dengue virus RNA-dependent RNA polymerase catalytic domain at 1.85-angstrom resolution. *J Virol* **81**, 4753-65 (2007).

62. Shen, B.W., Landthaler, M., Shub, D.A. & Stoddard, B.L. DNA binding and cleavage by the HNH homing endonuclease I-HmuI. *J Mol Biol* **342**, 43-56 (2004).

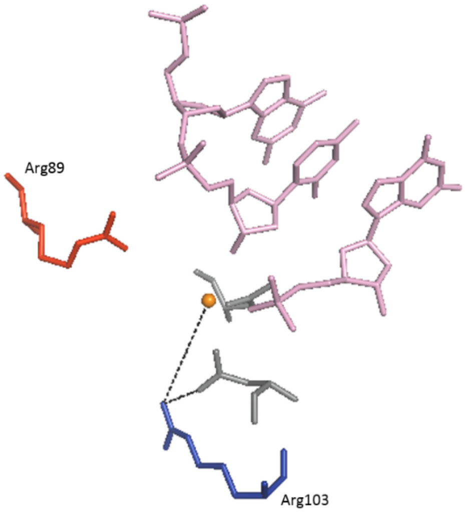
63. Jiang, F. et al. Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* **351**, 867-71 (2016).

Supplementary Figures and Tables

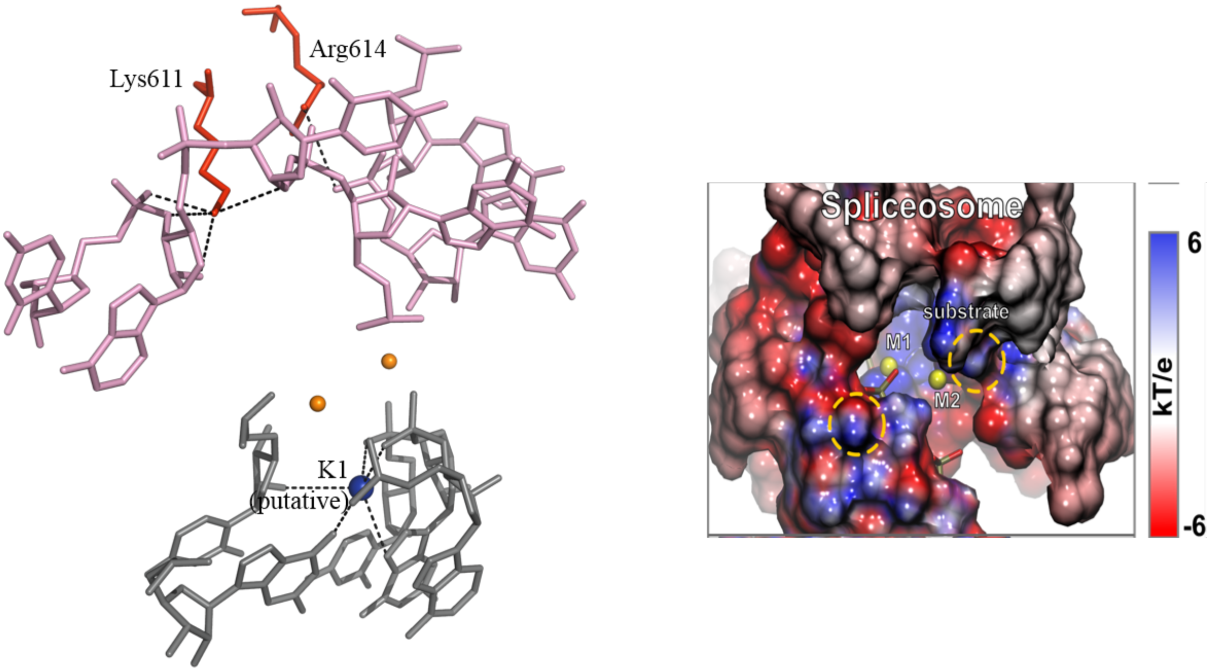
Supplementary Fig. 5.1: additional examples of nucleases and polymerases possessing K1- and K2-like residues. Active site representations of restriction endonucleases EcoRI (A, PDB id.: 1QPS), PvuII (B, 1F0O), BglI (C, 1DMU), MutH (D, 2AOR), RNase-H (E, PDB id.: 3O3H), and Norwalk virus RdRp (F, PDB id.: 3BSO). Structural elements are oriented and color coded as in Fig. 1.



Supplementary Fig. 5.2: K1- and K2-like residues in one-metal-ion enzymes. Active site representations of HNH nuclease I-Hmu1 (PDB id.: 1U3E). K1 and K2-like residues in one-metal-ion enzymes are not as well conserved in sequence and structure as in two-metal-ion enzymes (see main text for details). Structural elements are oriented and color coded as in Fig. 1.



Supplementary Fig. 5.3: K1- and K2-like residues in the spliceosome.
Representation of the spliceosome (PDB id.: 5LJ3) active site (A) and electrostatic distribution (B) around the active site. Structural elements are oriented and color coded as in Fig. 1 and 3.



Supplementary Table 5.1. DNA/RNA processing enzymes adopting a group II intron-like architecture of the active site.

Enzyme (Organism)	PDB id	Enzymatic classification (EC)	K1	d1 (Å)[a]	d-ac (Å)[b]	mutant	functional defect	K2	d2 (Å)[a]	d-sub(Å)[c]	mutant	functional defect
Spliceosome (*S. cerevisiae*)	5LJ3[212]	ribozyme	K1?[d]	~3.8	~2.6	n.a.	n.a.	Lys611 (Prp8) Arg614 (Prp8)	9.08 9.88	3.67 3.38	n.a. n.a.	n.a. n.a.
BglI (*B. subtilis*)	1DMU[213]	3.1.21.4	Arg118	7.51	via Pro115	n.a.	n.a.	Lys144	4.31	3.24	n.a.	n.a.
EcoRI (*E. coli*)	1QPS[e]	3.1.21.4	His114	6.23	via Lys113 of PD-(D/E)-XK motif	H114Y/F/S/T; H114P[214]	Lost substrate specificity; impaired activity	Lys148	11.50	2.85	computational model[215]	Reduced stability
PvuII (*P. vulgaris*)	1F0O[216]	3.1.21.4	Lys70	4.31	3.81	K70A[217]	No activity	Lys93	7.84	3.42	n.a.	n.a.
EcoRV (*E. coli*)	1STX, 1EO3[25,218]	3.1.21.4	Lys38 Arg49	6.39 5.13	7.14 3.81	K92A[219]	Reduced activity	Lys119 Lys173	18.15 9.39	3.05 7.05	K119A[220]	Reduced activity and reduced DNA binding affinity
MspI (*Moraxella sp.*)	1SA3[221]	3.1.21.4	Na^+ Lys32	x[f] x[f]	2.67 7.87	n.a.	n.a.	Lys97	x[f]	via H_2O	n.a.	n.a.
BpuSI (*B. pumilus*)	3S1S[222]	3.1.21.4	His26	4.19	3.36	n.a.	n.a.	Lys856	12.96	x[f]	n.a.	n.a.
DpnI (*S. pneumoniae*)	4KYW[223]	3.1.21.4	His44	8.00	3.28	n.a.	n.a.	Arg135	13.78	2.88	R135A[223]	Reduced DNA cleavage
FokI (*P. okeanokoites*)	1FOK[224]	3.1.21.4	His442	x[f]	3.11	n.a.	n.a.	Arg447 Arg487	x[f] x[f]	x[f] x[f]	n.a. R487A[225]	n.a. No activity
NaeI (*L. aerocolonigenes*)	1IAW[226]	3.1.21.4	His74	x[f]	4.23	n.a.	n.a.	Lys59	x[f]	3.58	n.a.	n.a.
NgoMIV (*N. gonorrhoeae*)	1FIU[227]	3.1.21.4	(Lys187)[g]	-	-	-	-	Arg102	12.12	via H_2O	n.a.	n.a.
SfiI (*S. fimbriatus*)	2EZV[228]	3.1.21.4	(Lys102)[g]	-	-	-	-	Arg218	9.78	3.30	n.a.	n.a.
SgrAI (*S. griseus*)	3DVO[229]	3.1.21.4	(Lys242)[g]	-	-	-	-	Arg152	6.29	2.83	n.a.	n.a.
FluA, endonuclease PA subunit (Influenza A virus)	2W69[179]	2.7.7.48	His41 Lys134	2.91 4.17	3.47 2.77	n.a. K134A[230]	Reduced thermal stability and no activity	Arg84 Lys137	6.74 13.33	x[f] x[f]	n.a. K137A[230]	Decreased activity
FluB, endonuclease PA subunit (Influenza B virus)	4WRT[231]	2.7.7.48	His41 Lys135	x[f] x[f]	x[f] x[f]	n.a.	n.a.	Arg85 Lys138	x[f] x[f]	x[f] x[f]	n.a.	n.a.
MutH (*H. influenzae*)	2AOR[232]	3.1.21.4	Lys79	4.67	3.51	K79E[233]	Reduced DNA binding affinity and no activity	Lys66 Lys116	7.27 6.90	5.00 5.35	K116E[233]	Reduced DNA binding affinity and no activity

Exo-I (*H. sapiens*)	3QE9[234]	3.1.11.3	Lys85	5.06	3.55	n.a.	n.a.	Arg92	7.88	3.24	n.a.	n.a.
Exo-1 (*E. coli*)	4HCC[e]	3.1.11.3	His181	5.62	4.24	H181A[235]	Strongly reduced activity	Arg165	11.20	4.28	n.a.	n.a.
PhoExo-I (*P. horikoshii*)	4YOY[236]	3.1.11.3	Arg142	6.71	2.79	n.a.	n.a.	Lys136	9.65	3.88	n.a.	n.a.
RNase-H (*T. maritima*)	3O3H[237]	3.1.26.4	Lys47	4.50	3.37	n.a.	n.a.	Lys122	6.78	6.65	n.a.	n.a.
RNase-H (*H. sapiens*)	3PUF[238]	3.1.26.4	Arg186	x[f]	x[f]	R186W[238]	Impaired substrate cleavage *in vitro* and Aicardi-Goutières Syndrome *in vivo*	Lys69	x[f]	x[f]	K69A[238]	Reduced enzymatic activity due to substrate mispositioning
Cas9, RuvC domain (*A. naeslundii*)	4OGC[160]	3.1.-.-	Arg730	5.98	3.63	n.a.	n.a.	Lys711	11.43	x[f]	n.a.	n.a.
Cas9, RuvC domain (*F. tularensis*)	5B2O[239]	3.1.-.-	Arg1137	6.95	2.61	n.a.	n.a.	Lys1142	21.36	x[f]	n.a.	n.a.
Cas9, RuvC domain (*S. aureus*)	5AXW[240]	3.1.-.-	Arg694	x[f]	2.87	n.a.	n.a.	Arg686	x[f]	x[f]	n.a.	n.a.
DNA Pol-β (*H. sapiens*)	1BPY[241]	2.7.7.7	Arg183 Arg149	6.82 8.48	2.96[h]	R183A R149A[242]	Reduced polymerization activity rate	Arg254	5.42	3.36	R254A[243]	Reduced polymerization activity rate
DNA Pol-ι (*H. sapiens*)	2ALZ[104]	2.7.7.7	Lys214	6.29	via H_2O	n.a.	n.a.	Lys207	6.00	2.54	n.a.	n.a.
DNA Pol-κ (*H. sapiens*)	1T94[108]	2.7.7.7	Lys328	x[f]	x[f]	n.a.	n.a.	Lys321	x[f]	x[f]	n.a.	n.a.
DNA Pol-η (*S. cerevisiae*)	3MFI[244]	2.7.7.7	Lys279	6.65	via H_2O	K279A[245]	Decreased efficiency	Lys272	4.88	3.21	n.a.	n.a.
DinB (*E. coli*)	4IRK[246]	2.7.7.7	Lys157	7.17	3.61	n.a.	n.a.	Lys150	9.77	2.46	n.a.	n.a.
Dpo4 (*S. solfataricus*)	2AGO[68]	2.7.7.7	Lys159	3.96	3.31	n.a.	n.a.	Lys152	5.70	4.46	n.a.	n.a.
Dbh (*S. solfataricus*)	1IM4[247]	2.7.7.7	Lys160	x[f]	x[f]	n.a.	n.a.	Lys153	x[f]	x[f]	n.a.	n.a.
RNA Pol-I (*S. cerevisiae*)	5M5Y[184]	2.7.7.6	Lys934 (RPA190) Arg957 (RPA135)	x[f] x[f]	8.49 5.13	n.a.	n.a.	Lys916 (RPA135) Lys924 (RPA135)	x[f] x[f]	3.44 2.36	n.a.	n.a.
RNA Pol-III (*S. cerevisiae*)	5FJ8[248]	2.7.7.6	Lys800 (Rpc1) Arg952(Rpc2)	x[f] x[f]	9.71 7.55	n.a.	n.a.	Lys911 (Rpc2) Lys919 (Rpc2)	x[f] x[f]	3.75 4.45	n.a.	n.a.
Norwalk RdRp (Norwalk virus)	3BSO[249]	2.7.7.48	Lys374	5.55	4.40	n.a.	n.a.	Arg392	8.29	2.58	n.a.	n.a.

Sapporo RdRp (Sapporo virus)	2UUW[e]	2.7.7.48	Lys380	x[f]	x[f]	n.a.	n.a.	Lys394	x[f]	x[f]	n.a.	n.a.
RHDV RdRp (Rabbit haemorragic disease virus)	1KHV[250]	2.7.7.48	Lys387	5.82	5.43	n.a.	n.a.	Lys403	3.96	x[f]	n.a.	n.a.
Poliovirus RdRp (Poliovirus)	1RDR[251]	2.7.7.48	Lys359	12.35	9.48	n.a.	n.a.	Lys375	9.77	x[f]	n.a.	n.a.
FMDV RdRp (Foot-and-mouse-disease virus)	1U09[252]	2.7.7.48	Lys369	x[f]	x[f]	n.a.	n.a.	Lys387	x[f]	x[f]	n.a.	n.a.
FluA, RdRp PB1 subunit (Influenza A virus)	4WSB[231]	2.7.7.48	Lys235 (PB1) Lys481 (PB1)[4]	x[f] x[f]	x[f] x[f]	n.a.	n.a.	Arg233 (PB1) Arg658 (PA)	x[f] x[f]	x[f] x[f]	n.a.	n.a.
FluB, RdRp PB1 subunit (Influenza B virus)	4WRT[231]	2.7.7.48	Lys235 (PB1) Lys480 (PB1)	x[f] x[f]	x[f] x[f]	n.a.	n.a.	Arg233 (PB1) Arg658 (PA)	x[f] x[f]	x[f] x[f]	n.a.	n.a.
HCV RdRp (Hepatitis C virus)	1QUV[253]	2.7.7.48	Arg222	x[f]	x[f]	n.a.	n.a.	Arg386	x[f]	x[f]	n.a.	n.a.
Dengue RdRp (Dengue virus)	2J7U[254]	2.7.7.48	Lys689	x[f]	x[f]	n.a.	n.a.	Arg729	x[f]	x[f]	n.a.	n.a.
1-Hmu1 (*B. virus* SPO1)	1U3E[255]	3.1.-.-	Arg89	6.48	3.30	n.a.	n.a.	Arg103	4.75	3.54	n.a.	n.a.
Cas9, HNH domain (A. *naeslundii*)	4OGC[160]	3.1.-.-	(Lys609)[f]	5.16	2.97	-	-	(Arg607)[g]	10.07[d]	9.71	-	-
Cas9, HNH domain (*S. pyogenes*)	4CMQ/5F9R[160,161]	3.1.-.-	(Lys866)[f]	x[f]	4.03	-	-	(Lys862)[g]	-	-	-	-
Cas9, HNH domain (*F. tularensis*)	5B2O[239]	3.1.-.-	(Lys598)[f]	x[f]	x[i]	-	-	(Lys996)[g]	-	-	-	-
Cas9, HNH domain (*S. aureus*)	5AXW[240]	3.1.-.-	(Lys583)[f]	x[f]	2.48	-	-	(Lys582)[g]	-	-	-	-

[a] closest distance between the ion / amino acid corresponding to K1 or K2 and the M_A-M_B center; [b] closest distance between the ion / amino acid corresponding to K1 and the acidic residues that chelate M_A-M_B; [c] closest distance between the ion / amino acid corresponding to K2 and the substrate; [d] putative, see main text for details; [e] unpublished; [f] no substrate or ions in the active site; [g] these residues are unlikely to play the same role as K1- or K2-like residues, see text for details; [h] from K2-like residue; [i] Lys598 is disordered in the structure; n.a. = not available.

Supplementary File 5.1: sequence alignments and conservation of K1-like and K2-like (in red) residues for different classes of two-metal-ion depend enzymes.

