Genetic and structural study of DNA-dependent RNA polymerase II of Trypanosoma brucei, towards the designing of novel antiparasitic agents.

Louis Papageorgiou [1, 2, 3] , **Vasileios Megalooikonomou** [3] , **Dimitrios Vlachakis** [Corresp. 2, 3]

[1] Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece
[2] Computational Biology & Medicine Group, Biomedical Research Foundation, Academy of Athens, Athens, Attica, Greece
[3] Computer Engineering and Informatics Department, University of Patras, Patra, Greece

Corresponding Author: Dimitrios Vlachakis
Email address: dvlachakis@bioacademy.gr

Trypanosoma brucei brucei (TBB) belongs to the unicellular parasitic protozoa organisms, specifically to the *Trypanosoma* genus of the *Trypanosomatidae* class. A variety of different vertebrate species can be infected by TBB including humans and animals. Under particular conditions, the TBB can be hosted by wild and domestic animals; thereby an important reservoir of infection always remains available to transmit through the tsetse flies. Although the TBB parasite is one of the leading causes of death in the most underdeveloped countries, to date, there is neither vaccination available nor any drug against TBB infection. The subunit RPB1 of the TBB DNA-directed RNA polymerase II (DdRpII) constitutes an ideal target for the design of novel inhibitors, since it is instrumental role is vital for the parasite's survival, proliferation, and transmission. A major goal of the described study is to provide insights for novel anti-TBB agents via a state of the art drug discovery approach of the TBB DdRpII RPB1. In an attempt to understand the function and action mechanisms of this parasite enzyme related to its molecular structure, an in-depth evolutionary study has been conducted in parallel to the *in silico* molecular designing of the 3D enzyme model, based on state of the art comparative modelling and molecular dynamics techniques. Based on theevolutionary studies results nine new invariant, first-time reported, highly conserved regions have been identified within the DdRpII family enzymes. Consequently, those patches have been examined both at the sequence and structural level and have been evaluated in regards to their pharmacological targeting appropriateness. Finally, the pharmacophore elucidation study enabled us to virtually *in silico* screenhundreds of compounds and evaluate their interaction capabilities with the enzyme. It was found that a series of Chlorine-rich set of compounds were the optimal inhibitors for the TBB DdRpII RPB1 enzyme. All-in-all, herein we present a series of new sites on the TBB DdRpII RPB1 of high pharmacological interest, alongside the

construction of the 3D model of the enzyme and the suggestion of a new *in silico* pharmacophore model for fast screening of potential inhibiting agents.

Genetic and structural study of DNA-dependent RNA polymerase II of Trypanosoma brucei, towards the designing of novel antiparasitic agents.

Louis Papageorgiou[1,2,3], Vasileios Megalooikonomou[2] and Dimitrios Vlachakis[1,2,#]

[1]Computational Biology & Medicine Group, Biomedical Research Foundation, Academy of Athens, Soranou Efessiou 4, Athens 11527, Greece

[2]Computer Engineering and Informatics Department, University of Patras, Patras University Campus, 26504, Greece

[3]Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, University Campus, Athens, 15784, Greece

#Correspondence to:
Dimitrios Vlachakis,
Bioinformatics & Medical Informatics Team, Biomedical Research Foundation, Academy of Athens, SoranouEfessiou 4, Athens 11527, Greece
Tel: + 30 210 6597 647, Fax: +30 210 6597 545
E-mail: dvlachakis@bioacademy.gr