1) Multiple sequence alignment of restriction endonuclease BamHI

Legend

```
>B_amyloliquefaciens   >2BAM:B|PDBID|CHAIN|SEQUENCE
>B_subtilis            >gi|503118864|ref|WP_013353547.1|
>B_cereus              >gi|822525455|ref|WP_046955369.1|
>G_stearothermophilus  >gi|696476303|ref|WP_033015965.1|
>O_valericigenes       >gi|503884873|ref|WP_014118867.1|
>Scytonema_sp_HK-05    >gi|1121323857|ref|WP_073633968.1|
>M_vaginatus           >gi|493682480|ref|WP_006632638.1|
>N_piscinale           >gi|1011379467|ref|WP_062292232.1|
>T_bouteillei          >gi|740242599|ref|WP_038083624.1|
>Marinomonas_sp_S3726  >gi|800993396|ref|WP_046019508.1|
>Nostoc_sp_NIES-3756   >gi|1056316658|ref|WP_067770513.1|
```

Alignment

```
B_amyloliquefaciens       MEVEKEFITDEAKELLSKDKLIQQAYNEVKTSICSPIWPATSKTFTINNTEKNCNGVVPI
B_subtilis                MEVEKEFITDEAKELLSKDKLIQQAYNEVKTSICSPIWPATSKTFTINNTEKNCNGVVPI
B_cereus                  MKIEKEFITETAKNLLLTDSLIEQAYKEVKTSICSPVWPMESKIFTVNNTKKNCNGVVPI
G_stearothermophilus      MKVDKVYMTDIAKQLITSDKLCKQAYEEVITSIRSSVWPKGSNIFTINNSEKNVNGVVPL
O_valericigenes           MKIAKFYISPVADNLMRTCPAAVEAYHEVEQSIIENTTP-GKDIFILNDTSKNCNGVVPV
Scytonema_sp_HK-05        MKIVQE-VSLISIGSFEESSDWSIIRSEIRSAISLIVYPPGTSSFTINPTKHG-NGVKPI
M_vaginatus               MKIVQE-VSLISRGSFEESEEWGVIKNEIRIAIDAIAWPVGASNFTINPTRHG-NGVKPI
N_piscinale               MKIVQE-VSLISIGSFEESSDWAIIRSEIRDAIALIVHPTGTSSFTINPKKHG-NGVKPI
T_bouteillei              MKIVQE-VSLMSRGSFEKSQQWVTIQNEIRSAIQLIVWPPGTSNFTINPTPHG-NGVKPI
Marinomonas_sp_S3726      MKITRT-EVLINSGGFFDTQQFNDVLAEIEESISKVVWPLNSQYFSINPTKKG-NGVKPI
Nostoc_sp_NIES-3756       MKIVQE-VSLINIGSFAESSDWSIIRAEIRNAISVIVHPPGTSSFTINPKKHG-NGVKPI
                          *:: :          :           *:  :*     *   . * :* . :. *** *:

B_amyloliquefaciens       KELCYTLLEDTYNWYREKPLDILKLEKK-KGGPIDVYKEFIEN----------------S
B_subtilis                KELCYTLLEDTYNWYREKPLDILKLEKK-KGGPIDVYKEF---------IENSE------
B_cereus                  KELCYTTLEETYNWYREKPLNVLKVEKK-KGGPIDVYKEFSAGLSEKKELDNLEGLNQVD
G_stearothermophilus      KENCYIMLEETYNWFREKPLDVLKYEKK-KGGPIDVYKEF---------RDGDT------
O_valericigenes           KERCYQILEEDHLWYREKPLSYFHDDAQ-KGGPIDVYKEF---------RTPSG------
Scytonema_sp_HK-05        KEACMTTLRDQFGWRLETPI---RYATK-SPGKVDATK-V---------IDDYL------
M_vaginatus               KNACMAALHDNFGWQLETKI---RFATR-APGRVDATK-I---------LDNHL------
N_piscinale               KEACMIALRDRFGWRLEAPI---NYATK-SPGKVDATK-V---------IDNYF------
T_bouteillei              KNACMAFLKETFGWQLETKI---TYATK-SPGRVDATK-A---------LNGHL------
Marinomonas_sp_S3726      KNGCMSHL-EGLGWQLEERL---RITSTMRPGPLDAVKTL---------PNGSV------
Nostoc_sp_NIES-3756       KEACMLALKDKFDWRLETAI---NYATK-SPGKVDATK-V---------IDNHL------
                          *: *   * :   *  *  :           * :*. *

B_amyloliquefaciens       ELKRVGMEFETGNISSAHRSMNKLLLGLKHGEIDLAIILMPIKQLAYYLTDRVTNFEELE
B_subtilis                -LKRVGMEFETGNISSAHRSMNKLLLGLKHGEIDLAIILMPIKQLAYYLTDRVTNFEELE
B_cereus                  NIRRVGMEFETGNISSAHRSMNKLLLGLKREEIDLAIILMPIKKLAYYLTDRVTNFEELE
G_stearothermophilus      -VRRVGLEFETGNISSAHRSMQKLLLGLNRKELDMAIILMPVFELAYYLTDRVTNYEELE
O_valericigenes           -FFRAGLEFETGNISSAHRSMNKLCVGILKGEIDLAMLMMPIKQMSYYLTDRVSNYEELE
Scytonema_sp_HK-05        ----FALEWETGNISSSHRAVNKLVLGLLRGVFLGAALVLPSRKLYPYLTDRIGNYEELE
M_vaginatus               ----FAFEWETGNISSSHRAVNKLVLGILRGIFLGTALVLPSRQLYPYLTDRIGNYEELE
N_piscinale               ----FALEWETGNISSSHRAVNKMVLGLIRGVFLGAALVLPSRKLYPYLTDRIGNYEELE
T_bouteillei              ----FALEWETGNISSSHRAVNKLVLGLLRGVFLGSALVLPSRKLYPYLTDRVGNYEELE
Marinomonas_sp_S3726      ----LALEWETGNISSSHRALNKMVLGILEGALTGGILILPSRAMYKYLTDRIGNFQEIE
Nostoc_sp_NIES-3756       ----FALEWETGNISSSHRAVNKLVLGLLRGIFLGAALVLPSRKLYPYLTDRIGNYEELE
                               .:*:*******:**:::*: :*: .  :    :::*   :  *****: *::*:*

B_amyloliquefaciens       PYFEL------TEGQPFIFIGFNAEAYNSNVPLIPKGSDGMSKRSIKKWKDKVENK  213
B_subtilis                PYFEL------TEGQPFIFIGFNAEAYNSNVPLIPKGSDGMSKRSIKKWKDKVENK  213
B_cereus                  PYFEL------TEGQPFMFIGFDAEAYNVNVPVIPKGSDGMSDRSIKKWKDKIEDL  229
G_stearothermophilus      PYFEN------AEGKAFVFIGFNADAFDSSVEIIPKGKDGMSKRSIKKWIQKKD--  211
O_valericigenes           PYFILL------DNYPFVVFGFDAEQYRNQAPFLPKGKDGMSPRTKRKWQNNHD--  210
Scytonema_sp_HK-05        PYFDVWRAVRIQEGFLAIFV-IEHDQVDTSVPRITKGTDGRALV------------  197
M_vaginatus               PYFDVWRAVNIQEGFLAIFV-IEHDALDSSVPTITKGTDGRALI------------  197
N_piscinale               PYFDVWRSVQLQEGFLAIFV-IEHDQLDSNVPTLTKGTDGRALV------------  197
T_bouteillei              PYFDVWRSVNINEGFLAIFV-VEHDAQDSNVARITKGTDGRALI------------  197
Marinomonas_sp_S3726      PYFPQWRRPDISNGYLAVIE-IEHEYEDENSPLIPKGTDGWAEFQGS---------  201
Nostoc_sp_NIES-3756       PYFDVWRSVHLSEGFLAIFV-IEHDQLDSNVPTLSKGTDGRALI------------  197
                          ***         :.   :.  .: :    .   : **.** :
```

2) Multiple sequence alignment of restriction endonuclease BglI

Legend

```
>B_subtilis                 >1DMU:A|PDBID|CHAIN|SEQUENCE
>B_atrophaeus               >gi|737721210|ref|WP_035690053.1|
>Y_intermedia               >gi|912843510|ref|WP_050288336.1|
>N_bacterium_HCH-1          >gi|983041745|gb|KWT84937.1|
>Geitlerinema_sp_PCC7105    >gi|1102700459|ref|WP_071592032.1|
>N_eutropha                 >gi|1124648614|ref|WP_074928903.1|
>Planktothricoides_sp_SR001 >gi|935599881|ref|WP_054465660.1|
>P_agardhii                 >gi|754791914|ref|WP_042155505.1|
>Cyanobacterium_ESFC-1      >gi|517210223|ref|WP_018399041.1|
```

Alignment

```
B_subtilis                      MYNLHREKIFMSYNQNKQYLEDNPEIQEKIELYGLNLLNEVISDNEEEIRADYNEANFLH 60
B_atrophaeus                    MYNLHREKIFMSYNQNKQYLEDNPEIQEKIELYGLNLLNEVISDNEEEIRADYNEANFLH 60
Y_intermedia                    MFNINRFQQYEAYNENRAYLIDNPDVLINLERFFLIKISELIKQHSLGIQSDYNEASFLY 60
N_bacterium_HCH-1               MFNKFRKEQQKIYINIRNHLSNRPEILINLENYFADYTGKLLQENIESIKSDYNEASYLY 60
Geitlerinema_sp_PCC7105         MFNTFRELQYYSYNRARNYFRDNYIQLVKLENFLSKQVFQLIQNYIIQIKEDYNEASYLY 60
N_eutropha                      MINIHRQNQFNIYNSARNHFIANPSSLIELEKFLTNYLVSIITANIVEIKQDYNEASYLY 60
Planktothricoides_sp_SR001      MFNKFRNSQYSIYKRARKYFIQNYNQLIDIEKFVSIKFYEIVNNNLQQIVSDFNEASNLY 60
P_agardhii                      MFNKFRNSQYSIYNQARNYFIQNYDQLIGIEKFIALKIYEIVNNNIQQIANDFNEASNLY 60
Cyanobacterium_ESFC-1           MFNKFRKSQYSLYNQARTHFLQHYNQLINIEKFVSTKIDEIINKNLQQIVNDFNEASNLY 60
                                * *  *      *   : ::  .      :* :      .::      *  *:***. *:

B_subtilis                      PFWMNYPPLDRGKMPKGDQIPWIEVGEKAVGSKLTRLVSQREDITVREIGLPTGPDERYL 120
B_atrophaeus                    PFWMNYPPLDRGKMPKGDQIPWIKVGEKAVGSKLTRLVSQREDITVREIGLPTGPDERYL 120
Y_intermedia                    PFWQNYPPDNRGRQPRGDQYPWIEVGEHSVGRKLSRHLA--EHFFVKDIGVPTGADERFL 118
N_bacterium_HCH-1               PFWQQYPPDNRGRQPRGDQFPWIEVGEHAIGDKLPRLFQ--KDFSLRDVGLPTGPDKRFI 118
Geitlerinema_sp_PCC7105         PFWQNYPPEERGRQPIGDQYPWIEVGEHSLGGKLYRLLS--LSFNIRDIGLPAGSDIRVV 118
N_eutropha                      PFWENYPPEDRGRQPIKDQYPWIEVGEHAIGSKLPRLLDS--VFRVRDTGLPTGSDQRFV 118
Planktothricoides_sp_SR001      PFWQNYPPDDRGRSPIGDQYPWIEVGEHTIGYKLPRLLE--PYFRIRDIGLPSGSDLRLV 118
P_agardhii                      PFWQNYPPEERGRYPIGDQYPWIEVGEHSIGDKLPRLLE--PYFSIRDVGLPTGADVRLV 118
Cyanobacterium_ESFC-1           PFWQNYPPDNRGRAPIGDQYPWIEVGEHTIGEKLPRLLE--PYFHIRDMGLPSGTDVRLV 118
                                *** :*** :**: *  ** ***:***:::* ** * .     : ::: *:*:* * * :

B_subtilis                      LTSPTIYSLTNGFTDSIMMFVDIKSVGPRDSDYDLVLSPNQVSGNGDWAQLEGGIQNNQQ 180
B_atrophaeus                    LTSPTIYSLTNGFTDSIMMFVDIKSVGPRDSDYDLVLSPNQVSGNGDWAQLEGGIQNNQQ 180
Y_intermedia                    ISSKEILNYSGGMYENVFLSIDIKSVGPRDDAHHAVMSHNQISGDGCWDNIEKGVYNTPL 178
N_bacterium_HCH-1               VTSPLIAEIT-GFTNSAWLFIDIKSVGPRDDADHTVMSHNQISGDGTWTDLEKGIKNSVM 177
Geitlerinema_sp_PCC7105         LSSDKIYQITEGFTNSCWLFVDIKSVGPRDDWNHAVMSHNQISGSGRWDSLLSGITNDVI 178
N_eutropha                      LTDDAIATATGGFTNSVWFFVDIKSVGPRDDQHHTVMSHNQVSGDGVWINPVDGVRNTIL 178
Planktothricoides_sp_SR001      LTHSEINKLTNSFTDTCWLFLDIKSVGPRDDQNHAVMSPNQISGSGRWDSADSGVVNDVI 178
P_agardhii                      LTHPEINNLTNSFTDTCWLFLDIKSVGPRDDQSHAVMSPNQISGSGIWDSVDGGVSNTVI 178
Cyanobacterium_ESFC-1           LTNAEINKLTNRLTDTCWLFLDIKSVGPRDDQNHAVMSPNQISGNGRWDAEDNGVVNDVI 178
                                ::   *   :  : :.  : :*********.  . *:* **:**.* *     *: *

B_subtilis                      TIQGPRSSQIFLPTIPPLYILSDGTIAPVVHLFIKPIYAMRSLT---KGDTGQSLYKIKL 237
B_atrophaeus                    TIQGPRSSQIFLPTIPPLYILSDGTIAPVVHLFIKPIYAMRSLT---KGDTGQSLYKIKL 237
Y_intermedia                    VARGIRAYHDFHCSLPPLYVLSDRTVAPVITIVVKPVYEMYSLYPTANYQSGQPLKRISL 238
N_bacterium_HCH-1               TAIGSREHHSFYCSIPPLYVLSDRTVAPVVILAIKPVYKMLGL--DNNTVSGQPLSRLEL 235
Geitlerinema_sp_PCC7105         IACGKRSSHPFYCSIPPIYVLSDGTVVPVVIIILKPVYDMLSLE-FDVSDGGQPLSRISC 237
N_eutropha                      QATGARASHDFHASLPPVFVLSDGTIAPLVMIALKPVYRMLQPNVVGARNDGQPLERIDI 238
Planktothricoides_sp_SR001      VAKGKRKSQAFYCSIPPIYILSDGTMIPVIILIVKPVYRMLSLE-ENSKDGGQPLGRISL 237
P_agardhii                      VAKGRNKSHLFHASIPPIYILSDGTVIPVIIVILKPVYKMLSLE-EQSEDGGQPLGRISF 237
Cyanobacterium_ESFC-1           VAKGKRKSQDFYCSIPPIYVLSDGTILPVIILIVKPVYRMLSLE-ENSKDGGQPLGRISL 237
                                   * .  : *  ::**:::*** *: *:: : :**:* *           ** * ::.

B_subtilis                      ASVPNGLGLFCNPGYAFDSAYKFLFRPGKDDRTKSLLQKRVRVDLRVLDKIGP-RVMTID 296
B_atrophaeus                    ASVPNGLGLFCNPGYAFDSAYKFLFRPGKDDRTKSLLQKRVRVDLRVLDKIGP-RVMTID 296
Y_intermedia                    ASIPNGLLLNVNPNYL--GLYPGLFYPGKDDKGKSPLKVRARVDFNILQKIASWRYYDIF 296
N_bacterium_HCH-1               VSIPNGLLMEVNPRYL--KKYPNLLYPGKDDKSKNPLKMRCRISFALLREIAAWRVQEFL 293
Geitlerinema_sp_PCC7105         ATVPNGLLLCERPNYL--SEFPELFFPGKDDKTKNPRKKRCRVSFEILKQIEYWRFREIL 295
N_eutropha                      ACIPNGLLLTQQPNYL--GAYNGLLFPGKDDKSKDPRKLRARVSFELLKNIAPWRVQTIQ 296
Planktothricoides_sp_SR001      ATVPNGLLLQENPNYL--QQYPNLFFPGKDDRSTNYLKKRCRISFDVLKSIDNWRFKEIV 295
P_agardhii                      ATVPNGLLLHEQPNYL--AQYPNLFFPGKDDKNTNPQKMRCRVSFEVLKSIANWRFQEIV 295
Cyanobacterium_ESFC-1           ASVPNGILLQENPNYL--QQYPNLFFPGKDDQSTNPLKKRCRISFDVLKSIEAWRFKEIV 295
                                . :***: :  .* *     :  *: *****: ..  : * *:.: :* .*   *   :

B_subtilis                      MDK-      299
B_atrophaeus                    MDK-      299
Y_intermedia                    F---      297
N_bacterium_HCH-1               ISDA      297
Geitlerinema_sp_PCC7105         VS--      297
N_eutropha                      VPFP      300
Planktothricoides_sp_SR001      LP--      297
P_agardhii                      LK--      297
Cyanobacterium_ESFC-1           LP--      297
                                .
```

3) Multiple sequence alignment of restriction endonuclease EcoRI

Legend

```
>E_coli                 >1QPS:A|PDBID|CHAIN|SEQUENCE
>S_enterica             >gi|970558700|ref|WP_058820205.1|
>A_baumannii            >gi|746197619|ref|WP_039255571.1|
>C_achromatium_palustre >gi|860362112|ref|WP_048398711.1|
>R_anatipestifer        >gi|491055006|ref|WP_004916653.1|
>P_pleuritidis_F0068    >gi|544002746|gb|ERK02129.1|
>S_suis                 >gi|758839608|ref|WP_043025063.1|
>S_pneumoniae           >gi|895296590|emb|COC92604.1|
>S_moniliformis         >gi|1035745704|ref|WP_064580150.1|
```