Abstract

Trypanosoma brucei brucei (TBB) belongs to the unicellular parasitic protozoa organisms, specifically to the *Trypanosoma* genus of the *Trypanosomatidae* class. A variety of different vertebrate species can be infected by TBB including humans and animals. Under particular conditions, the TBB can be hosted by wild and domestic animals; thereby an important reservoir of infection always remains available to transmit through the tsetse flies. Although the TBB parasite is one of the leading causes of death in the most underdeveloped countries, to date, there is neither vaccination available nor any drug against TBB infection. TBB DNA-dependent RNA polymerase II (DdRpII subunit RPB1) is an ideal target for the design of novel inhibitors against TBB. This enzyme plays a critical role in parasite's survival, proliferation, and transmission. A major goal of the described study is to provide insights for novel anti-TBB agents via a state of the art drug discovery approach of the TBB DdRpII RPB1. In an attempt to understand the function and action mechanisms of this parasite enzyme related to its molecular structure, an in-depth evolutionary study has been conducted in parallel to the *in silico* molecular designing of the 3D enzyme model, based on state of the art comparative modelling and molecular dynamics techniques. Based on the evolutionary studies results nine new invariant, first-time reported, highly conserved regions have been identified within the DdRpII family enzymes. Consequently, those patches have been examined both at the sequence and structural level and have been evaluated in regards to their pharmacological targeting appropriateness. Finally, a 3D pharmacophore model was constructed specifically for the TBB DdRpII RPB1 enzyme. All-in-all, herein we present a series of new sites on the TBB DdRpII RPB1 of high pharmacological interest, alongside the construction of the 3D model of the enzyme and the suggestion of a new *in silico* pharmacophore model for fast screening of potential inhibiting agents.

Introduction

African trypanosome parasites cause human sleeping sickness and nagana in Africa, Asia, and South America. More than 95% of reported cases are caused by two subspecies of *Trypanosoma brucei brucei* (TBB), the *Trypanosoma brucei gambiense* (TBG) and the *Trypanosoma brucei rhodesiense* (TBR) which is found in western and central Africa (Berriman et al. 2005; World Health Organization 2015). The parasitic infection is transmitted by tsetse flies, which breed in warm and humid areas. Tsetse flies are found living in 36 countries in sub-Saharan Africa, thus putting 60 million people at risk. Currently, about 10,000 new cases each year are reported by the World Health Organization (WHO). Moreover, it is believed that many cases are undiagnosed and unreported. Sleeping sickness can be curable with medication, but it may be fatal if it is left untreated. It is estimated that Human deaths caused by Sleeping sickness are of about 48,000 annually. Bites by the tsetse fly erupt into a red sore on the skin and in the following weeks, the person may have to deal with several symptoms including fever, swollen lymph glands, aching muscles, headaches, and irritability. In advanced stages, the TBB parasite attacks the central nervous system of the host, and in general consul some disorders in personality, circadian rhythm, serenity, speech, and difficulties in walking. Despite the significant treatment advances for patients with sleeping sickness, the parasite's progression is often inevitable and needs more treatment options. Until today, drugs can only be used in the early stages of the disease and without providing 100% reassurance for full convalesce of the patient (Ridley 2002; Ross et al. 2007; Trouiller et al. 2002). The TBB parasite starts its activity after each invasion through its proteins, specifically with its replication enzymes including helicases and polymerases. Such enzymes are ideal targets for inhibitor design since those proteins are crucial for the TBB parasite survival. Being already in possession of the widely known sequence of the DNA-dependent RNA polymerase II (DdRpII) RPB1 (Chung et al. 1993) which plays a significant role in the replication of the parasite, our primary goal is to suppress its function towards replication itself when it infects a human. Although TBB has been reported many times in the past, the three-dimensional structure of its essential enzymes like DdRpII remains unknown so far (Malvy & Chappuis 2011).

Protein structure has been found to be three to ten times more conserved than sequence (Illergard et al. 2009). Thus, when possible, it is preferable to study an enzyme's 3D structure rather than its sequence. Knowledge of the tertiary structure can assist in the understanding of relationships between structure and function (Berg et al. 2002). Herein, the three-dimensional structure of DdRpII subunit RPB1 has been modelled, in an effort to predict the 3D molecular structure that is linked to the function of this enzyme (Bayele 2009; Koch et al. 2016). Two molecular models have been constructed using conventional molecular modelling techniques and two different homolog 3D structures as templates. The established molecular models of the DdRpII RPB1 enzyme of TBB exhibits all known structural motifs that are unique to the DdRpII RPB1 enzymes.

Upon successful completion of the 3D structure prediction of the TBB DdRpII RPB1 protein, molecular dynamics simulations have been performed to structurally improve and benchmark the quality of the 3D models. Moreover, the reliability and viability of the TBB DdRpII RPB1 models were checked using several *in silico* scoring tools such as MOE and Procheck. After the model validation process, a *de novo* structure-based drug design approach has been performed based on two models, which led to the establishment of a 3D novel

pharmacophore model that is highly specific for the DdRpII RPB1 enzyme of TBB. The generated pharmacophore model may be used in future experiments involving the high throughput virtual screening of large compound databases towards the identification of novel anti-TBB agents (Loukatou et al. 2014). The present work opens the field for the design of novel compounds with improved biochemical and clinical characteristics in the future.