Alignment

```
E_coli                      ----------------SQGVIGIFGDYAKAHDLAVGEVSKLVKKALSNEYPQLSFRYRDS     44
S_enterica                  MANKNQSNRLTDQHKLSQGVIGIFGDYAKAHDLAVGEVSKLVKDALGKEYPQLSFRYRDS     60
A_baumannii                 MANKNQSNRLTDQHKLSQGVIGIFGDYAKAHDLAVGEVSKLVKDALGKEYPQLSFRYRDS     60
C_achromatium_palustre      MAKNNQSNRLTSQHKDSHGIVGIFGLKAKYHDMTVEKISHSVIKQLKNEYPQLSFRYRTS     60
R_anatipestifer             MAKKNQSTRLTVQHKKSQGVVGIFGEKAKLHDLTLGEISHLVIKQLEEEYPQLTFQYKTS     60
P_pleuritidis_F0068         --------------------------------MTVGEVSHLALKQLQEEYPQLEFQYRTS     28
S_suis                      MAKKNQSSRLTNQHKASKGVVGIFGDEARTHDSAVGTISHLVKYELEQKYPKLEFRFRKS     60
S_pneumoniae                MAKKNQSNRLTNQHKDSRGVVGIFGEDAKSHDIAVGTISHLVKSKLEELYPMLEFRFRKS     60
S_moniliformis              M-KKGQSNRLTEQQKEGQGPITIFHEDAQVHDKEVYNTSITVKEKLEEEFPMLTFRYRKD     59
                                                              :   *  .   * : :* * *::: .

E_coli                      IKKTEINEALKKIDPDLGGTLFVSNSSIKPDGGIVEVKDDYGEWRVVLVAEAKHQGKDII     104
S_enterica                  IRKAEINEALKKIDPELGGTLFVSNSSIKPDGGIVEVKDDNDEWRVVLVTEAKHQGKDII     120
A_baumannii                 IRKAEINEALKKIDPELGGTLFVSNSSIKPDGGIVEVKDDNDEWRVVLVTEAKHQGKDII     120
C_achromatium_palustre      IKKEEINEALKKVDPELGQTLFVSNSSIIPDGGIIEVKDDNSNWRVVLVSEAKHQGKDID     120
R_anatipestifer             IRKEEINKALRKIDGELGQTLFVPNSSVKPDGGIIEVKDDNGNWRIILVSEAKHQGKDIE     120
P_pleuritidis_F0068         IKKEEINKALKKIDPGLGKTLFVSNSSIIPDGGIVEVKDDNGEWRIVLVTEAKHQGKDIE     88
S_suis                      ISKKEINDSLRKIDSELGQTLFNHNANIIPDGGIIEVRDDYGNWRVILVTEAKHQGKDIE     120
S_pneumoniae                VSKKEINHYLSKLDKDLGKTLFTQNASIIPDGGIIEVKDDSGSWRVVLVTEAKHQGKDIE     120
S_moniliformis              LSKKEINESLQKIDTYLGQTLFVDNAKIKPDGGIIEVKDDQGNWRVVLVSEAKHQGKDIE     119
                            : * ***. * *:*  ** ***  *:.: *****:**:** ..**::**:*********

E_coli                      NIRNGLLVGKRGDQDLMAAGNAIERSHKNISEIANFMLSESHFPYVLFLEGSNFLTENIS     164
S_enterica                  NIKNGILVGKTGSQDLMAAGNAIERSHKNISEIANFMLYESHFPYILFLEGSNFLTETIS     180
A_baumannii                 NIKNGILVGKTSSQDLMAAGNAIERSHKNISEIANFMLYESHFPYILFLEGSNFLTETIS     180
C_achromatium_palustre      NIKNGVLVGKDNNQDLMAAGNAIERSHKNISEIANLMLSESHFPYVLFLEGSNFLTETIS     180
R_anatipestifer             NIRKGRLVGKANNQDLMAAGNAIERSHKNISEIANFMLLESHFPYVLFLEGSNFLTETVS     180
P_pleuritidis_F0068         NIKAGKLVGAKNDQDLMAAGNAIERSHKNISEIANLMLAESHFPYVLFLEGSNFLTETIS     148
S_suis                      NIRAGKLVGKNNDQDLMAAGNAIERSHKNIAELANFMLSEIHFPYVIFLEGSNFLTETIS     180
S_pneumoniae                NIKSGKLVGKNNDQDLMAAGNAIERSHKNIAEIANFMLSEEHFPYIIFLEGSNFLTQTIS     180
S_moniliformis              NIMVGKLVGKKGNQDLMVAGNAIERAYKNINEIANFMLSERHFPYILFLEGSNFLTQNVT     179
                            **  * ***  ..****.*******::*** *:**:** * ****::*********:.::

E_coli                      ITRPDGRVVNLEYNSGILNRLDRLTAANYGMPINSNLCINKFVNHKDKSIMLQAASIYTQ     224
S_enterica                  ITRPDGRVVKLEYNSGMLNRLDRLTAANYGLPINSNLCVNKFVKHKDKTIMLQAASIYTQ     240
A_baumannii                 ITRPDGRVVKLEYNSGMLNRLDRLTAANYGLPINSNLCVNKFVKHKDKTIMLQAASIYTQ     240
C_achromatium_palustre      ITRPDGRVVTLEYNSGMLNRLDRLTSANYGMPINTNLCKNKFIKHKDKTIMLQATSIYTQ     240
R_anatipestifer             IERPDGRIVRLEYNSGILNRLDRLTAANYGMPINTNLCVNKFIKHKDKTIMLQATSIYTQ     240
P_pleuritidis_F0068         VKRPDGRVVVLEYNSGMLNRLDRLTAANYGMPINKNLCENKFVKHNEKTIMLQAASIYTQ     208
S_suis                      VERPDGRVVTLEYNSGILNRLDRLTAANYGLPFNTNLCRNKFVKCENRSIMLQAASMYTT     240
S_pneumoniae                VIRPDGRSVILEHDSGILNRLDRLTAANYGMPINTNLCKNKFIKNKDKSIMLQATSIYTT     240
S_moniliformis              IARPDGREVTLIYKDGALNRLDRLTAANYGMPLNTNLCENRFVNGNDVNIMLQAASIYTK     239
                            : ***** * * :..* ********:****:*:*.*** *:*:: :: .*****:*:**

E_coli                      GDGREWDSKIMFEIMFDISTTSLRVLGRDLFEQLTSK-------    261
S_enterica                  GNGEKWDAKIMFEIMFDVSTSSLRVLGRDLFNQLT------AK-    277
A_baumannii                 GNGEKWDAKIMFEIMFDVSTSSLRVLGRDLFNQLT------AK-    277
C_achromatium_palustre      GNGEKWDIQKMFDIMLEVSKTSLQVLGSDLFNQIT-----KSK-    278
R_anatipestifer             GKGEKWDNKEMFDIMLTISKTSLKVLGSDLFNQIT-----KNK-    278
P_pleuritidis_F0068         GNGEKWKVDKMLEIMLDISRTSLQMLGRDLFSQLT--KNKKSK-    249
S_suis                      GNGSHWESQAMLEIMLDVAQTSLNILGSDLFNQLTIEEEKNEKE    284
S_pneumoniae                GDGSH---------------------------------------    245
S_moniliformis              GSGGHWNDNDMINIMLEVARTSLKILGSDIFDQL-RKTTV----    278
                            *.* .
```

4) Multiple sequence alignment of restriction endonuclease PvuII

Legend

```
>P_vulgaris                 >1F0O:B|PDBID|CHAIN|SEQUENCE
>Aeromonas_sp_EERV15        >gi|1057711761|ref|WP_068979542.1|
>A_ferrivorans              >gi|917378215|ref|WP_051984927.1|
>P_savastanoi               >gi|929544999|ref|WP_054083891.1|
>Bradyrhizobium_sp_ORS375   >gi|918664917|ref|WP_052536119.1|
>Synechococcus_sp_PCC8807   >gi|1047320646|ref|WP_065716609.1|
>A_baumannii                >gi|914250353|ref|WP_050560546.1|
>R_leguminosarum            >gi|739279779|ref|WP_037142586.1|
>Arthrospira_sp_PCC8005     >gi|495328681|ref|WP_008053422.1|
>V_barjaei                  >gi|1027659813|ref|WP_063605258.1|
```

Alignment

```
P_vulgaris                    ------------------------------------------------------------   0
Aeromonas_sp_EERV15           ------------------------------------------------------------   0
A_ferrivorans                 ------------------------------------------------------------   0
P_savastanoi                  ------------------------------------------------------------   0
Bradyrhizobium_sp_ORS375      ------------------------------------------------------------   0
Synechococcus_sp_PCC8807      ------------------------------------------------------------   0
A_baumannii                   MIEEV-----FARNLKKIREHMNLSQEQLAEKCDLDRTYIGILERGEKVPTLTTVEKCAN   55
R_leguminosarum               ------------------------------------------------------------   0
Arthrospira_sp_PCC8005        ------------------------------------------------------------   0
V_barjaei                     MINKEGLLLQLADNVRLLRQELNWTQEYLAEVSDLSPRQISRVENLENEPSLDVVCAIAN   60


P_vulgaris                    -----------------MSHPDLNKLLELWPHIQEYQDLALKHGINDIFQDNGGKLLQVL   43
Aeromonas_sp_EERV15           ---------MA-H-------PDIEHLRALWPYIREYQSLATKHGINDIFQDNGGKLLQVL   43
A_ferrivorans                 ---------MTPH-------DDYARLLEIWPSVQEYQALATKHGIDDIFQDNGGKLLQVL   44
P_savastanoi                  ---------M-PH-------DDYNILTSVWPAVKEYQALATAHGIDDIFQDNGGKLLQVL   43
Bradyrhizobium_sp_ORS375      ---------MKPSSQED--VDEFKRL---WPSIQAYQDLASKHGIQDIFQDNGGKIVQVL   46
Synechococcus_sp_PCC8807      ---------MNFHP-------DKRILDELFPYIQRYQELASKHGINDIFQDNGGKLLQVL   44
A_baumannii                   ALGVSVIDLLSNDLYLDSFSSDKETLEAIWPFIRKYQELATKNGINDIFQDNGGKLLQVL   115
R_leguminosarum               ---------MKPHE-------DKARMDELMPAIKEFQTLATKHGIGDVFQDNGGKLLQVI   44
Arthrospira_sp_PCC8005        ---------MKSHP-------DKAILDELFPYIQQYQALATKHGINDIFQDNGGKLLQVI   44
V_barjaei                     AFNLHPSKLYEPIF----FESKIEHLNHIFPSIREMELLAKEEGIKDIFQDNGGKLLQVL   116
                                                   .   :    * ::  : **  .** *:*******::**:

P_vulgaris                    LITGLTVLPGREGNDAVDNAGQEYELKSINIDLTKGFSTHHHMNPVIIAKYRQVPWIFAI   103
Aeromonas_sp_EERV15           LITNLKVLPGREGNDAVDLSGQEYELKSINIDLTKAFSTHHHMNPTIIAKYRKVPWVFAI   103
A_ferrivorans                 LLMGLKILPGREGNDAVDASGREYELKSVNIELTKGFSTHHHMNPTIIAKYRQVPWAFAI   104
P_savastanoi                  LLMNLRVLPGREGNDAVDSSGREYELKSVNIELTRGFSTHHHMNPVIIAKYRKVPWVFAM   103
Bradyrhizobium_sp_ORS375      LLLGLTNIAGREGNDAVDSDGREYELKSVNIELTTGVSTHHHMNPTIIAKYRQVDWIFAI   106
Synechococcus_sp_PCC8807      LITGLEVLPGREGNDAKDSDGNEYELKSVNIQLTKSFSTHHHINPRIIEKYRKVDWLFAV   104
A_baumannii                   LNLDLKVLEGREGNDAVDESGQEYELKSLNIELVKGFSTHHHMNPVIIAKYRQVPWIFAI   175
R_leguminosarum               LTLGLTVLPGREGNDAKDDEDREYELKSVNLDLTKGFSTHHHLNPVILKKYRQVDWIFAI   104
Arthrospira_sp_PCC8005        LVTGLQIIPGREGNDAKDADGNEFELKSVNIALTKSFSTHHHINPRIIDKYRQVDWIFAV   104
V_barjaei                     LVTGLKDLPGREGNDAVDANGYEYELKSLNVNLVRGFSTHHHMNPKIIEKYREVDWVFAV   176
                              *  .*  : ******* *  . *:****:*: *. ..*****:** *: ***:* * **:

P_vulgaris                    YRGIAIEAIYRLEPKDLEFYYDKWERKWYSDGHKDINNPKIPVKYVMEHGTKIY------   157
Aeromonas_sp_EERV15           YSGIEIKSIFLLKPDDLEPYYSAWEKKWHDDGNKDINNPKIPVRYVMEVGKQIF------   157
A_ferrivorans                 YRHIALQAVYLLEPSDLEFYFTKWEEKWHADGEKDINNPKIPVVYVMKHGRLIHGTSPEI   164
P_savastanoi                  YRHIELHAVYILEPSDLEFYFSKWEEKWYADGHKDINNPKIPVAYVMQHGRLVYGNAPVF   163
Bradyrhizobium_sp_ORS375      YRNIELQRVYKLSPKDLEPYFSKWEEKWRSDGGKDINNPKIPIKFVKEVGTLVYETDVSV   166
Synechococcus_sp_PCC8807      YSGINLVSIYQLTPADLEFYYEQWERKWNDSGGKDINNPKISLTYVREHGFLIYEA----   160
A_baumannii                   YKNIELQAIYRLEPEDLEEMYVKWEEKWHRDGGKDINNPKIPLKYVMEKGKLMSGTVPII   235
R_leguminosarum               YRGIELICVHRLKPWQLEPIFERWEKKWHDEGGKDINNPKIPVAYVLEQGELLDGAVPDI   164
Arthrospira_sp_PCC8005        YRGINLLYIYKLTSSDLEYFYRKWEEKWYSKGGKDINNP---------------------   143
V_barjaei                     YKDIELQEIWLLTPEDLEFYYDKWTRQWHDRGGKDINNPKIPLTYVQDNGTLVY------   230
                              *  * :  :  *   :**  :  * .:*   * ******

P_vulgaris                    ----------------------------------   157
Aeromonas_sp_EERV15           ----------------------------------   157
A_ferrivorans                 SMRRK-----------------------------   169
P_savastanoi                  QSRRRVFTPPDAA-------GFGPDDI-------   183
Bradyrhizobium_sp_ORS375      T---EIAEDAKEAQDTVKE---------------   182
Synechococcus_sp_PCC8807      ----------------------------------   160
A_baumannii                   PSKTK--KKKDLGGEHLIEGGFKPDED-------   260
R_leguminosarum               KYRPSGIVPKKPPID-----PFMPDDEQPEQKAS   193
Arthrospira_sp_PCC8005        ----------------------------------   143
V_barjaei                     -------EKSDDGDLYFAD---ILDFT-------   247
```

5) Multiple sequence alignment of RNase-H

Legend

```
>T_maritima      >3O3H:A|PDBID|CHAIN|SEQUENCE
>F_pennivorans   >gi|1028459368|gb|ANE42301.1|
>P_elfii         >gi|655460282|ref|WP_028843263.1|
>P_lettingae     >gi|500834657|ref|WP_012002944.1|
>M_infera        >gi|973111565|gb|KUK68074.1|
>A_tabaci        >gi|737464781|ref|WP_035444766.1|
>B_cereus        >gi|488066538|ref|WP_002137935.1|
>E_cecorum       >gi|828150775|ref|WP_047242528.1|
>A_suicloacalis  >gi|1120435097|ref|WP_073297719.1|
>T_halophilus    >gi|1007070803|ref|WP_061840533.1|
```

Alignment

```
T_maritima          ------------------------------------------------------------    0
F_pennivorans       ---------MNE------------------------------------------------    3
P_elfii             ------------------------------------------------------------    0
P_lettingae         ------------------------------------------------------------    0
M_infera            --------------------------------------------MAFASCAEAYEMVMYN    16
A_tabaci            MTQPNKKKSISEIKELLKTITDEKDERLA-LIAQDERLGVQKALVSWK------K---AR    50
B_cereus            --M--QKMTIHEAECLLQEIMNEEDERFQ-MLVKDERKGVQKLILKWY------KQKELA    49
E_cecorum           --MS---ETIAQIKEKLAQIHDAQDAYVL-QLRKDERAGVQKLIQQFE------NRLAKE    48
A_suicloacalis      --M---KMTINEIKERLQYVTDYEDEFLL-ACQADERKGVQLAVKQWQ------KRQVAK    48
T_halophilus        --MA--KEAISEIKTHLLATETMTDPYVK-QLQMDERKGVQKLLLQLE------RRLAQK    49
                                                 *

T_maritima          ---------------GIDELYKKEFGIVAGVDEAGRGCLAGPVVAAAVVLEK--EIEGIN    43
F_pennivorans       --EKSKKVNWAEE-------KKHIDYSIIGVDEAGRGPLFGPVVAAAVYFDEGVYIEGIA    54
P_elfii             --MRSELFAYDRF-------YKQNFGTVIGVDEAGRGCLAGPVVAAAVILEVQ--LD-VF    48
P_lettingae         --MRSELFAYDRF-------YKQNFGTVIGVDEAGRGCLAGPVVAAAVILEVQ--LD-VF    48
M_infera            FSMKLSEEEFNKLVRF-DAAYRVDHKIVAGVDEAGRGPLAGPVVAAAVIVLNP--VEGVY    73
A_tabaci            QKIKEAEEERDRMLVYEKALWEREFHYVAGIDEVGRGPLAGPVVTASVVLPPDVSLVGIR    110
B_cereus            QKEREKFLEMSKY---EDELREKGLTYIAGIDEVGRGPLAGPVVTAAVVLPEDFYIPGLN    106
E_cecorum           QAMINKAKTMRQF---ENELLAKGYQAICGIDEVGRGPLAGPVVAAAVILPNDELILGLN    105
A_suicloacalis      QKLYDRYQRMNQL---EEN-YATTYSLIAGIDEVGRGPLAGPVVAAAVILDQDKQILGLN    104
T_halophilus        ETLKEQFYTMQAF---ERSCYKQGHHLIAGIDEVGRGPLAGPVVAAAVILPEGSEILGLN    106
                                               : *:**.*** * ****:*:* .     :  :

T_maritima          DSKQLSPAKRERLLDEIME-KAAVGIGIASPEEIDLYNIFNATKLAMNRALENLS--VKP    100
F_pennivorans       DSKALSEKQREELYNEIFARA-KFGLGLATPEEIDLYNIFHATELAMNRALEILSQFVEI    113
P_elfii             DSKQLTAQKREELFLQIMNSA-EVGIGIATPEEIDLYNIFNATKIAMNRALASLN--KKD    105
P_lettingae         DSKQLTAQKREELFLQIMNSA-EVGIGIATPEEIDLYNIFNATKIAMNRALASLN--KKD    105
M_infera            DSKALSRKIRESLFERIIENS-IVGIGLSSPEEIDLINVLAATRLAMNRALSVLS--ERP    130
A_tabaci            DSKKLSVSKREKLYEEIMDAAVSVGIGVVDAAQIDELNILQATKAAMKKSIEQLT--VQP    168
B_cereus            DSKKLSEAKRERFYDEIKEHAIAIGVGIISPQVIDEINIYQATKQAMLDAVANLS--CTP    164
E_cecorum           DSKQLSEKKRESLYQIIQEKAVAIGIGVVDETTIDAINIYQAARLAMTKAVEQLA--VQP    163
A_suicloacalis      DSKKLSLAKRVELFTKIKKEAVAVGIGVVSAEEIDKYNILGATKIAMKKAVSNLN--KQP    162
T_halophilus        DSKQLSEKKRLDLDNKIKEQATAIGIGEISAEQIDQVNIYQASKMAMTKAVENLA--VKA    164
                    *** *:   *  :   *      .*:*      **  *:  *:. **  ::  *

T_maritima          SFVLVNGKGIELSVPGTCLVKGDQKSKLIGAASIVAKVFRDRLMSEFHRMYPQFSFHKHK    160
F_pennivorans       KNVFVDGKNLKLNIPAVCVVKGDSKIYQISAASILAKVTRDKIMEKFHAQYPEYNLIKHK    173
P_elfii             AYVLVDGKSLNLSQQGVCIVKGDEKSASIAAASIVAKVLRDRIMVAYDRIYPCYGFSKHK    165
P_lettingae         AYVLVDGKSLNLSQQGVCIVKGDEKSASIAAASIVAKVLRDRIMVAHDRIYPCYGFSKHK    165
M_infera            DYVIVDGKWLRLDVEGECVVRGDRKSASIASASIIAKVFRDRIMDSLDSLYPEYGYRRHK    190
A_tabaci            DYLLVDALELPLPIPQTSIIKGDATSLSIAAASIIAKVTRDRMMTEYDELYPGYGFSKNA    228
B_cereus            EYLLIDAMKLPTSIPQTSIIKGDAKSVSISAASIIAKVTRDRMMKELGGKYPAYGFEQHM    224
E_cecorum           DYLLIDAMELDLDIQQTSLIKGDARSQSIAAASIIAKVYRDHLMVELDKQYPGYGFGKNA    223
A_suicloacalis      ELLLVDAVQLQTPIPQETIIKGDLKSNSIAAASIIAKVTRDEMMVEYGREFPGYGFANNA    222
T_halophilus        DHLLIDAMSIDINIPQEKIIKGDARSVSIAAASIIAKVYRDQLMKEYHKIHPHYAFDKNA    224
                      ::::.  :        :::**     *.:***:*** **.:*      .* :   .:

T_maritima          GYATKEHLNEIRKNGVLPIHRLSFEPVLELLTDDLLREFFEKGLISENRFERILN-LLGA    219
F_pennivorans       GYPTQEHLELLRKYGPTPFHRLSFEPVINLVSKELLDDWLDRRLITEQR-YRHLLNLLEV    232
P_elfii             GYGTVHHLNAIREFGPTVFHRLSFSPVLSNLSVKKVHDLFSENINCE-R-AKVILRKLSS    223
P_lettingae         GYGTVHHLNAIREFGPTVFHRLSFSPVLSNLSVKKVHDLFSENINCE-R-AKVILRKLSS    223
M_infera            GYCTEMHLNALREFGPTTWHRLTYRPIRELIPKELVSRWSEENEVSNARLFRAGLATLEV    250
A_tabaci            GYGTQEHLNGLKNQGPSPIHRYSFSPVTQYKK----------------------------    260
B_cereus            GYGTKQHLEAIEVHGVLEEHRKSFAPIKDMIQK---------------------------    257
E_cecorum           GYGTKEHLEGLEKYGVTPIHRKTFAPIKDMI-----------------------------    254
A_suicloacalis      GYGTKEHLEAMKRLGITPIHRRSFSPVKKYI-----------------------------    253
T_halophilus        GYGTKAHLTGLKEYGITAIHRKSYAPIKKYL-----------------------------    255
                    ** *  **  :.  *    ** :: *: .

T_maritima          RKS-------------      222
F_pennivorans       DLFGNTRRSKRKTRVK      248
P_elfii             S---------------      224
P_lettingae         S---------------      224
M_infera            KN--------------      252
A_tabaci            ----------------      260
B_cereus            ----------------      257
```

```
E_cecorum          ----------------      254
A_suicloacalis     ----------------      253
T_halophilus       ----------------      255
```

6) Multiple sequence alignment of MutH

Legend

```
>H_influenzae      >2AOR:B|PDBID|CHAIN|SEQUENCE
>H_haemolyticus    >gi|822511222|ref|WP_046942105.1|
>S_pneumoniae      >gi|987289758|emb|CVP28239.1|
>P_pneumotropica   >gi|980941964|ref|WP_059365670.1|
>A_aphrophilus     >gi|491980978|ref|WP_005704304.1|
>P_dagmatis        >gi|492155710|ref|WP_005765349.1|
>A_paragallinarum  >gi|737717974|ref|WP_035686820.1|
>Mannheimia_sp_MG13 >gi|764739431|ref|WP_044470572.1|
>P_multocida       >gi|1124583343|ref|WP_074865438.1|
>M_haemolytica     >gi|544865651|ref|WP_021279504.1|
>G_anatis          >gi|917516715|ref|WP_052123132.1|
```

Alignment

```
H_influenzae             ----MIPQTLEQLLSQAQSIAGLTFGELADELHIPVPIDLKRDKGWVGMLLERALGATAG 56
H_haemolyticus           ---MI-PQTLEQLLSQAQSIAGLTFGELADELHIPVPPDLKRDKGWVGMLLERALGATAG 56
S_pneumoniae             ---MI-PQTLEQLLSQAQSIAGLTFGELADELHIPVPIDLKRDKGWVGMLLERALGATAG 56
P_pneumotropica          ---MM-PQTLTQLLQRAQSIAGLTFGELADELTIPVPPNLKRDKGWVGMLLETALGATAG 56
A_aphrophilus            ---MI--QTEQQLLARAQSIAGMTFGELAQQLQIPVPPNLKRDKGWVGILIEMALGATAG 55
P_dagmatis               ---MI-PRTEQELLQKAQNIAGLRLAELAEELYIPIPSDLKRNKGWVGMLIETALGAAAG 56
A_paragallinarum         ---MNSPKTEQDLLALAQSLAGLTFGELAQDLSLPVPPNLKRDKGWVGTLLETALGATAG 57
Mannheimia_sp_MG13       --MITSPQSKQQLLHRARTIAGLSFGELAEELNIIVPPDLKRDKGWVGQLIETALGAKAG 58
P_multocida              ---MT-PKTEQELLQRAQAIAGLRFAELAQSLHMLVPPDLKRDKGWVGMLIETALGATAG 56
M_haemolytica            MQLSTFSPTEQALLSKAEWLAGFTLGEIAEMLHIPIPADLKRDKGWVGTLIETALGAKAG 60
G_anatis                 -------------MQRAYQLAGRTFVDIAEQLHIVVPQDLKRDKGWVGNLIECALGASAG 47
                                      :  *  :**  : ::*: * : :* :***:***** *:* **** **

H_influenzae             SKAEQDFSHLGVELKTLPINAEGYPLETTFVSLAPLVQNSGVKWENSHVRHKLSCVLWMP 116
H_haemolyticus           SKAEQDFSHLGVELKTLPINAEGYPLETTFVSLAPLVQNSGVKWENSHVRHKLSCVLWMP 116
S_pneumoniae             SKAEQDFSHLGVELKTLPINAEGYPLETTFVSLAPLVQNSGVKWENSHVRHKLSCVLWMP 116
P_pneumotropica          SKAEQDFAHLGVELKTLPINEQGFPLETTFVSLAPLTQNSGVQWENSHVRHKLSCVLWIP 116
A_aphrophilus            SKAEQDFEHLGIELKTIPINAQGFPLETTFVSLAPLIQNSGVNWQNSHVRHKLSKVLWIP 115
P_dagmatis               SKAERDFAHLGIELKTLPINSKGYPLETTFVSLAPLIQNTGVTWQNSHVKHKLSRVLWIP 116
A_paragallinarum         SKAEQDFAHLGIELKTIPIDSQGKPLETTFVSLAPLIHNSGITWQSSHVKHKLSKVLWIP 117
Mannheimia_sp_MG13       SKPEQDFAHLGIELKTIPINSQGYPLETTFVSLAPLIQTAGVNWQNSHLRYKLSQVLWIP 118
P_multocida              SKAEQDFAHLGIELKTLPINAQGMPLETTFVSLAPLTQNVGVSWENSHVRHKLSKVLWIL 116
M_haemolytica            SKAEQDFAHLGIELKTIPVNQKGLPLETTFVSLAPLTQNNGITWETSHVKHKLSRVLWIP 120
G_anatis                 SKPEQDFAHLGIELKTVPINRFGEPLETTFVSLAPLIDNNGIVWECSHVRYKLSKVLWIP 107
                         ** *:** ***:****:*::  * ************ .. *: *: **:::*** ***:

H_influenzae             IEGSRHIPLRERHIGAPIFWKPTAEQERQLKQDWEELMDLIVLGKLDQITARIGEVMQLR 176
H_haemolyticus           IEGSRHIPLRERHIGAPILWKPTAEQERQLKQDWEELMDLIVLGKLDQITARIGEVMQLR 176
S_pneumoniae             IEGSRHIPLRERHIGAPIFWKPTTGQERQLKQDWEELMDLIVLGKLEQITARIGEVMQLR 176
P_pneumotropica          IEGSRHIALRERHIGTPILWQPSSEQEQQLKQDWEELMEYITMGRLDEITARIGEVMQLR 176
A_aphrophilus            IEGERHIPLAERHIGAPILWQPSPHQEARLRQDWEELMDYIVLGKLDQITARLGDVLQLR 175
P_dagmatis               VEGERQIPLVDRHIGQGILWQPTAEQEWRLQRDWEELMEYITLGKLDQITARLGEVLQLR 176
A_paragallinarum         IEGERHIPLAQRHIGQPILWQPSREQEQRLQQDWEELMEYIVFGRLDEITARLGEVLQLR 177
Mannheimia_sp_MG13       IQGERNIPLASRLIGSPILWQPNAEQEAQLQQDWEELMDYIVLGKVHLITAKIGKVLQLR 178
P_multocida              VEGERQIPLSERRVGQPILWQPSAQQELRLKRDWEELMEYISLGKLEQINATLGEVLQLR 176
M_haemolytica            VEGERKIPLAERHIGQPILWSPTITQESRLQQDWEELMELIVLGELHKINATLGEVLQLR 180
G_anatis                 IEGERTIPLAQRRVGQPILWQPSAEEEAQLKQDWEELMDLIVLGEVKKIDARIGEVMQLR 167
                         ::*.* * * .* :*  *:*.*.  :* :*::******: * :*.:. * * :*.*:***

H_influenzae             PKGANSRAVTKGIGKNGEIIDTLPLGFYLRKEFTAQILNAFLETKSL---- 223
H_haemolyticus           PKGANSRAVTKGIGKNGEIIDTLPLGFYLRKEFTAQILNAFLETKSL---- 223
S_pneumoniae             PKGVNSRAVTKGIGKNGEIIDTLPLGFYLRKEFTAQILNAFLETKSL---- 223
P_pneumotropica          PKGANSKAITKGIGKNGEVIDTLPLGFYLRKEFTAGILREFLNDEFFRW-- 225
A_aphrophilus            PKGANSKALTKGIGKNGEIIDTLPLGFYLRKAFTHEILQQFIQQTI----- 221
P_dagmatis               PKGANSRSLTKGIGKKGEIIDTLPLGFYLRKEFTYEILQNFLTN------- 220
A_paragallinarum         PKGANSKALTKGIGRNGEIIDTLPLGFYLRKNFTHEILQQFQQPR------ 222
Mannheimia_sp_MG13       PKGANSRAKTKGIGPQGEVIDTLPLGFYLRKEFTATILQNFLQNR------ 223
P_multocida              PKGANSKALTRGIGKHGEMIDTLPLGFYLRKTFTAEILQQFLLGTG----- 222
M_haemolytica            PKGRNNRSITSAINAKGEIVQSIPLGFYLRKHFTAEILQNFLHCPL----- 226
G_anatis                 PKGNNNRALTKAIGQQGQVIDTLPLGFYLRKNFTAKILRNFQNNGYSLSPR 218
                         *** *.:: * .*. :*::::::******** **  **. *
```

7) Multiple sequence alignment of Exo-λ

Legend

```
>Escherichia_virus_lambda     >4WUZ:A|PDBID|CHAIN|SEQUENCE
>Enterobacteria_phage_HK630   >gi|428782822|ref|YP_007112573.1|
>Stx2-converting_phage_1717   >gi|209447136|ref|YP_002274221.1|
>Achromobacter_sp_ATCC35328   >gi|928599693|emb|CUK07497.1|
>Cronobacter_phage_ENT47670   >gi|431810511|ref|YP_007237585.1|
>S_pneumoniae                 >gi|1062527211|ref|WP_069289300.1|
>Vibrio_phage_2E1             >gi|1067529875|gb|AOQ26699.1|
>P_multocida                  >gi|1096225174|ref|WP_071171091.1|
>H_parasuis                   >gi|737515692|ref|WP_035494618.1|
>H_influenzae                 >gi|491918549|ref|WP_005671505.1|
>G_anatis                     >gi|746077620|ref|WP_039145071.1|
```

Alignment

```
Escherichia_virus_lambda        GSHMTPDIILQRTGIDVRAVEQGDDAWHKLRLGVITASEVHNVIAKPRS----------- 49
Enterobacteria_phage_HK630      ---MTPDIILQRTGIDVRAVEQGEDAWHKLRLGVITASEVHNVIAKPRS----------- 46
Stx2-converting_phage_1717      ---MTPDIILQRTGIDVRAVEQGDDAWHKLRLGVITASEVHNVIAKPRS----------- 46
Achromobacter_sp_ATCC35328      ---MTPDIILQRTGIDVRAVEQGDDAWHKLRLGVITASEVHNVIAKPRS----------- 46
Cronobacter_phage_ENT47670      ---MTPAIILERTGIDVLTVEQGDEAWQRLRLGVITASDVHNVISKPRS----------- 46
S_pneumoniae                    ---MTPDIILQRTGIDVRAVEQGDDAWHKLRLGVITASEVHNVIAKPRS----------- 46
Vibrio_phage_2E1                -----MINVNHITGVNTFNIEQGTEEWLRHRAGTITASRAHLVIADDITPPMPDDVEIIP 55
P_multocida                     ----------MIDNLITLDCEQGTEEWLVARLGIPTATGIKNIVTPSG------------ 38
H_parasuis                      ---------------MTLNCEQGTEEWLTARLGIPTATGVSNIVTPSG------------ 33
H_influenzae                    ----------MLDKLITLDCEQGTEEWLAARCGIPTATGISNIVTPTG------------ 38
G_anatis                        ----------MIDGLITLDCEQGTEEWLVARLGIPTATGIKNIVNNSG------------ 38
                                                .   *** : *   * *  **:    ::

Escherichia_virus_lambda        -------------------------------GKKWPDMKMSYFHTLLAEVCTGVAPEVN- 77
Enterobacteria_phage_HK630      -------------------------------GKKWPDMKMSYFHTLLAEVCTGVAPEVN- 74
Stx2-converting_phage_1717      -------------------------------GKKWPDMKMSYFHTLLAEVCTGVAPEVN- 74
Achromobacter_sp_ATCC35328      -------------------------------GKKWPDMKMSYFHTLLAEVCTGVAPEVN- 74
Cronobacter_phage_ENT47670      -------------------------------GKKWPDMKMSYFHTLLAEVCTGVAPEVN- 74
S_pneumoniae                    -------------------------------GKKWPDMKMSYFHTLLAEVCTGVAPEVN- 74
Vibrio_phage_2E1                TEKRGVNDVSYNGESFQGTKANCEKWVRSKLEPVMPDGKKSYMLELIAQIATGNVPESAS 115
P_multocida                     ---------------------------------QKSGGWISYLAELVAESVEGVTEGFK- 64
H_parasuis                      ---------------------------------KKSSAWTSYLAELVAESIEGLTEGFK- 59
H_influenzae                    ---------------------------------KKSGSYLPYLAELIAESIEGLKENYK- 64
G_anatis                        ---------------------------------KKSSGWFTYLAELVAESIEGGNNIIK- 64
                                                                    .    *:  *:*:   *

Escherichia_virus_lambda        AKALAWGKQYENDARTLFEFTSGVNVTESPIIYRDESMRTACSPDGLCSD-GNGLELKCP 136
Enterobacteria_phage_HK630      AKALAWGKQYENDARTLFEFTSGVNVTESPIIYRDESMRTACSPDGLCSD-GNGLELKCP 133
Stx2-converting_phage_1717      AKALAWGKQYENDARTLFEFTSGVNVTESPIIYRDESMRTACSPDGLCSD-GNGLELKCP 133
Achromobacter_sp_ATCC35328      AKALAWGKQYENDARVLFEFTSGVNVTESPIIYRDESMRTACSPDGLCSD-GNGLELKCP 133
Cronobacter_phage_ENT47670      ARALAWGKQYEDDARALFEFTAGVQVTESPIIYKDETMRTACSPDGLCSD-GRGLELKCP 133
S_pneumoniae                    AKALAWGKQYENDARTLFEFTSGVNVTESPIIYRDESMRTACSPDGLCSD-GNGLELKCP 133
Vibrio_phage_2E1                FKQAEWGHLNEPLARDAFEAKNFCIVTEAGLIYKDESLRCAISPDGLLMDEKQGLEIKSP 175
P_multocida                     SQHMERGNELEPLARMAYEFETGHDVTQVGGVYLNEKKELMVSPDGLILSHQKGLEIKCP 124
H_parasuis                      STDMLRGNLLEEQARMAYEFATGNDVVQVGGVYRNADKDMMVSPDGLIPTLRKGLEIKCP 119
H_influenzae                    SEDMARGNELEPFARAAYEFETGNAVIQVGGVYLNADKDLMISPDGLIPNLRKGLEIKCP 124
G_anatis                        TVDMERGNELEPKARMAYEFLTDNTVVQVGGVYLNEQKELMISPDGLIPNLKKGLEIKCP 124
                                      *:  *  **  :*      * :   :* :       *****     .***:*.*

Escherichia_virus_lambda        FTSRDFMKFRLGGFEAIKSAYMAQVQYSMWVTRKNAWYFANYDPRMKR---EGLHYVVIE 193
Enterobacteria_phage_HK630      FTSRDFMKFRLGGFEAIKSAYMAQVQYSMWVTRKNAWYFANYDPRMKR---EGLHYVVIE 190
Stx2-converting_phage_1717      FTSRDFMKFRLGGFEAIKSAYMAQVQYSMWVTRKDAWYFANYDPRMKR---EGLHYVVVE 190
Achromobacter_sp_ATCC35328      FTSRDFMKFRLGGFEAIKSAYMAQVQYSMWVTRKDAWYFANYDPRMKR---EGLHYVVIE 190
Cronobacter_phage_ENT47670      FTSRDFMKFRLGGFEAIKSAYMAQVQFSMWVTGKDAWYFSNYDPRMRR---EGLHHVVVE 190
S_pneumoniae                    FTSRDFMKFRLGGFEAIKSAYMAQVQYSMWVTRKNAWYFANYDPRMKR---EGLHYVVIE 190
Vibrio_phage_2E1                YTTQVHLDTVLNG--KIKPEYLIQCQFSMWVTGWDKWHFCSYDHRLRGSSQNRLHTVVIE 233
P_multocida                     KM-KTHIKYILEG--GVPSEYIIQVQVAMWVTGYKSWDFVSYCPEYQK---QTLYLYTAT 178
H_parasuis                      KM-KTHIKYIIEG--VVPSEYIIQVQVALWVTGYDSWDFVSYCPEYQK---QTLFIHTEN 173
H_influenzae                    QI-KTHIKYLLQG--GVPQEYLIQVQSALWVTGYETWDFVSYCPEYYK---QPFYLFTAQ 178
G_anatis                        KM-KTHIKYLLES--GVPSEYLMQVQSALWVTGYETWDFVSYCPDYQK---QPLYLYTAE 178
                                   : .:.  : .   :   *: * * ::***  . * * .*         : :.  .

Escherichia_virus_lambda        RDEKYMASFDEIVPEFIEKMDEALAEIGFVFGEQWR--      229
Enterobacteria_phage_HK630      RDEKYMASFDEIVPEFIEKMDEALAEIGFVFGEQWR--      226
Stx2-converting_phage_1717      RDEKYMASFDEMVPEFIEKMDEALAEIGFVFGEQWR--      226
Achromobacter_sp_ATCC35328      RNEKYMASFDEMVPEFIEKMDEALAEIGFVFGEQWR--      226
Cronobacter_phage_ENT47670      RDEKYMEDFTEAVPEFIEKMDMALAEIGFTFGEQWR--      226
S_pneumoniae                    RDEKY---------------------------------      195
Vibrio_phage_2E1                RDESIMAKFDKYIPKFIDEMDRQLKKLNFEFNDQWREF      271
P_multocida                     RDEMLMKAFDEYIPQFLKSLRA------LRDG------      204
H_parasuis                      PDPVLMKAFDKYIPQFIETLKA------LKVN------      199
H_influenzae                    RDPNLMKSFDRLIPEFIKTLKA------YKSTE-----      205
G_anatis                        RDPILMKAFDKYIPEFLNALKA------LKGENQWQA-      209
```