Methods

Database sequence search

The full-length protein sequences related to the DdRpII family were extracted from the NCBI database. In total, 36 DdRpII protein sequences were downloaded from several species with fully sequenced genomes (Supplementary data 1).

Genetic and evolutionary analyses

Multiple sequence alignment of the DdRpII protein family sequences were performed using two different programs, MUSCLE (Edgar 2004) and CLUSTALW (Chenna et al. 2003; Thompson et al. 1994). In the next step, multiple sequence alignment was checked with ProtTest3 (Darriba et al. 2011) to estimate the appropriate model of sequence evolution. Phylogenetic analyses were performed by two different ways, and two representative phylogenetic trees were constructed for the DdRpII dataset (Vlachakis et al. 2014b). The first phylogenetic tree was constructed using the MEGA software (Stecher et al. 2014) utilizing Bayesian and Maximum Likelihood statistical methods as described in with 100 bootstrap replicates (Figure 1 and Supplementary data 2). The second phylogenetic tree was constructed using the Jalview software (Waterhouse et al. 2009) utilizing the neighbour joining statistical method in with 100 bootstrap replicates (Supplementary Figure 1 and Supplementary data 3).

Conserved motifs exploration

The phylogenetic trees that derived from the phylogenetic analyses (Jalview and MEGA) were separated in sub-trees, in order to extract the most highly related protein sequences of the TBB DdRpII RPB1 family for the conserved motifs exploration (Figure 2). The full-length amino acid sequences of the closely related proteins with the TPP DdRpII RPB1 protein were aligned using the CLUSTALW (Thompson et al. 1994) statistical method. The evolutionary conserved sequences motifs that were derived from the multiple sequence alignment were identified through the consensus sequence and logo graph where generated using Jalview (Waterhouse et al. 2009) (Figure 2).

Molecular modelling

All calculations and visual constructions were performed using the Molecular Operating Environment (MOE) version 2013.08 software package developed by Chemical Computing Group (Montreal, Canada) on a cloud-based multi core High Performance Computing (HPC) cluster (Loukatou et al. 2014).

Identification of templates structures and sequence alignment

The amino acid sequence of the TBB DdRpII RPB1 was retrieved from the conceptual translation of the trypanosomal RNA polymerase largest subunit genes at the NCBI database (http://www.ncbi.nlm.nih.gov/) (UniProtKB/Swiss-Prot: P17545.1) (Das et al. 2006; Evers et al. 1989). The blastp algorithm (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to identify homologous structures by searching in the Protein Data Bank (PDB). The multiple sequence alignment was performed using MOE (Vilar et al. 2008).

Homology Modelling

The homology modelling of the Tbb DdRPII RPB1 was carried out using MOE. The selection of template crystal structures for homology modelling was based on the primary sequence identity and similarity (Figure 3, Supplementary Figures 2 and 3), and the crystal resolution (Nayeem et al. 2006). The crystal structure of *Schizosaccharomyces pombe* DdRpII RPB1 (PDB: 3H0G) was used as template structure for the model A, while the crystal structure of *Bos taurus* DdRpII RPB1 (PDB: 5FLM) was used for building model B. The MOE homology model method is separated into four main steps. First, comes a primary fragment geometry specification. Second the insertion and deletions task. The third step is the loop selection and the side-chain packing, and the last step is the final model selection and refinement (Figures 4, 5 and Supplementary data 4 and 5) (Papageorgiou et al. 2014; Vlachakis et al. 2013b). Subsequently, energy minimization was done in MOE initially using the Amber99 (Wang et al. 2000) force-field as implemented into the same package. The energy minimization process was applied up to a gradient of 0.0001, in an effort to remove the geometrical strain (Vlachakis et al. 2013a).

Molecular electrostatic potential

Molecular electrostatic potential surfaces were calculated by solving the non-linear Poisson – Boltzmann equation using finite difference method as implemented into the MOE and PyMol Software (Seeliger & de Groot 2010; Vilar et al. 2008). The potential was calculated on solid points per side. Protein contact potential is an automated representation where the false red/blue charge-smoothed surface is shown on the protein (Figure 4). Amber99 charges and atomic radii were used for this calculation.