:

8) Structure-based sequence alignment of DNA polymerases

Legend

```
>H_sapiens_Pol-eta    >4ECS:A|PDBID|CHAIN|SEQUENCE
>S_cerevisiae_Pol-eta >3MFI:A|PDBID|CHAIN|SEQUENCE
>H_sapiens_Pol-iota   >2ALZ:A|PDBID|CHAIN|SEQUENCE
>H_sapiens_Pol-kappa  >1T94:A|PDBID|CHAIN|SEQUENCE
>S_solfataricus_DPO4  >2AGO:A|PDBID|CHAIN|SEQUENCE
>S_solfataricus_DBH   >1IM4:A|PDBID|CHAIN|SEQUENCE
>E_coli_DinB          >4IRK:A|PDBID|CHAIN|SEQUENCE
```

Alignment (extract)

```
H_sapiens_Pol-eta     ----------------------------------------DLQLTVGAVIVEEMRAAIER 209
S_cerevisiae_Pol-eta  ----------------------------------------DVILALGSQVCKGIRDSIKD 261
H_sapiens_Pol-iota    ----------------------------------------HIRLLVGSQIAAEMREAMYN 165
H_sapiens_Pol-kappa   HERSISPLLFEESPSDVQPPGDPFQVNFEEQNNPQILQNSVVFGTSAQEVVKEIRFRIEQ 236
S_solfataricus_DPO4   ------------------------------------------DYREAYNLGLEIKNKILE 134
S_solfataricus_DBH    -----------------------------------------GNFENGIELARKIKQEILE 140
E_coli_DinB           --------------------------------------------GSATLIAQEIRQTIFN 133
                                                                    .  :   ::  :

H_sapiens_Pol-eta     ETGFQCSAGISHNKVLAKLACGLNKPNRQTLVSHG--SVPQLFSQ--MP-IRKIRSLGGK 264
S_cerevisiae_Pol-eta  ILGYTTSCGLSSTKNVCKLASNYKKPDAQTIVKND--CLLDFLDCGKF-EITSFWTLGGV 318
H_sapiens_Pol-iota    QLGLTGCAGVASNKLLAKLVSGVFKPNQQTVLLPE--SCQHLIHS--LNHIKEIPGIGYK 221
H_sapiens_Pol-kappa   KTTLTASAGIAPNTMLAKVCSDKNKPNGQYQILPNRQAVMDFIKD--L-PIRKVSGIGKV 293
S_solfataricus_DPO4   KEKITVTVGISKNKVFAKIAADMAKPNGIKVIDDE--EVKRLIRE--L-DIADVPGIGNI 189
S_solfataricus_DBH    KEKITVTVGVAPNKILAKIIADKSKPNGLGVIRPT--EVQDFLNE--L-DIDEIPGIGSV 195
E_coli_DinB           ELQLTASAGVAPVKFLAKIASDMNKPNGQFVITP--AEVPAFLQT--L-PLAKIPGVGKV 188
                              *::  . ..*: ..  **:    :         ::    :  : ..  :*

H_sapiens_Pol-eta     LGASVIEILG---------------------------IEYMGELTQFTESQLQSHFGEKN 297
S_cerevisiae_Pol-eta  LGKELID-VLDLPHENSIKHIRETWPDNAGQLKEFLDAKVKQSDYDRSTSNIDPLKTADL 377
H_sapiens_Pol-iota    TAKCLEA-LG---------------------------INSVRDLQTFSPKILEKELGISV 253
H_sapiens_Pol-kappa   TEKMLKA-LG---------------------------IITCTELYQ-QRALLSLLFSETS 324
S_solfataricus_DPO4   TAEKLKK-LG---------------------------INKLVDTLSIEFDKLKGMIGEAK 221
S_solfataricus_DBH    LARRLNE-LG---------------------------IQKLRDILSKNYNELEK------ 221
E_coli_DinB           SAAKLEA-MG---------------------------LRTCGDVQACDLVMLLKRF-GKF 219
                          :   :                                 .        :
```

9) Multiple sequence alignment of RNA polymerase Pol-II Rbp1 subunit and structure-based sequence alignment of RPA190, Rbp1, and Rpc1 subunits of RNA polymerases I, II, and III.

Legend

```
>S_cerevisiae_PolII-Rbp1             >2E2H:A|PDBID|CHAIN|SEQUENCE
>S_arboricola_H6_PolII-Rbp1          >gi|401626472|gb|EJS44418.1|
>C_glabrata_CBS138_PolII-Rbp1        >gi|50289967|ref|XP_447415.1|
>H_valbyensis_NRRL-Y-1626_PolII-Rbp1 >gi|1037356197|gb|OBA26116.1|
>Y_lipolytica_PolII-Rbp1             >XP_501909.2 YALI0C16566p
>A_rubescens_DSM1968_PolII-Rbp1      >ODV60282.1
>O_parapolymorpha_DL-1_PolII-Rbp1    >XP_013932372.1
>R_norvegicus_PolII-Rbp1             >gi|109488292|ref|XP_343923.3|
>H_sapiens_PolII-Rbp1                >gi|4505939|ref|NP_000928.1|
>B_taurus_PolII-Rbp1                 >gi|329663165|ref|NP_001193242.1|
>O_garnettii_PolII-Rbp1              >gi|395836506|ref|XP_003791195.1|
>M_mulatta_PolII-Rbp1                >gi|355568192|gb|EHH24473.1|
>M_musculus_PolII-Rbp1               >gi|200794|gb|AAA40071.1|
>S_cerevisiae_PolI-RPA190            >5M5Y:A|PDBID|CHAIN|SEQUENCE
>S_cerevisiae_PolIII-Rpc1            >5FJ8:A|PDBID|CHAIN|SEQUENCE
```

Multiple sequence alignment (extract)

```
S_cerevisiae_PolII-Rbp1                  TMREITETIAEAKKKVLDVTKEAQANLLTAKHGMTLRESFEDNVVRFLNEARDKAGRLAE
        734
S_arboricola_H6_PolII-Rbp1               TMREITETIAEAKKKVLDVTKEAQANLLTAKHGMTLRESFEDNVVRFLNEARDKAGRLAE
        734
C_glabrata_CBS138_PolII-Rbp1             TMREISETIAEAKQKVEAVTKEAQANLLTAKHGMTLRESFEDNVVRFLNEARDRAGRLAE
        734
H_valbyensis_NRRL-Y-1626_PolII-Rbp1      TMLEITEAIAIAKVKVEEVTKEAQENLLSAKHGMTLRESFEDNVVRFLNEARDKAGRSAE
        738
Y_lipolytica_PolII-Rbp1                  TMRDVTETIEEAKKKVKEIILEAHANTLTAEAGMTMRESFEHNVSRVLNQARDTAGRSAE
        739
A_rubescens_DSM1968_PolII-Rbp1           TMKKITTTIAEAKQKVQDLILDAQANRLELEPGMTLRESFESKVSRVLNQARDDAGHCAQ
        732
O_parapolymorpha_DL-1_PolII-Rbp1         TMKTITETIAIAKEKVQEVIMDAQKNLLEAEPGMTVRESFEQKVSKLLNEARDSAGKSAE
        694
R_norvegicus_PolII-Rbp1                  TYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGNTLRQTFENQVNRILNDARDKTGSSAQ
        757
H_sapiens_PolII-Rbp1                     TYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGNTLRQTFENQVNRILNDARDKTGSSAQ
        757
B_taurus_PolII-Rbp1                      TYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGNTLRQTFENQVNRILNDARDKTGSSAQ
        757
O_garnettii_PolII-Rbp1                   TYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGNTLRQTFENQVNRILNDARDKTGSSAQ
        757
M_mulatta_PolII-Rbp1                     TYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGNTLRQTFENQVNRILNDARDKTGSSAQ
        767
M_musculus_PolII-Rbp1                    TYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGNTLRQTFENQVNRILNDARDKTGSSAQ
        757
                                         *   :  :*  ** .*  :  .*: * *    * *:*::** :* :.**:*** :*  *:

S_cerevisiae_PolII-Rbp1                  VNLKDLNNVKQMVMAGSKGSFINIAQMSACVGQQSVEGKRIAFGFVDRTLPHFSKDDYSP
        794
S_arboricola_H6_PolII-Rbp1               VNLKDLNNVKQMVMAGSKGSFINIAQMSACVGQQSVEGKRIAFGFVDRTLPHFSKDDYSP
        794
C_glabrata_CBS138_PolII-Rbp1             MNLKDLNNVKQMVSAGSKGSFINIAQMSACVGQQSVEGKRIGFGFVDRTLPHFSKDDYSP
        794
H_valbyensis_NRRL-Y-1626_PolII-Rbp1      VNLKSLNNVKQMVSSGSKGSFINIAQMSACVGQQSVEGKRIPFGFADRTLPHFSKDDYSP
        798
Y_lipolytica_PolII-Rbp1                  MSLKDLNNVKQMVVAGSKGSFINISQMSACVGQQMVEGKRVPFGFADRTLPHFCKDDYSP
        799
A_rubescens_DSM1968_PolII-Rbp1           MNLKELNNVKQMVVSGSKGSFINISQMSACVGQQIVEGKRIPFGFADRTLPHFTKDDFSP
        792
O_parapolymorpha_DL-1_PolII-Rbp1         TSLKDSNNVKQMVTAGSKGSYINISQMSACVGQQIVEGKRINFGFADRSLPHFTKDDYSA
        754
R_norvegicus_PolII-Rbp1                  KSLSEYNNFKSMVVSGAKGSKINISQVIAVVGQQNVEGKRIPFGFKHRTLPHFIKDDYGP
        817
H_sapiens_PolII-Rbp1                     KSLSEYNNFKSMVVSGAKGSKINISQVIAVVGQQNVEGKRIPFGFKHRTLPHFIKDDYGP
        817
B_taurus_PolII-Rbp1                      KSLSEYNNFKSMVVSGAKGSKINISQVIAVVGQQNVEGKRIPFGFKHRTLPHFIKDDYGP
        817
O_garnettii_PolII-Rbp1                   KSLSEYNNFKSMVVSGAKGSKINISQVIAVVGQQNVEGKRIPFGFKHRTLPHFIKDDYGP
        817
M_mulatta_PolII-Rbp1                     KSLSEYNNFKSMVVSGAKGSKINISQVIAVVGQQNVEGKRIPFGFKHRTLPHFIKDDYGP
        827
M_musculus_PolII-Rbp1                    KSLSEYNNFKSMVVSGAKGSKINISQVIAVVGQQNVEGKRIPFGFKHRTLPHFIKDDYGP
        817
                                          .*..  **.*.** :*:*** ***:*: * **** *****: *** .*:**** ***:.
```

```
S_cerevisiae_PolII-Rbp1                  ESKGFVENSYLRGLTPQEFFFHAMGGREGLIDTAVKTAETGYIQRRLVKALEDIMVHYDN
        854
S_arboricola_H6_PolII-Rbp1               ESKGFVENSYLRGLTPQEFFFHAMGGREGLIDTAVKTAETGYIQRRLVKALEDIMVHYDN
        854
C_glabrata_CBS138_PolII-Rbp1             ESKGFVENSYLRGLTPQEFFFHAMGGREGLIDTAVKTAETGYIQRRLVKALEDIMVHYDG
        854
H_valbyensis_NRRL-Y-1626_PolII-Rbp1      ESKGFVENSYLRGLTPQEFFFHAMGGREGLIDTAVKTAETGYIQRRLVKALEDIMVHYDG
        858
Y_lipolytica_PolII-Rbp1                  ESKGFIENSYLRGLTPQEFFFHAMAGREGLIDTAVKTAETGYIQRRLVKALEDVMVQYDG
        859
A_rubescens_DSM1968_PolII-Rbp1           ESKGFVENSYLRGLTPQEFFFHAMAGREGLIDTAVKTAETGYIQRRLVKALEDIMVHYDG
        852
O_parapolymorpha_DL-1_PolII-Rbp1         ESKGFVENSYLRGLTPQEFFFHAMAGREGLIDTAVKTAETGYIQRRLLKALEDIMVHYDG
        814
R_norvegicus_PolII-Rbp1                  ESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYIQRRLIKSMESVMVKYDA
        877
H_sapiens_PolII-Rbp1                     ESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYIQRRLIKSMESVMVKYDA
        877
B_taurus_PolII-Rbp1                      ESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYIQRRLIKSMESVMVKYDA
        877
O_garnettii_PolII-Rbp1                   ESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYIQRRLIKSMESVMVKYDA
        877
M_mulatta_PolII-Rbp1                     ESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYIQRRLIKSMESVMVKYDA
        887
M_musculus_PolII-Rbp1                    ESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYIQRRLIKSMESVMVKYDA
        877
                                         **:**:***** **** *******.**********************:*::*.:**:**
```

Structure-based alignment (extract)

```
S_cerevisiae_PolII-Rbp1             AEAKKKVLDVTKEAQANL-------LTAKHGMTLRESFEDNVVRFLNEAR 726
S_cerevisiae_PolI-RPA190            REAAAEVTNLDKDTPADDPELLKRLQEILRDNNKSGILDAVTSSKVNAIT 904
S_cerevisiae_PolIII-Rpc1            EIAYHKCDELITLFNKGE-------LETQPGCNEEQTLEAKIGGLLSKVR 774
                                      *  :  :: .    .             . .    ::      :.

S_cerevisiae_PolII-Rbp1             DKAGRLAEV----NLKDLNNVKQMVMAGSKGSFINIAQMSACVGQQSVEG 772
S_cerevisiae_PolI-RPA190            SQVVSKCVPDGTMKKFPCNSMQAMALSGAKGSNVNVSQIMCLLGQQALEG 954
S_cerevisiae_PolIII-Rpc1            EEVGDVCIN----ELDNWNAPLIMATCGSKGSTLNVSQMVAVVGQQIISG 820
                                    .:.   .      :    *    *. .*:*** :*::*: . :*** :.*

S_cerevisiae_PolII-Rbp1             KRIAFGFVDRTLPHFSKDDYSPESKGFVENSYLRGLTPQEFFFHAMGGRE 822
S_cerevisiae_PolI-RPA190            RRVPVMVSGKTLPSFKPYETDAMAGGYVKGRFYSGIKPQEYYFHCMAGRE 1004
S_cerevisiae_PolIII-Rpc1            NRVPDGFQDRSLPHFPKNSKTPQSKGFVRNSFFSGLSPPEFLFHAISGRE 870
                                    .*:.  . .::** *   .  . : *:*.. :  *:.* *: **.:.***
```

10) Multiple sequence alignment of RNA polymerase Pol-II Rbp2 subunit and structure-based sequence alignment of RPA135, Rbp2, and Rpc2 subunits of RNA polymerases I, II, and III

Legend

```
>S_cerevisiae_PolII-Rbp2             >2E2H:B|PDBID|CHAIN|SEQUENCE
>S_arboricola_H6_PolII-Rbp2          >gi|401623583|gb|EJS41677.1|
>C_glabrata_CBS138_PolII-Rbp2        >gi|50293059|ref|XP_448959.1|
>H_valbyensis_NRRL-Y-1626_PolII-Rbp2 >gi|1037356434|gb|OBA26349.1|
>Y_lipolytica_PolII-Rbp2             >gi|50549809|ref|XP_502376.1|
>A_rubescens_DSM1968_PolII-Rbp2      >gi|1064972083|gb|ODV62510.1|
>O_parapolymorpha_DL-1_PolII-Rbp2    >gi|927376581|ref|XP_013934141.1|
>R_norvegicus_PolII-Rbp2             >gi|300798312|ref|NP_001178807.1|
>H_sapiens_PolII-Rbp2                >gi|4505941|ref|NP_000929.1|
>B_taurus_PolII-Rbp2                 >gi|154707850|ref|NP_001092552.1|
>O_garnettii_PolII-Rbp2              >gi|395819924|ref|XP_003783328.1|
>M_mulatta_PolII-Rbp2                >gi|355687402|gb|EHH25986.1|
>M_musculus_PolII-Rbp2               >gi|226958589|ref|NP_722493.2|
>S_cerevisiae_PolI-RPA135            >5M5Y:B|PDBID|CHAIN|SEQUENCE
>S_cerevisiae_PolIII-Rpc2            >5FJ8:B|PDBID|CHAIN|SEQUENCE
```