Molecular dynamics

The Molecular Dynamics simulations of both TBB DdRPII RPB1 3D models A and B were executed in a periodic cell, which was explicitly solvated with simple point charge (SPC) water. The truncated octahedron box was chosen for solvating the models, with a set distance of 7Å clear of the protein. The molecular dynamic simulations were conducted at 300 K, 1 atm with a set 2 fsecond step size for a total of one hundred nanoseconds. For the purposes of this study we opted for a NVT ensemble in a canonical environment (Vlachakis et al. 2014a). NVT stands for Number of atoms, Volume, and Temperature that remain constant throughout the calculation (Vlachakis 2009). The intricate zinc ions were included in the molecular dynamics simulations as integral parts of the modelled biological system (Chakravorty & Merz 2014; Temiz et al. 2010). However, due to the nature of the ions, we had to limit the allowed degrees of freedom for those molecules. Thus, the potential of the zinc ions was constrained in the three dimensional conformational space in the vicinity of the TBB DdRPII RPB1 3D models. The

ions were prepositioned in the 3D models of TBB DdRPII RPB1, after structural superposition to the template x-ray structure. The models were structurally optimized and adjusted locally by subsequent energy minimizations, in an effort to eliminate any molecular clashes and minimize the constrain energy. A radius of 6Å around each ion was given full degrees of freedom during the abovementioned structural optimizations. Provided that the TBB DdRPII RPB1 is a nucleotide processing enzyme, whose structure coordinates a repertoire of ions (e.g. Zinc, Mg++), the AMBER99 forcefield was selected (Figure 6). The AMBER99 forcefield is fully parameterized for our biological system as it implements ff10 parameters for amino acids and nucleic acids as well as EHT for small molecules, such as ions/cations at the same time (Vilar et al. 2008). AM1-BCC charges were applied since the molecular system included the ion molecules. The results of the molecular dynamics simulations for both models were collected into a database by MOE for further analysis. The full simulation trajectories and molecular dynamics graphs for both models are presented in Figures 7-11.

Model evaluation

The produced models were initially evaluated within the MOE package by a residue packing quality function, which depends on the number of buried non-polar side-chain groups and on hydrogen bonding. Moreover, the suite PROCHECK (Laskowski et al. 1996) was employed to further evaluate the quality of the produced models. Finally, MOE and its build in protein check module was used to evaluate whether the models of DdRpII RPB1 domains are similar to known protein structures of this family (Supplementary data 6, 7 and 8).

Pharmacophore Elucidation

A pharmacophoric feature characterizes a particular property and is not tied to a specific chemical structure; indeed different chemical groups may share the same property and so be represented by the same feature (Vlachakis et al. 2013a). It is thus a mistake to name as pharmacophoric features chemical functionalities such as guanidines or sulfonamides or typical structural skeletons such as flavones or steroids.

The term pharmacophore modeling refers to the generation of a pharmacophore hypothesis for the binding interactions in a particular active site (Vlachakis et al. 2015). Several different pharmacophore models for the same active site can be overlaid and reduced to their shared features so that common interactions are retained. Such a consensus pharmacophore can be considered as the largest common denominator shared by a set of active molecules.

In MOE, the computerized representation of a hypothesized pharmacophore is called a pharmacophore query. A MOE pharmacophore query is a set of query features that are typically created from ligand annotation points. Annotation points are markers in space that show the location and type of biologically important atoms and groups, such as hydrogen donors and acceptors, aromatic centers, projected positions of possible interaction partners or R-groups, charged groups, and bioisosteres. The annotation points on a ligand are the potential locations of the features that will constitute the pharmacophore query. Annotation points relevant to the pharmacophore are converted into query features with the addition of an extra parameter: a non-zero radius that encodes the permissible variation in the pharmacophore query's geometry.

Once generated, a pharmacophore query can be used to screen virtual compound libraries for novel ligands. Pharmacophore queries can also be used to filter conformer databases, e.g. output from molecular docking runs, for biologically active conformations.

Results

Phylogenetic Analysis

In the present study, two phylogenetic analyses of DdRpII family proteins in all available genomes, with putative full-length protein sequences were performed using two different statistical methods from the Jalview and MEGA software. Based on findings, putative members of the DdRpII family were identified in the *Animalia*, *Fungi*, *Plantae, Protista* and C*hromalveolata* kingdom major eukaryotic taxonomic division, as well as viruses (Figure 1 and Supplementary Figure 1). In our analyses, in agreement with previous reports (Smith et al. 1989), we found that DdRpII family is split into two main subunits the RPB1 and the RPB2. The two subunits of the DdRpII family are clearly separated in the phylogenetic trees as two major sub-trees were obtained for each one of them (Figure 1 and Supplementary Figure 1). The monophyletic sub-tree of the RPB1 subunit contains the TBB DdRpII RPB1, as well as another 17 leaves, which are related to RPB1 subunit. Furthermore, in the phylogenetic trees, the TBB DdRpII RBP1 forms a distinct monophyletic branch with the *Euplotes octocarinatus* DdRpII RPB1 and the *Plasmodium falciparum* DdRpII RPB1, which is basal to a clade that corresponds to other parasites. The Newick format of the phylogenetic trees is provided (Supplementary Data 1 and 2).

Conserved motifs exploration

Multiple sequence alignment of the DdRpII subunit RPB1 protein sequences from a variety of several species were included in the first sub-tree, highlights important conserved functional domains as described previously by Janet L. Smith and Judith R. Levin (Smith et al. 1989). Good conservation is evident throughout the whole length of the sequence, especially among species that belong to the same taxonomic division (Figure 2).

In this study, an effort has been done to suggest motifs that were probably included in the DdRpII of the subunit RPB1. Regions conserved across all species (eukaryotic and viruses) are indicative of important functional domains of the DdRpII RPB1 enzyme. Finally, the consensus sequence of the multiple sequence alignment highlights nine conserved motifs which are conserved between all species. All of the conserved motifs identified here have not been reported previously, and indisputably deserve further study (Figures 2 and 3). It is remarkable that all 18 polymerases, from the phylogenetic sub-tree of the subunit RPB1, have high identity score and remain undamaged during the evolution (Figures 1 and 2). The highly conserved motifs in protein families are directly related to their active sites and functionality (Koonin & Galperin 2003; Papageorgiou et al. 2016).

3D models A and B of the *Trypanosoma brucei brucei* DdRpII RPB1

Homologous solved 3D structures from the Protein Data Bank (PDB) have been identified from the Protein Data Bank (PDB) using the NCBI/BLASTp algorithm. Based on BLASTp report many 3D structures were determined suitable as templates for the homology modelling including the

crystal structure of the *Schizosaccharomyces pombe* DdRpII RPB1 (PDB: 3H0G) (Spahr et al. 2009), the crystal structure of the *Saccharomyces cerevisiae* DdRpII RPB1 (PDB: 4A3C and 1I3Q) (Cheung et al. 2011; Cramer et al. 2001), the electron microscopy structure *Bos taurus* DdRpII RPB1 (PDB: 5FLM) and the electron microscopy structure of the *Human* DdRpII RPB1 (PDB: 3J0K) (Bernecky et al. 2011). The final choice of a template structure was not only based on the percent sequence identity/similarity and the structure resolution, but also on the results of the phylogenetic trees. Two models were prepared. Model A was based on the *Schizosaccharomyces pombe* DdRpII RPB1 x-ray structure, while model B was based on the *Bos taurus* DdRpII RPB1 x-ray structure (Figure 3). Although the *Human* DdRpII RPB1 could also be used to build the *Trypanosoma brucei* DdRpII RPB1 3D model, it was avoided in an effort to minimize potential toxicity issues during the drug design process. Nonetheless, the sequence of the *Human* DdRpII and the corresponding sequence of the *Trypanosoma brucei* and *Bos taurus* were aligned in an effort to identify sequence-based differences and/or similarities for the modelling and drug design process (Supplementary Figure 2). A multiple sequence alignment was constructed including the *Trypanosoma brucei brucei* DdRpII RPB1 (NCBI: P17545.1) (Das et al. 2006), the *Trypanosoma brucei gambiense* DdRpII RPB1 (NCBI: XP_011773113.1) (Jackson et al. 2010), the crystal structure of *Schizosaccharomyces pombe* DdRpII RPB1 (PDB: 3H0G A chain) (Spahr et al. 2009), the crystal structure of *Saccharomyces cerevisiae* DdRpII RPB1 (PDB: 1I3Q A chain) (Cramer et al. 2001) , *Bos taurus* DdRpII RPB1 (PDB: 5FLM) (Bernecky et al. 2016). and the crystal structure of *Human* DdRpII RPB1 (PDB: 3J0K A chain) (Bernecky et al. 2011) towards to identify all the suggested conserved motifs within the highlighted domains of the RPB1 and the major sequences differences and similarities (Supplementary Figure 2).