Multiple sequence alignment (extract)

```
S_cerevisiae_PolII-Rbp2                  GKTTPISPDEEELGQRT-AYHSKRDASTPLRSTENGIVDQVLVTTNQDGLKFVKVRVRTT
        971
S_arboricola_H6_PolII-Rbp2               GKTTPISPDEEELGQRT-AYHSKRDASTPLRSTENGIVDQVLVTTNQDGLKFVKVRVRTT
        971
C_glabrata_CBS138_PolII-Rbp2             GKTTPIAPDEEELGQRT-AYHSKRDASTPLRSTENGIVDQVLITTNQDGLKFVKVRVRTT
        970
H_valbyensis_NRRL-Y-1626_PolII-Rbp2      GKTTPIKPDNEELGLRT-AFHTKRDASTPLRSTESGIIDQVLITTNEEGLRFVKVRVRTT
        1016
Y_lipolytica_PolII-Rbp2                  GKTAPIPPDAEELGQRT-KYHTKRDASTPLRSTENGIVDQVLLTTNQEGLRFVKVRMRTT
        973
```

```
A_rubescens_DSM1968_PolII-Rbp2           GKTTPISPDNEELGRKT-QFHTKRDASMPLRSTENGIVDQVVLTTNHEGLKFVKVRMRTT
        966
O_parapolymorpha_DL-1_PolII-Rbp2         GKTVPIPPDTEELGQRT-KYHTKRDASTPLRTTESGIVDQVLLTTNAEGLKFAKVRMRTT
        1020
R_norvegicus_PolII-Rbp2                  NKSMPTVTQIPLEGSNVPQQPQYKDVPITYKGATDSYIEKVMISSNAEDAFLIKMLLRQT
        888
H_sapiens_PolII-Rbp2                     GKTVTLPENEDELESTN-RRYTKRDCSTFLRTSETGIVDQVMVTLNQEGYKFCKIRVRSV
        926
B_taurus_PolII-Rbp2                      GKTVTLPENEDELEGTN-RRYTKRDCSTFLRTSETGIVDQVMVTLNQEGYKFCKIRVRSV
        926
O_garnettii_PolII-Rbp2                   NKSMPTVTQIPLEGSNVPQQPQYKDVPITYKGATDSYIEKVMISSNAEDAFLIKMLLRQT
        888
M_mulatta_PolII-Rbp2                     GKTVTLPENEDELESTN-RRYTKRDCSTFLRTSETGIVDQVMVTLNQEGYKFCKIRVRSV
        926
M_musculus_PolII-Rbp2                    GKTVTLPENEDELESTN-RRYTKRDCSTFLRTSETGIVDQVMVTLNQEGYKFCKIRVRSV
        926
                                         .*:     :              :*     : :  . :::*::: * :.  : *: :* .

S_cerevisiae_PolII-Rbp2                  KIPQIGDKFASRHGQKGTIGITYRREDMPFTAEGIVPDLIINPHAIPSRMTVAHLIECLL
        1031
S_arboricola_H6_PolII-Rbp2               KVPQIGDKFASRHGQKGTIGITYRREDMPFTAEGIVPDLIINPHAIPSRMTVAHLIECLL
        1031
C_glabrata_CBS138_PolII-Rbp2             KVPQIGDKFASRHGQKGTIGITYRREDMPFTAEGIVPDLIINPHAIPSRMTVAHLIECLL
        1030
H_valbyensis_NRRL-Y-1626_PolII-Rbp2      KVPQIGDKFASRHGQKGTIGITYRREDMPFTAEGIVPDLIINPHAIPSRMTVAHLIECLL
        1076
Y_lipolytica_PolII-Rbp2                  KIPQIGDKFASRHGQKGTIGVTYRHEDMPFSAEGVVPDIIINPHAIPSRMTVAHLIECLL
        1033
A_rubescens_DSM1968_PolII-Rbp2           KVPQIGDKFASRHGQKGTIGITYRHEDMPFSREGIVPDLIINPHAIPSRMTVAHLIECLL
        1026
O_parapolymorpha_DL-1_PolII-Rbp2         KVPQIGDKFASRHGQKGTIGITYRHEDMPFTAQGIVPDLIINPHAIPSRMTVAHLIECLL
        1080
R_norvegicus_PolII-Rbp2                  RRPEIGDKFSSRHGQKGVCGLIVPQEDMPFCDSGICPDIIMNPHGFPSRMTVGKLIELLA
        948
H_sapiens_PolII-Rbp2                     RIPQIGDKFASRHGQKGTCGIQYRQEDMPFTCEGITPDIIINPHAIPSRMTIGHLIECLQ
        986
B_taurus_PolII-Rbp2                      RIPQIGDKFASRHGQKGTCGIQYRQEDMPFTCEGITPDIIINPHAIPSRMTIGHLIECLQ
        986
O_garnettii_PolII-Rbp2                   RRPEIGDKFSSRHGQKGVCGLIVPQEDMPFCDSGICPDIIMNPHGFPSRMTVGKLIELLA
        948
M_mulatta_PolII-Rbp2                     RIPQIGDKFASRHGQKGTCGIQYRQEDMPFTCEGITPDIIINPHAIPSRMTIGHLIECLQ
        986
M_musculus_PolII-Rbp2                    RIPQIGDKFASRHGQKGTCGIQYRQEDMPFTCEGITPDIIINPHAIPSRMTIGHLIECLQ
        986
                                         : *:*****:*******. *:   :*****  .*: **:*:***.:*****:.:*** *

S_cerevisiae_PolII-Rbp2                  SKVAALSGNEGDASPFT-DITVEGISKLLREHGYQSRGFEVMYNGHTGKKLMAQIFFGPT
        1090
S_arboricola_H6_PolII-Rbp2               SKVAALSGNEGDASPFT-DITVEGISKLLREHGYQSRGFEVMYNGHTGKKLMAQIFFGPT
        1090
C_glabrata_CBS138_PolII-Rbp2             SKVAALSGNEGDASPFT-DITVEGISKLLREHGYQSRGFEVMYNGHTGKKLMAQIFFGPT
        1089
H_valbyensis_NRRL-Y-1626_PolII-Rbp2      SKVAAINGQEGDASPFV-DVTVDSISDLLRQAGYQSRGFEVMYNGHTGKKLMAQIFFGPT
        1135
Y_lipolytica_PolII-Rbp2                  SKVSCLSGLEGDATPFT-DVTVDAISKLLRSHGYQSRGFEVMYHGHTGKKIMAQCFLGPT
        1092
A_rubescens_DSM1968_PolII-Rbp2           SKVCALMGTEGDATPFNDDITVDAISDILYTFGYQSRGFEVLYNGHTGKKLMAQVFFGPT
        1086
O_parapolymorpha_DL-1_PolII-Rbp2         SKVASMRGYEGDATPFT-DLTVDAVSKLLRENGYQSRGFEVMYNGHTGKKLMAQVFFGPT
        1139
R_norvegicus_PolII-Rbp2                  GKAGVLDGRFHYGTAFG-GSKVKDVCEDLVRHGYNYLGKDYVTSGITGEPLEAYIYFGPV
        1007
H_sapiens_PolII-Rbp2                     GKVSANKGEIGDATPFNDAVNVQKISNLLSDYGYHLRGNEVLYNGFTGRKITSQIFIGPT
        1046
B_taurus_PolII-Rbp2                      GKVSANKGEIGDATPFNDAVNVQKISNLLSDYGYHLRGNEVLYNGFTGRKITSQIFIGPT
        1046
O_garnettii_PolII-Rbp2                   GKAGVLDGRFHYGTAFG-GSKVKDVCEDLVRHGYNYLGKDYVTSGITGEPLEAYIYFGPV
        1007
M_mulatta_PolII-Rbp2                     GKVSANKGEIGDATPFNDAVNVQKISNLLSDYGYHLRGKEVLYNGFTGRKITSQIFIGPT
        1046
M_musculus_PolII-Rbp2                    GKVSANKGEIGDATPFNDAVNVQKISNLLSDYGYHLRGNEVLYNGFTGRKITSQIFIGPT
        1046
                                         .*.    *    .: *    .*. :.. *   **:  * : :  * **. : :  ::**.
```

Structure-based alignment (extract)

```
S_cerevisiae_PolII-Rbp2                 -PDEEELGQRTAYHSKRDASTPLRSTENGIVDQVLVTTN----QDGLKFV 964
S_cerevisiae_PolI-RPA135                --YFDDTLNKT-------KIKTYHSSEPAYIEEVNLIGDESNKFQELQTV 901
S_cerevisiae_PolIII-Rpc2                SADAPNPNNVNVQTQYREAPVIYRGPEPSHIDQVMMSVS----DNDQALI 896
                                             :  : .            :..* . :::* :  .     :    :
S_cerevisiae_PolII-Rbp2                 KVRVRTTKIPQIGDKFASRHGQKGTIGITYRREDMPFTAEGIVPDLIINP 1014
S_cerevisiae_PolI-RPA135                SIKYRIRRTPQIGDKFSSRHGQKGVCSRKWPTIDMPFSETGIQPDIIINP 951
S_cerevisiae_PolIII-Rpc2                KVLLRQNRRPELGDKFSSRHGQKGVCGIIVKQEDMPFNDQGIVPDIIMNP 946
                                        .:  *  : *::*****:*******. .      ****.  ** **:*:**
```

```
S_cerevisiae_PolII-Rbp2                 HAIPSRMTVAHLIECLLSKVAALSGNEGDASPFT----DITVEGISKLLR 1060
S_cerevisiae_PolI-RPA135                HAFPSRMTIGMFVESLAGKAGALHGIAQDSTPWIFNEDDTPADYFGEQLA 1001
S_cerevisiae_PolIII-Rpc2                HGFPSRMTVGKMIELISGKAGVLNGTLEYGTCFG----GSKLEDMSKILV 992
                                        *.:*****:. ::* : .*...* *    .: :     .   : :.: *

S_cerevisiae_PolII-Rbp2                 EHGYQSRGFEVMYNGHTGKKLMAQIFFGPTYYQRLRHMVDDKIHARARGP 1110
S_cerevisiae_PolI-RPA135                KAGYNYHGNEPMYSGATGEELRADIYVGVVYYQRLRHMVNDKFQVRSTGP 1051
S_cerevisiae_PolIII-Rpc2                DQGFNYSGKDMLYSGITGECLQAYIFFGPIYYQKLKHMVLDKMHARARGP 1042
                                        . *::  * : :*.* **: * * *:.*  ***:*:*** **::.*: **
```

11) Viral RdRps

Legend

>Norwalk	>3BSO:A|PDBID|CHAIN|SEQUENCE
>Sapporo	>2UUW:A|PDBID|CHAIN|SEQUENCE
>RHDV	>1KHV:A|PDBID|CHAIN|SEQUENCE
>Polio	>1RDR:A|PDBID|CHAIN|SEQUENCE
>FMDV	>1U09:A|PDBID|CHAIN|SEQUENCE

Structure-based alignment (extract)

```
Norwalk         WLLTLCALSEV-TNL----SPDIIQANSLFSFYGDDEIVST-DIK-LDPE 356
Sapporo         MIYVAAAILQAYESHNVPYTGNVFQVET-IHTYGGGCMYSVCPATASIFH 362
RHDV            WLLWSAAVYKSCAEIGLH-CSNLYEDAP-FYTYGDDGVYAMTPMMVSLLP 369
Polio           NLIIRTLLLKTYKGIDL-------DHLK-MIAYGDDVIASY-PHE-VDAS 341
FMDV            NIYVLYALRRHYEGVEL-------DTYT-MISYGDDIVVAS-DYD-LDFE 351
                 :     : .              :    :  **.. : :

Norwalk         KLTAKLKEYGLKPTRPDKTEGPLVISEDLNGLTFLRRTVTRDP---AGWF 403
Sapporo         TVLANLTSYGLKPTAADKSDAIK----PTNTPVFLKRTFTQTP---HGIR 405
RHDV            AIIENLRDYGLSPTAADKTEFIDV--CPLNKISFLKRTFELTD---IGWV 414
Polio           LLAQSGKDYGLTMTPADKSATFE--TVTWENVTFLKRFFRADEKYPFLIH 389
FMDV            ALKPHFKSLGQTITPADKSDKGFVLGHSITDVTFLKRHFHMDYGT--GFY 399
                 :     . * . * .**:              **:* .

Norwalk         GK-LEQSSILRQMYWTRGP-NHEDPSET-MIPHSQRPIQLMSLLGEAALH 450
Sapporo         AL-LDITSITRQFYWLKAN-RTSDPSSPPAFDRQARSAQLENALAYASQH 453
RHDV            SK-LDKSSILRQLEWSKTTSRHMVIEETYDLAKEERGVQLEELQVAAAAH 463
Polio           PV-MPMKEIHESIRWTKDP-RNT-------------QDHVRSLCLLAWHN 424
FMDV            KPVMASKTLEAILSFARRG-T-I-------------QEKLISVAGLAVHS 434
                   :  . :   : : :                     :: .    *
```

12) Multiple sequence alignment of spliceosomal Prp8 subunit

Legend

>S_cerevisiae	>CAA80854.1
>S_pombe	>NP_593861.1
>C_thermophilum	>XP_006697396.1
>A_thaliana	>NP_178124.2
>H_sapiens	>NP_006436.3
>C_elegans	>CCD66122.1

Structure-based alignment (extract)

```
S_cerevisiae        EEYPVKVKVSYQKLLKNYVLNELHPTLPTNHNKTKLLKSLKNTKYFQQTTIDWVEAGLQL    571
S_pombe             PNQPVKVRVSYQKLLKSHVMNKLHMAHPKSHTNRSLLRQLKNTKFFQSTSIDWVEAGLQV    519
C_thermophilum      PKQPVKVRVSYQKLLKTYVLNELHRKRPKSMQKQSLLRTLKQTKFFQQTTIDWVEAGLQV    469
A_thaliana          PAYPVKVRVSYQKLLKCYVLNELHHRPPKAQKKKHLFRSLAATKFFQSTELDWVEVGLQV    519
H_sapiens           AGQPVKVRVSYQKLLKYYVLNALKHRPPKAQKKRYLFRSFKATKFFQSTKLDWVEVGLQV    496
C_elegans           AGMPVKVRVSYQKLLKVFVLNALKHRPPKPQKRRYLFRSFKATKFFQTTTLDWVEAGLQV    488
                       ****:******** .*:* *:   *.   .  *:: :  **:** * :****.***:

S_cerevisiae        CRQGHNMLNLLIHRKGLTYLHLDYNFNLKPTKTLTTKERKKSRLGNSFHLMRELLKMMKL    631
S_pombe             CRQGYNMLQLLIHRKGLTYLHLDYNCNLKPTKTLTTKERKKSRFGNAFHLMREILRLTKL    579
C_thermophilum      CRQGFNMLNLLIHRKNLTYLHLDYNFNLKPVKTLTTKERKKSRFGNAFHLMREILRLTKL    529
A_thaliana          CRQGYNMLNLLIHRKNLNYLHLDYNFNLKPVKTLTTKERKKSRFGNAFHLCREILRLTKL    579
H_sapiens           CRQGYNMLNLLIHRKNLNYLHLDYNFNLKPVKTLTTKERKKSRFGNAFHLCREVLRLTKL    556
C_elegans           LRQGYNMLNLLIHRKNLNYLHLDYNFNLKPVKTLTTKERKKSRFGNAFHLCREILRLTKL    548
                     ***.***:******.*.******* ****.************:**:*** **:*:: **

S_cerevisiae        IVDTHVQFRLGNVDAFQLADGIHYILNHIGQLTGIYRYKYKVMHQIRACKDLKHIIYYKF    691
S_pombe             IVDSHVQYRLGNIDAYQLADGLHYIFNHVGQLTGMYRYKYRLMRQIRACKDFKHLIYYRF    639
C_thermophilum      IVDAQVQYRLGNIDAFQLADGIHYAFNHVGQLTGMYRYKYKLMHQIRSCKDLKHLIYYRF    589
A_thaliana          VVDANVQFRLGNVDAFQLADGLQYIFSHVGQLTGMYRYKYRLMRQIRMCKDLKHLIYYRF    639
H_sapiens           VVDSHVQYRLGNVDAFQLADGLQYIFAHVGQLTGMYRYKYKLMRQIRMCKDLKHLIYYRF    616
C_elegans           VVDAHVQYRLNNVDAYQLADGLQYIFAHVGQLTGMYRYKYKLMRQVRMCKDLKHLIYYRF    608
                    :**::**:**.*:**:*****::* : *:*****:*****::*:*:* ***:**:***:*
```

Bibliography

(1) Palermo, G.; Cavalli, A.; Klein, M. L.; Alfonso-Prieto, M.; Dal Peraro, M.; De Vivo, M. *Acc. Chem. Res.* **2015**, *48*, 220.

(2) Fragkos, M.; Ganier, O.; Coulombe, P.; Mechali, M. *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 360.

(3) Brosh, R. M. *Nat. Rev. Cancer* **2013**, *13*, 542.

(4) Wang, J. C. *Nat. Rev. Mol. Cell Bio.* **2002**, *3*, 430.

(5) Lange, S. S.; Takata, K.; Wood, R. D. *Nat. Rev. Cancer* **2011**, *11*, 96.

(6) Johnson, K. A. *Bba-Proteins Proteom.* **2010**, *1804*, 1041.

(7) Beard, W. A.; Shock, D. D.; Vande Berg, B. J.; Wilson, S. H. *J. Biol. Chem.* **2002**, *277*, 47393.

(8) Nakamura, T.; Zhao, Y.; Yamagata, Y.; Hua, Y. J.; Yang, W. *Nature* **2012**, *487*, 196.

(9) McCulloch, S. D.; Kunkel, T. A. *Cell Res.* **2008**, *18*, 148.

(10) Bergoglio, V.; Boyer, A. S.; Walsh, E.; Naim, V.; Legube, G.; Lee, M. Y.; Rey, L.; Rosselli, F.; Cazaux, C.; Eckert, K. A.; Hoffmann, J. S. *J. Cell Biol.* **2013**, *201*, 395.

(11) Broughton, B. C.; Cordonnier, A.; Kleijer, W. J.; Jaspers, N. G.; Fawcett, H.; Raams, A.; Garritsen, V. H.; Stary, A.; Avril, M. F.; Boudsocq, F.; Masutani, C.; Hanaoka, F.; Fuchs, R. P.; Sarasin, A.; Lehmann, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 815.

(12) Woodgate, R.; Levine, A. S. *Cancer Surv.* **1996**, *28*, 117.

(13) Yang, W. *Biochemistry* **2014**, *53*, 2793.

(14) Yang, W. *FEBS Lett* **2005**, *579*, 868.

(15) Patra, A.; Banerjee, S.; Johnson Salyard, T. L.; Malik, C. K.; Christov, P. P.; Rizzo, C. J.; Stone, M. P.; Egli, M. *J. Am. Chem. Soc.* **2015**, *137*, 7011.

(16) Pata, J. D. *Biochim. biophys. acta* **2010**, *1804*, 1124.

(17) Chou, K. M. *Antioxid Redox Sign.* **2011**, *14*, 2521.

(18) Steitz, T. A. *Curr. Opin. Struc. Biol.* **1993**, *3*, 31.

(19) Steitz, T. A.; Steitz, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6498.

(20) Yang, W.; Lee, J. Y.; Nowotny, M. *Mol. Cell* **2006**, *22*, 5.

(21) Zhao, C.; Rajashankar, K. R.; Marcia, M.; Pyle, A. M. *Nat. Chem. Biol.* **2015**, *11*, 967.

(22) Gao, Y.; Yang, W. *Science* **2016**, *352*, 1334.

(23) Perera, L.; Freudenthal, B. D.; Beard, W. A.; Shock, D. D.; Pedersen, L. G.; Wilson, S. H. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, E5228.

(24) Ivanov, I.; Tainer, J. A.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1465.

(25) Horton, N. C.; Perona, J. J. *Biochemistry* **2004**, *43*, 6841.

(26) Vyas, R.; Reed, A. J.; Tokarsky, E. J.; Suo, Z. C. *J. Am. Chem. Soc.* **2015**, *137*, 5225.

(27) Kelley, M. R.; Fishel, M. L. *Anti-Cancer Agent. Med.* **2008**, *8*, 417.

(28) Sale, J. E.; Lehmann, A. R.; Woodgate, R. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 141.

(29) Masutani, C.; Kusumoto, R.; Yamada, A.; Dohmae, N.; Yokoi, M.; Yuasa, M.; Araki, M.; Iwai, S.; Takio, K.; Hanaoka, F. *Nature* **1999**, *399*, 700.

(30) Srivastava, A. K.; Han, C.; Zhao, R.; Cui, T.; Dai, Y.; Mao, C.; Zhao, W.; Zhang, X.; Yu, J.; Wang, Q. E. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 4411.

(31) Chaney, S. G.; Campbell, S. L.; Bassett, E.; Wu, Y. *Crit. Rev. Oncol. Hematol.* **2005**, *53*, 3.

(32) Liu, Y.; Yang, Y.; Tang, T. S.; Zhang, H.; Wang, Z.; Friedberg, E.; Yang, W.; Guo, C. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 1789.

(33) Inui, H.; Oh, K. S.; Nadem, C.; Ueda, T.; Khan, S. G.; Metin, A.; Gozukara, E.; Emmert, S.; Slor, H.; Busch, D. B.; Baker, C. C.; DiGiovanna, J. J.; Tamura, D.; Seitz, C. S.; Gratchev, A.; Wu, W. H.; Chung, K. Y.; Chung, H. J.; Azizi, E.; Woodgate, R.; Schneider, T. D.; Kraemer, K. H. *J. Invest. Dermat.* **2008**, *128*, 2055.

(34) Marcia, M.; Pyle, A. M. *Cell* **2012**, *151*, 497.

(35) Alder, B. J., Wainwrite, T.E. *J. Chem. Phys.* **1957**, *27*, 1208.

(36) Fermi, E., Pasta, J., Ulam, S. *Los Alamos report* **1955**, 1940.

(37) Frenkel, D., Smith, B. *Academic Press San Diego* **2002**.

(38) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1996**, *118*, 2309.

(39) Scott, W. R. P.; Hunenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Kruger, P.; van Gunsteren, W. F. *J. Phys. Chem. A* **1999**, *103*, 3596.

(40) Brooks, B. R., Bruccoleri, R. E., Olfason, B. D., State, D. J., Swaminathan, S., Karplus, M. *J. Comp. Chem.* **1983**, *4*, 187.

(41) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., 3rd; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817.

(42) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269.

(43) Ewald, P. *Ann. Phys.* **1921**, *64*, 253.

(44) Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Dinola, A., Haak, J.R. *J. Chem. Phys.* **1984**, *81*, 3684.

(45) Hoover, W. G. *Phys Rev A* **1985**, *31*, 1695.

(46) Kohn, W., Sham, L. *Phys. Rev. A* **1965**, *140*, A1133.

(47) Vosko, S., Wilk, L., Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.

(48) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(49) Becke, A. D. *J. Chem. Phys.* **1988**, *88*, 1053.

(50) Becke, A. D. *Phys. Rev A* **1988**, *38*, 3098.

(51) Koch, W., Holthausen, M. *Wiley* **1999**.

(52) Car, R.; Parrinello, M. *Phy. Review. Lett.* **1985**, *55*, 2471.

(53) Rothlisberger, U.; Carloni, P. *Lect. Notes. Phys.* **2006**, *704*, 449.

(54) Carloni, P.; Rothlisberger, U.; Parrinello, M. *Acc. Chem. Res.* **2002**, *35*, 455.

(55) Genna, V.; Vidossich, P.; Ippoliti, E.; Carloni, P.; De Vivo, M. *J. Am. Chem. Soc.* **2016**, *138*, 14592.

(56) Warshel, A., Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227.

(57) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941.

(58) Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. *J. Chem. Theory Comput.* **2005**, *1*, 1176.

(59) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*.

(60) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Phys. Chem. B* **2002**, *106*, 7300.

(61) Biertumpfel, C.; Zhao, Y.; Kondo, Y.; Ramon-Maiques, S.; Gregory, M.; Lee, J. Y.; Masutani, C.; Lehmann, A. R.; Hanaoka, F.; Yang, W. *Nature* **2010**, *465*, 1044.

(62) Ucisik, M. N.; Hammes-Schiffer, S. *J. Am. Chem. Soc.* **2015**, *137*, 13240.

(63) McCulloch, S. D.; Kokoska, R. J.; Masutani, C.; Iwai, S.; Hanaoka, F.; Kunkel, T. A. *Nature* **2004**, *428*, 97.

(64) Cruet-Hennequart, S.; Gallagher, K.; Sokol, A. M.; Villalan, S.; Prendergast, A. M.; Carty, M. P. *Subcell Biochem.* **2010**, *50*, 189.

(65) Woodgate, R. *DNA Repair* **2001**, *485*, 83.

(66) Cruet-Hennequart, S.; Villalan, S.; Kaczmarczyk, A.; O'Meara, E.; Sokol, A. M.; Carty, M. P. *Cell Cycle* **2009**, *8*, 3039.

(67) Ummat, A.; Silverstein, T. D.; Jain, R.; Buku, A.; Johnson, R. E.; Prakash, L.; Prakash, S.; Aggarwal, A. K. *J. Mol. Biol.* **2012**, *415*, 627.

(68) Vaisman, A.; Ling, H.; Woodgate, R.; Yang, W. *EMBO J* **2005**, *24*, 2957.

(69) Parsons, J. L.; Nicolay, N. H.; Sharma, R. A. *Antioxid. Redox Sign.* **2013**, *18*, 851.

(70) Roberts, S. A.; Gordenin, D. A. *Nat. Rev. Cancer* **2014**, *14*, 786.

(71) Zhao, Y.; Gregory, M. T.; Biertumpfel, C.; Hua, Y. J.; Hanaoka, F.; Yang, W. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 8146.

(72) Lin, Q.; Clark, A. B.; McCulloch, S. D.; Yuan, T.; Bronson, R. T.; Kunkel, T. A.; Kucherlapati, R. *Cancer Res.* **2006**, *66*, 87.

(73) Kulaksiz, G.; Reardon, J. T.; Sancar, A. *Mol. Cell Biol.* **2005**, *25*, 9784.

(74) Tomicic, M. T.; Aasland, D.; Naumann, S. C.; Meise, R.; Barckhausen, C.; Kaina, B.; Christmann, M. *Cancer Res.* **2014**, *74*, 5585.

(75) Ling, H.; Boudsocq, F.; Woodgate, R.; Yang, W. *Cell* **2001**, *107*, 91.

(76) De Vivo, M.; Dal Peraro, M.; Klein, M. L. *J. Am. Chem. Soc.* **2008**, *130*, 10955.

(77) Wang, L. H.; Yu, X. Y.; Hu, P.; Broyde, S.; Zhang, Y. K. *J. Am. Chem. Soc.* **2007**, *129*, 4731.

(78) Lior-Hoffmann, L.; Wang, L. H.; Wang, S. L.; Geacintov, N. E.; Broyde, S.; Zhang, Y. K. *Nucleic Acids Res.* **2012**, *40*, 9193.

(79) Andrei A. Golosov, J. J. W., ; Lorena S. Beese,; Karplus M. *Structure* **2010**, *18*, 83.

(80) Su, Y.; Patra, A.; Harp, J. M.; Egli, M.; Guengerich, F. P. *J. Biol. Chem.* **2015**, *290*, 15921.

(81) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins* **2010**, *78*, 1950.

(82) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157.

(83) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comp. Chem.* **2005**, *26*, 1781.

(84) Grest, G. S.; Kremer, K. *Phys. Rev. A* **1986**, *33*, 3628.

(85) Dal Peraro, M.; Spiegel, K.; Lamoureux, G.; De Vivo, M.; DeGrado, W. F.; Klein, M. L. *J. Struct. Biol.* **2007**, *157*, 444.

(86) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

(87) Barducci, A.; Bussi, G.; Parrinello, M. *Phys. Rev. Lett.* **2008**, *100*, 020603.

(88) Ho, M. H.; De Vivo, M.; Dal Peraro, M.; Klein, M. L. *J. Am. Chem. Soc.* **2010**, *132*, 13702.

(89) Palermo, G.; Stenta, M.; Cavalli, A.; Dal Peraro, M.; De Vivo, M. *J. Chem. Theory Comput.* **2013**, *9*, 857.

(90) De Vivo, M.; Ensing, B.; Dal Peraro, M.; Gomez, G. A.; Christianson, D. W.; Klein, M. L. *J. Am. Chem. Soc.* **2007**, *129*, 387.

(91) Vidossich, P.; Magistrato, A. *Biomolecules* **2014**, *4*, 616.

(92) Zhao, Y.; Biertumpfel, C.; Gregory, M. T.; Hua, Y. J.; Hanaoka, F.; Yang, W. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 7269.

(93) Patra, A.; Nagy, L. D.; Zhang, Q. Q.; Su, Y.; Muller, L.; Guengerich, F. P.; Egli, M. *J. Biol. Chem.* **2014**, *289*, 16867.

(94) Wheate, N. J.; Walker, S.; Craig, G. E.; Oun, R. *Dalton Trans.* **2010**, *39*, 8113.

(95) Degan, P.; Shigenaga, M. K.; Park, E. M.; Alperin, P. E.; Ames, B. N. *Carcinogenesis* **1991**, *12*, 865.

(96) Shimoda, R.; Nagashima, M.; Sakamoto, M.; Yamaguchi, N.; Takagi, H.; Mori, M.; Hirohashi, S.; Yokota, J.; Kasai, H. *Gastroenterology* **1994**, *106*, A983.

(97) Fraga, C. G.; Motchnik, P. A.; Shigenaga, M. K.; Helbock, H. J.; Jacob, R. A.; Ames, B. N. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 11003.

(98) Katafuchi, A.; Sassa, A.; Niimi, N.; Gruz, P.; Fujimoto, H.; Masutani, C.; Hanaoka, F.; Ohta, T.; Nohmi, T. *Nucleic Acids Res.* **2010**, *38*, 859.

(99) Ucisik, M. N.; Hammes-Schiffer, S. *J. Chem. Inf. Model.* **2015**, *55*, 2672.

(100) Makino, Y.; Iwasaki, K.; Hirata, T. *J Agr Eng Res* **1997**, *67*, 47.

(101) Abdelaal, Y. A. I.; Hammock, B. D. *Acs. Sym. Ser.* **1985**, *276*, 135.

(102) Zhao, L. L.; Pence, M. G.; Eoff, R. L.; Yuan, S.; Fercu, C. A.; Guengerich, F. P. *Febs J* **2014**, *281*, 4394.

(103) Beckman, J. W.; Wang, Q. X.; Guengerich, F. P. *J. Biol. Chem.* **2008**, *283*, 36711.

(104) Nair, D. T.; Johnson, R. E.; Prakash, L.; Prakash, S.; Aggarwal, A. K. *Structure* **2005**, *13*, 1569.

(105) Maxwell, B. A.; Suo, Z. C. *Biochemistry* **2014**, *53*, 2804.

(106) Bunting, K. A.; Roe, S. M.; Pearl, L. H. *EMBO J* **2003**, *22*, 5883.

(107) Jiang, Q.; Karata, K.; Woodgate, R.; Cox, M. M.; Goodman, M. F. *Nature* **2009**, *460*, 359.

(108) Uljon, S. N.; Johnson, R. E.; Edwards, T. A.; Prakash, S.; Prakash, L.; Aggarwal, A. K. *Structure* **2004**, *12*, 1395.

(109) Lone, S.; Townson, S. A.; Uljon, S. N.; Johnson, R. E.; Brahma, A.; Nair, D. T.; Prakash, S.; Prakash, L.; Aggarwal, A. K. *Mol. Cell* **2007**, *25*, 601.

(110) Swan, M. K.; Johnson, R. E.; Prakash, L.; Prakash, S.; Aggarwal, A. K. *J. Mol. Biol.* **2009**, *390*, 699.

(111) Reich, S.; Guilligay, D.; Pflug, A.; Malet, H.; Berger, I.; Crepin, T.; Hart, D.; Lunardi, T.; Nanao, M.; Ruigrok, R. W. H.; Cusack, S. *Nature* **2014**, *516*, 361.

(112) Balzarini, J.; Das, K.; Bernatchez, J. A.; Martinez, S. E.; Ngure, M.; Keane, S.; Ford, A.; Maguire, N.; Mullins, N.; John, J.; Kim, Y.; Dehaen, W.; Vande Voorde, J.; Liekens, S.; Naesens, L.; Gotte, M.; Maguire, A. R.; Arnold, E. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 3475.

(113) Appleby, T. C.; Perry, J. K.; Murakami, E.; Barauskas, O.; Feng, J.; Cho, A.; Fox, D.; Wetmore, D. R.; McGrath, M. E.; Ray, A. S.; Sofia, M. J.; Swaminathan, S.; Edwards, T. E. *Science* **2015**, *347*, 771.

(114) Broyde, S.; Wang, L.; Rechkoblit, O.; Geacintov, N. E.; Patel, D. J. *Trends Biochem. Sci.* **2008**, *33*, 209.

(115) Franklin, M. C.; Wang, J.; Steitz, T. A. *Cell* **2001**, *105*, 657.

(116) Nilforoushan, A.; Furrer, A.; Wyss, L. A.; van Loon, B.; Sturla, S. J. *J. Am. Chem. Soc.* **2015**, *137*, 4728.

(117) Lodola, A.; De Vivo, M. *Adv. Protein. Chem. Struct. Biol.* **2012**, *87*, 337.

(118) Bae, B.; Nayak, D.; Ray, A.; Mustaev, A.; Landick, R.; Darst, S. A. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, E4178.

(119) Warren, T. K.; Wells, J.; Panchal, R. G.; Stuthman, K. S.; Garza, N. L.; Van Tongeren, S. A.; Dong, L.; Retterer, C. J.; Eaton, B. P.; Pegoraro, G.; Honnold, S.; Bantia, S.; Kotian, P.; Chen, X.; Taubenheim, B. R.; Welch, L. S.; Minning, D. M.; Babu, Y. S.; Sheridan, W. P.; Bavari, S. *Nature* **2014**, *508*, 402.

(120) Hou, G. H.; Cui, Q. *J. Am. Chem. Soc.* **2013**, *135*, 10457.

(121) Florian, J.; Goodman, M. F.; Warshel, A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6819.

(122) Sucato, C. A.; Upton, T. G.; Kashemirov, B. A.; Batra, V. K.; Martinek, V.; Xiang, Y.; Beard, W. A.; Pedersen, L. C.; Wilson, S. H.; McKenna, C. E.; Florian, J.; Warshel, A.; Goodman, M. F. *Biochemistry* **2007**, *46*, 461.

(123) Florian, J.; Goodman, M. F.; Warshel, A. *J. Am. Chem. Soc.* **2003**, *125*, 8163.

(124) Harrison, C. B.; Schulten, K. *J. Chem. Theory Comput.* **2012**, *8*, 2328.

(125) McGinnis, J. L.; Dunkle, J. A.; Cate, J. H. D.; Weeks, K. M. *J. Am. Chem. Soc.* **2012**, *134*, 6617.

(126) Genna, V.; Gaspari, R.; Dal Peraro, M.; De Vivo, M. *Nucleic Acids Res.* **2016**, *44*, 2827.

(127) Iannuzzi, M.; Laio, A.; Parrinello, M. *Phys. Rev. Lett.* **2003**, *90*.

(128) Senn, H. M.; Thiel, W. *Angew. Chem. Int. Edit.* **2009**, *48*, 1198.

(129) Lonsdale, R.; Ranaghan, K. E.; Mulholland, A. J. *Chem. Commun.* **2010**, *46*, 2354.

(130) Monard, G.; Merz, K. M. *Acc. Chem. Res.* **1999**, *32*, 904.

(131) Campomanes, P.; Rothlisberger, U.; Alfonso-Prieto, M.; Rovira, C. *J. Am. Chem. Soc.* **2015**, *137*, 11170.

(132) Dal Peraro, M.; Ruggerone, P.; Raugei, S.; Gervasio, F. L.; Carloni, P. *Curr. Opin. Struct. Biol.* **2007**, *17*, 149.

(133) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *Phys. Chem. Chem. Phys.* **2012**, *14*, 12431.

(134) Castro, C.; Smidansky, E.; Maksimchuk, K. R.; Arnold, J. J.; Korneeva, V. S.; Gotte, M.; Konigsberg, W.; Cameron, C. E. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4267.

(135) Harris, T. K.; Turner, G. J. *IUBMB Life* **2002**, *53*, 85.

(136) Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. *Chem. Rev.* **2006**, *106*, 3210.

(137) Hammes, G. G.; Benkovic, S. J.; Hammes-Schiffer, S. *Biochemistry* **2011**, *50*, 10422.

(138) Bhabha, G.; Lee, J.; Ekiert, D. C.; Gam, J.; Wilson, I. A.; Dyson, H. J.; Benkovic, S. J.; Wright, P. E. *Science* **2011**, *332*, 234.