The above-mentioned sequence alignments were used to identify all the nine canonical and conserved motifs as expected (Figures 2 and 3). The model of TPP DdRpII was first structurally superimposed and subsequently structurally compared to its template using the MOE software (Figure 4). The TPP DdRpII model exhibited an alpha-carbon RMSD lower than 1.3 angstroms (Figure 5 and Supplementary Data 8). Furthermore, the model was evaluated in regards to its geometry and its compatibility with the template structure using the build in protein check module of MOE (Supplementary Data 8). These results, confirmed the structural viability of the 3D *in silico* model.

Comparison of the *Trypanosoma brucei brucei* DdRPII RPB1 model A and model B.

It was decided to produce two models using the aforementioned template structures. Model A was build based on the *Schizosaccharomyces pombe* DdRpII RPB1 (PDB: 3H0G) X-ray structure and model B was based on the *Bos taurus* DdRpII RPB1 (PDB: 5FLM) structure. *Bos taurus* DdRPII RPB1 is a new released electron microscopy structure with 3.4 Å resolution, homolog to *Trypanosoma brucei brucei* DdRPII RPB1. The sequence alignment between the *Trypanosoma brucei brucei* DdRpII RPB1 and the *Bos taurus* DdRPII RPB1 template revealed 40% Identity and 56% similarity, same scores with the *Schizosaccharomyces pombe* DdRpII crystal structure, but the overall sequence alignment length was shorter than the *Schizosaccharomyces pombe* DdRpII crystal structure about 100 amino acids (Supplementary Figure 3). Furthermore, in the sequence alignment of the *Trypanosoma brucei* DdRpII RPB1 and *Bos taurus* DdRPII RPB1 all nine conserved motifs were identified, as expected. The root mean square deviation (RMSD) between model A and its template is 1.3 Å whereas the RMSD between model B and *Bos taurus*

template is 2.7 Å. Nevertheless, the overall RMSD between the two models and the two templates isn't bigger than 2,7 Å. (Figure 5 and Supplementary Data 8). Overall, we used to prepare in parallel a 3D model based on the *Bos taurus* structure as it bears better validation statistics and its sequence similarity to the *Trypanosoma brucei brucei* is higher. However, after performing another full coarse of MDs for model B, it was concluded that the added value of model B, when compared to model A is not significant ,as models A and B are quite similar indeed (Figures 7-11).

Discussion

Description of the *Trypanosoma brucei brucei* DdRPII RPB1 models.

RNA Polymerase II is a multi-subunit enzyme that transcribes protein-coding genes in eukaryotes (Sentenac 1985). Transcription in eukaryotes is dependent by three classes of nuclear RNA polymerases I-III. The genes encoding the largest subunits of eukaryotic RNA polymerases I, II and III have been isolated and are single copy genes, except *Trypanosoma* RNA polymerase II which contain two alleles (Smith et al. 1989). Structural and sequence differences between the two alleles are minor, but the C-terminal domain of those enzymes has a highly unusual structure. TBB DdRpII RPB1 model is the first protein subunit of the ten subunits multi-complex of RNA Polymerase II (Hahn 2004; Suh et al. 2013). The RPB1 subunit is very critical in RNA polymerase formation and function. The RPB1 active site and the RPB2 hybrid-binding region combine in a single fold that forms the active centre of the RpII (Figure 4). There are two metal ions at the RNA polymerase II active site. It has been previously reported that a Mg metal ion interacts with the three invariant aspartates of RPB1 (Cramer et al. 2001). The latter aspartate residues, which were found in all RPB1 sequences were aligned and fitted in a motifs exploration study. Consequently, those residues have now been marked as motif 4b in the TBB DdRpII RPB1 3D models.