(139) Kamerlin, S. C.; Haranczyk, M.; Warshel, A. *J. Phys. Chem. B* **2009**, *113*, 1253.

(140) Xiang, Y.; Goodman, M. F.; Beard, W. A.; Wilson, S. H.; Warshel, A. *Proteins* **2008**, *70*, 231.

(141) Warshel, A.; Russell, S. T. *Q. Rev. Biophys.* **1984**, *17*, 283.

(142) Basu, R. S.; Murakami, K. S. *J. Biol. Chem.* **2013**, *288*, 3305.

(143) Rosta, E.; Yang, W.; Hummer, G. *J. Am. Chem. Soc.* **2014**, *136*, 3137.

(144) Golosov, A. A.; Warren, J. J.; Beese, L. S.; Karplus, M. *Structure* **2010**, *18*, 83.

(145) Silva, D. A.; Weiss, D. R.; Avila, F. P.; Da, L. T.; Levitt, M.; Wang, D.; Huang, X. H. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 7665.

(146) Wang, Y.; Schlick, T. *J. Am. Chem. Soc.* **2008**, *130*, 13240.

(147) Da, L. T.; Wang, D.; Huang, X. H. *J. Am. Chem. Soc.* **2012**, *134*, 2399.

(148) Lopez-Canut, V.; Roca, M.; Bertran, J.; Moliner, V.; Tunon, I. *J. Am. Chem. Soc.* **2011**, *133*, 12050.

(149) Andrews, L. D.; Fenn, T. D.; Herschlag, D. *Plos. Biol.* **2013**, *11*.

(150) Miller, B. R.; Beese, L. S.; Parish, C. A.; Wu, E. Y. *Structure* **2015**, *23*, 1609.

(151) Strater, N.; Lipscomb, W. N.; Klabunde, T.; Krebs, B. *Angew. Chem. Int. Edit.* **1996**, *35*, 2024.

(152) Livesay, D. R.; Jambeck, P.; Rojnuckarin, A.; Subramaniam, S. *Biochemistry* **2003**, *42*, 3464.

(153) Steitz, T. A.; Steitz, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6498.

(154) Yang, W. *Nat. Struct. Mol. Bio.* **2008**, *15*, 1228.

(155) Marcia, M.; Somarowthu, S.; Pyle, A. M. *Mobile DNA* **2013**, *4*, 14.

(156) Marcia, M.; Pyle, A. M. *Rna* **2014**, *20*, 516.

(157) Viadiu, H.; Aggarwal, A. K. *Nature Struct. Bio.* **1998**, *5*, 910.

(158) Sun, J.; Viadiu, H.; Aggarwal, A. K.; Weinstein, H. *Biophys J* **2003**, *84*, 3317.

(159) Zhang, J.; McCabe, K. A.; Bell, C. E. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 11872.

(160) Jinek, M.; Jiang, F.; Taylor, D. W.; Sternberg, S. H.; Kaya, E.; Ma, E.; Anders, C.; Hauer, M.; Zhou, K.; Lin, S.; Kaplan, M.; Iavarone, A. T.; Charpentier, E.; Nogales, E.; Doudna, J. A. *Science* **2014**, *343*, 1247997.

(161) Jiang, F.; Taylor, D. W.; Chen, J. S.; Kornfeld, J. E.; Zhou, K.; Thompson, A. J.; Nogales, E.; Doudna, J. A. *Science* **2016**, *351*, 867.

(162) Zhang, F.; Gao, L.; Zetsche, B.; Slaymaker, I.; The Broad Institute Inc., M. I. O. T., Ed. 2016.

(163) Wang, D.; Bushnell, D. A.; Westover, K. D.; Kaplan, C. D.; Kornberg, R. D. *Cell* **2006**, *127*, 941.

(164) Strathern, J.; Malagon, F.; Irvin, J.; Gotte, D.; Shafer, B.; Kireeva, M.; Lubkowska, L.; Jin, D. J.; Kashlev, M. *J. Biol. Chem.* **2013**, *288*, 2689.

(165) Treich, I.; Carles, C.; Sentenac, A.; Riva, M. *Nucleic Acids Res.* **1992**, *20*, 4721.

(166) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215*, 403.

(167) Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Soding, J.; Thompson, J. D.; Higgins, D. G. *Mol. Syst. Biol.* **2011**, *7*, 539.

(168) Notredame, C.; Higgins, D. G.; Heringa, J. *J. Mol. Biol.* **2000**, *302*, 205.

(169) Emsley, P.; Cowtan, K. *Acta Cryst. Biol. Crystallogr.* **2004**, *60*, 2126.

(170) Schrodinger, LLC 2010.

(171) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037.

(172) Misra, V. K.; Draper, D. E. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 12456.

(173) Jeltsch, A.; Alves, J.; Wolfes, H.; Maass, G.; Pingoud, A. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 8499.

(174) Ramachandrakurup, S.; Ammapalli, S.; Ramakrishnan, V. *J. Biomol. Struct. Dyn.* **2016**, 1.

(175) Kurpiewski, M. R.; Engler, L. E.; Wozniak, L. A.; Kobylanska, A.; Koziolkiewicz, M.; Stec, W. J.; Jen-Jacobson, L. *Structure* **2004**, *12*, 1775.

(176) Hanoian, P.; Liu, C. T.; Hammes-Schiffer, S.; Benkovic, S. *Acc. Chem. Res.* **2015**, *48*, 482.

(177) Uyar, A.; Kurkcuoglu, O.; Nilsson, L.; Doruker, P. *Phys. Bio.* **2011**, *8*, 056001.

(178) Pingoud, A.; Jeltsch, A. *Nucleic Acids Res.* **2001**, *29*, 3705.

(179) Dias, A.; Bouvier, D.; Crepin, T.; McCarthy, A. A.; Hart, D. J.; Baudin, F.; Cusack, S.; Ruigrok, R. W. *Nature* **2009**, *458*, 914.

(180) Jiang, F.; Doudna, J. A. *Curr. Opin. Struct. Biol.* **2015**, *30*, 100.

(181) Wright, A. V.; Nunez, J. K.; Doudna, J. A. *Cell* **2016**, *164*, 29.

(182) Yang, W.; Chen, W. Y.; Wang, H.; Ho, J. W. S.; Huang, J. D.; Woo, P. C. Y.; Lau, S. K. P.; Yuen, K. Y.; Zhang, Q. L.; Zhou, W. H.; Bartlam, M.; Watt, R. M.; Rao, Z. H. *Nucleic Acids Res.* **2011**, *39*, 9803.

(183) Cramer, P.; Bushnell, D. A.; Kornberg, R. D. *Science* **2001**, *292*, 1863.

(184) Tafur, L.; Sadian, Y.; Hoffmann, N. A.; Jakobi, A. J.; Wetzel, R.; Hagen, W. J.; Sachse, C.; Muller, C. W. *Mol. Cell* **2016**, *64*, 1135.

(185) Baltimore, D. *Bact. Rev.* **1971**, *35*, 235.

(186) Gerlach, P.; Malet, H.; Cusack, S.; Reguera, J. *Cell* **2015**, *161*, 1267.

(187) Muller, R.; Poch, O.; Delarue, M.; Bishop, D. H.; Bouloy, M. *J. Gen. Vir.* **1994**, *75*, 1345.

(188) Wu, S.; Beard, W. A.; Pedersen, L. G.; Wilson, S. H. *Chem. Rev.* **2014**, *114*, 2759.

(189) Sharp, P. A. *Science* **1991**, *254*, 663.

(190) Pyle, A. M.; Lambowitz, A. M. In *The RNA World*; 3rd ed.; Gesteland, R. F., Cech, T. R., Atkins, J. F., Eds.; Cold Spring Harbor Press: Cold Spring Harbor, 2006, p 469.

(191) Hardy, S. F.; Grabowski, P. J.; Padgett, R. A.; Sharp, P. A. *Nature* **1984**, *308*, 375.

(192) Tseng, C. K.; Cheng, S. C. *Rna* **2013**, *19*, 971.

(193) Staub, E.; Fiziev, P.; Rosenthal, A.; Hinzmann, B. *BioEssays : news and reviews in molecular, cellular and developmental biology* **2004**, *26*, 567.

(194) Berdis, A. J. *Biochemistry* **2008**, *47*, 8253.

(195) Li, H.; Labean, T. H.; Leong, K. W. *Interface focus* **2011**, *1*, 702.

(196) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471.

(197) Sham, W. K. a. L. J. *Phys. Rev. Lett.* **1965**, *140*.

(198) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.

(199) De Vivo, M.; Ensing, B.; Klein, M. L. *J. Am. Chem. Soc.* **2005**, *127*, 11226.

(200) J. Wang, P. C. a. P. A. K. *J. Comput. Chem.* **2000**, *21*, 1049.

(201) De Vivo, M. *Front Biosci (Landmark Ed)* **2011**, *16*, 1619.

(202) De Vivo, M.; Ensing, B.; Dal Peraro, M.; Gomez, G. A.; Christianson, D. W.; Klein, M. L. *J. Am. Chem. Soc.* **2007**, *129*, 387.

(203) Troullier, N.; Martins, J. L. *Phys. Rev. B, Cond. Mat.* **1991**, *43*, 1993.

(204) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *J. Chem. Phys.* **2005**, *122*, 14113.

(205) Glenn J Martyna, M. E. T. *J. Chem. Phys.* **1999**, *110*, 2810.

(206) De Vivo, M.; Cavalli, A.; Carloni, P.; Recanatini, M. *Chemistry* **2007**, *13*, 8437.

(207) Cavalli, A.; De Vivo, M.; Recanatini, M. *Chem. Comm.* **2003**, 1308.

(208) W. F. van Gunsteren, S. R. B., A. A. Eising, P. H. Hünenberger, P. Krüger, A. E. Mark, W. R. P. Scott, and I. G. Tironi *Vdf Hochschulverlag AG an der ETH Zürich* **1996**, 1.

(209) Marcia, M.; Ermler, U.; Peng, G.; Michel, H. *Proteins* **2010**, *78*, 1073.

(210) Kazantsev, A. V.; Krivenko, A. A.; Pace, N. R. *Rna* **2009**, *15*, 266.

(211) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins* **2010**, *78*, 1950.

(212) Galej, W. P.; Wilkinson, M. E.; Fica, S. M.; Oubridge, C.; Newman, A. J.; Nagai, K. *Nature* **2016**, doi: 10.1038/nature19316.

(213) Newman, M.; Lunnen, K.; Wilson, G.; Greci, J.; Schildkraut, I.; Phillips, S. E. *EMBO J* **1998**, *17*, 5466.

(214) Flores, H.; Osuna, J.; Heitman, J.; Soberon, X. *Gene* **1995**, *157*, 295.

(215) Sen, S.; Nilsson, L. *Biophys. J* **1999**, *77*, 1782.

(216) Horton, J. R.; Cheng, X. *J. Mol. Biol.* **2000**, *300*, 1049.

(217) Nastri, H. G.; Evans, P. D.; Walker, I. H.; Riggs, P. D. *J. Biol. Chem.* **1997**, *272*, 25761.

(218) Horton, N. C.; Connolly, B. A.; Perona, J. J. *J. Am. Chem. Soc.* **2000**, *122*, 3314.

(219) Horton, N. C.; Otey, C.; Lusetti, S.; Sam, M. D.; Kohn, J.; Martin, A. M.; Ananthnarayan, V.; Perona, J. J. *Biochemistry* **2002**, *41*, 10754.

(220) Wenz, C.; Jeltsch, A.; Pingoud, A. *J Biol Chem* **1996**, *271*, 5565.

(221) Xu, Q. S.; Kucera, R. B.; Roberts, R. J.; Guo, H. C. *Structure* **2004**, *12*, 1741.

(222) Shen, B. W.; Xu, D.; Chan, S. H.; Zheng, Y.; Zhu, Z.; Xu, S. Y.; Stoddard, B. L. *Nucleic Acids Res.* **2011**, *39*, 8223.

(223) Mierzejewska, K.; Siwek, W.; Czapinska, H.; Kaus-Drobek, M.; Radlinska, M.; Skowronek, K.; Bujnicki, J. M.; Dadlez, M.; Bochtler, M. *Nucleic Acids Res.* **2014**, *42*, 8745.

(224) Wah, D. A.; Hirsch, J. A.; Dorner, L. F.; Schildkraut, I.; Aggarwal, A. K. *Nature* **1997**, *388*, 97.

(225) Bitinaite, J.; Wah, D. A.; Aggarwal, A. K.; Schildkraut, I. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 10570.

(226) Huai, Q.; Colandene, J. D.; Topal, M. D.; Ke, H. *Nat. Struct. Bio.* **2001**, *8*, 665.

(227) Deibert, M.; Grazulis, S.; Sasnauskas, G.; Siksnys, V.; Huber, R. *Nat. Struct. Bio.* **2000**, *7*, 792.

(228) Vanamee, E. S.; Viadiu, H.; Kucera, R.; Dorner, L.; Picone, S.; Schildkraut, I.; Aggarwal, A. K. *EMBO J* **2005**, *24*, 4198.

(229) Dunten, P. W.; Little, E. J.; Gregory, M. T.; Manohar, V. M.; Dalton, M.; Hough, D.; Bitinaite, J.; Horton, N. C. *Nucleic Acids Res.* **2008**, *36*, 5405.

(230) Crepin, T.; Dias, A.; Palencia, A.; Swale, C.; Cusack, S.; Ruigrok, R. W. *J. Virol.* **2010**, *84*, 9096.

(231) Reich, S.; Guilligay, D.; Pflug, A.; Malet, H.; Berger, I.; Crepin, T.; Hart, D.; Lunardi, T.; Nanao, M.; Ruigrok, R. W.; Cusack, S. *Nature* **2014**, *516*, 361.

(232) Lee, J. Y.; Chang, J.; Joseph, N.; Ghirlando, R.; Rao, D. N.; Yang, W. *Mol. Cell* **2005**, *20*, 155.

(233) Junop, M. S.; Yang, W.; Funchain, P.; Clendenin, W.; Miller, J. H. *DNA Repair* **2003**, *2*, 387.

(234) Orans, J.; McSweeney, E. A.; Iyer, R. R.; Hast, M. A.; Hellinga, H. W.; Modrich, P.; Beese, L. S. *Cell* **2011**, *145*, 212.

(235) Korada, S. K. C.; Johns, T. D.; Smith, C. E.; Jones, N. D.; McCabe, K. A.; Bell, C. E. *Nucleic Acids Res.* **2013**, *41*, 5887.

(236) Miyazono, K.; Ishino, S.; Tsutsumi, K.; Ito, T.; Ishino, Y.; Tanokura, M. *Nucleic Acids Res.* **2015**, *43*, 7122.

(237) Rychlik, M. P.; Chon, H.; Cerritelli, S. M.; Klimek, P.; Crouch, R. J.; Nowotny, M. *Mol. Cell* **2010**, *40*, 658.

(238) Figiel, M.; Chon, H.; Cerritelli, S. M.; Cybulska, M.; Crouch, R. J.; Nowotny, M. *J. Biol. Chem.* **2011**, *286*, 10540.

(239) Hirano, H.; Gootenberg, J. S.; Horii, T.; Abudayyeh, O. O.; Kimura, M.; Hsu, P. D.; Nakane, T.; Ishitani, R.; Hatada, I.; Zhang, F.; Nishimasu, H.; Nureki, O. *Cell* **2016**, *164*, 950.

(240) Nishimasu, H.; Cong, L.; Yan, W. X.; Ran, F. A.; Zetsche, B.; Li, Y.; Kurabayashi, A.; Ishitani, R.; Zhang, F.; Nureki, O. *Cell* **2015**, *162*, 1113.

(241) Sawaya, M. R.; Prasad, R.; Wilson, S. H.; Kraut, J.; Pelletier, H. *Biochemistry* **1997**, *36*, 11205.

(242) Kraynov, V. S.; Showalter, A. K.; Liu, J.; Zhong, X.; Tsai, M. D. *Biochemistry* **2000**, *39*, 16008.

(243) Xiang, Y.; Oelschlaeger, P.; Florian, J.; Goodman, M. F.; Warshel, A. *Biochemistry* **2006**, *45*, 7036.

(244) Silverstein, T. D.; Johnson, R. E.; Jain, R.; Prakash, L.; Prakash, S.; Aggarwal, A. K. *Nature* **2010**, *465*, 1039.

(245) Johnson, R. E.; Trincao, J.; Aggarwal, A. K.; Prakash, S.; Prakash, L. *Mol. Cell Biol.* **2003**, *23*, 3008.

(246) Sharma, A.; Kottur, J.; Narayanan, N.; Nair, D. T. *Nucleic Acids Res.* **2013**, *41*, 5104.

(247) Zhou, B. L.; Pata, J. D.; Steitz, T. A. *Mol. Cell* **2001**, *8*, 427.

(248) Hoffmann, N. A.; Jakobi, A. J.; Moreno-Morcillo, M.; Glatt, S.; Kosinski, J.; Hagen, W. J.; Sachse, C.; Muller, C. W. *Nature* **2015**, *528*, 231.

(249) Zamyatkin, D. F.; Parra, F.; Alonso, J. M. M.; Harki, D. A.; Peterson, B. R.; Grochulski, P.; Ng, K. K. S. *J. Biol. Chem.* **2008**, *283*, 7705.

(250) Ng, K. K.; Cherney, M. M.; Vazquez, A. L.; Machin, A.; Alonso, J. M.; Parra, F.; James, M. N. *J. Biol. Chem.* **2002**, *277*, 1381.

(251) Hansen, J. L.; Long, A. M.; Schultz, S. C. *Structure* **1997**, *5*, 1109.

(252) Ferrer-Orta, C.; Arias, A.; Perez-Luque, R.; Escarmis, C.; Domingo, E.; Verdaguer, N. *J. Biol. Chem.* **2004**, *279*, 47212.

(253) Ago, H.; Adachi, T.; Yoshida, A.; Yamamoto, M.; Habuka, N.; Yatsunami, K.; Miyano, M. *Structure* **1999**, *7*, 1417.

(254) Yap, T. L.; Xu, T.; Chen, Y. L.; Malet, H.; Egloff, M. P.; Canard, B.; Vasudevan, S. G.; Lescar, J. *J. Virol.* **2007**, *81*, 4753.

(255) Shen, B. W.; Landthaler, M.; Shub, D. A.; Stoddard, B. L. *J. Mol. Biol.* **2004**, *342*, 43.