The swinging motion of the clamp dictates the degree of opening of the cleft in DdRpII and permits the insertion of promoter DNA for the initiation of transcription (Suh et al. 2013). Based on previous studies, it is established that, upon closure of a transcribing complex, the RPB1 clamp serves as a multi-functional tool, sensing the DNA/RNA hybrid conformation and splitting DNA and RNA strands at the upstream end of the transcription complex (Cramer et al. 2001). The clamp is formed by N- and C-terminal regions of RPB1 and a part of the C-terminal region of RPB2 (Chen et al. 2007; Hahn 2004; Li et al. 2014). The clamp is primarily stabilized by three Zn ions within the RPB1 subunit (also marked in the TPP DdRpII RPB1) which forms zinc – finger conformations; two within the "clamp core" and one in the "clamp head". Accordingly, two Zinc-finger formations were identified and highlighted in the TBB DdRpII RPB1 model (Figure 6). The first formation can be recognized between a Zn ion and four cysteine residues in the suggested motif 1a, also known as CX(2)CXnCX2C/H (Das et al. 2006) (Figure 3). Mutations in the first Zn-finger formation confer a lethal phenotype of RNA polymerase II (Donaldson & Friesen 2000). The second Zinc –finger can be recognized in the next four cysteine residues (Figures 3 and 6). In the proposed motif 1b, the first two cysteine residues were identified, which constitute part of the second Zing finger formation. Finally, according to our molecular dynamics simulations, the main role of the Rpb1 and Rpb2 subunits is to provide stability within the overall structure formation of the RNA polymerase II molecule in the 3D space.

366
367 **3D Pharmacophore Elucidation**
368 3D Pharmacophore design techniques take into account both the three-dimensional structures
369 and binding modes of receptors and inhibitors towards identifying regions that are favorable or
370 not for a particular receptor-inhibitor interaction (Vlachakis & Kossida 2013). The description of
371 the receptor-inhibitor interaction pattern is determined through a correlation between the
372 specific properties of the inhibitors and their action on enzymatic activity (Balatsos et al. 2009;
373 Vlachakis et al. 2012). The pharmacophore for TBB DdRpII RPB1 (Figure 12) was based on
374 structural information from the enzyme's catalytic site including all steric and electronic
375 features that are necessary to ensure optimal non-covalent interactions. The pharmacophoric
376 features were investigated including positively or negatively ionized regions, hydrogen bond
377 donors and acceptors, aromatic regions and hydrophobic areas. Firstly, there should be one
378 electron-donating group in the proximity of the Ser1172 (colored green). The electron-donating
379 region indicates a particular property of the inhibitor and is not necessarily confined to a
380 specific chemical structure. Moreover, this interaction site may not strictly represent a
381 hydrogen bond, but water or ion mediated bridges since the distance from the catalytic amino
382 acids varies between 3-9 Å. An aromatic PAP (colored orange) was positioned in the proximity
383 of Phe1179, which established pi-stacking interactions. Two electron accepting PAPs (colored
384 red) were positioned in the proximity of the two Arginine residues (Arg1171 and Arg1203).
385 Finally, a set of two adjacent PAPs were positioned in the center of the active site, where the
386 Zn++ is coordinated in the crystal structure. Those yellow-colored PAPs are indicative of S-S
387 bonds and bridges or even S-C interactions, following the Michael acceptor moiety pattern. The
388 surrounding Cysteines are Cys1173, Cys1155, Cys1152, and Cys1270. However, the most
389 important factor of the latter PAPs was the optimal positioning of these groups in the 3D
390 conformational space of the TBB DdRpII RPB1 active site, rather than the amount of
391 conjugation or interaction with the protein.
392
393 **Conclusion**
394 The *Trypanosoma brucei brucei* DdRpII RPB1 enzyme was evolutionary analyzed, and nine new
395 conserved motifs were identified. Using the X-ray crystal structure of the *Schizosaccharomyces*
396 *pombe* DdRpII RPB1, the 3D model of the *Trypanosoma brucei brucei* DdRpII RPB1 was designed
397 using homology modelling techniques. The model was *in silico* evaluated and displayed high
398 conservation of the functional domains previously reported in other DdRpII subunit RPB1
399 species. The *Trypanosoma brucei brucei* DdRpII RPB1 model structure provides a basis for
400 interpretation of available data and the design of new experiments towards the *Trypanosoma*
401 *brucei brucei* inhibition. We, therefore, propose the use of the *Trypanosoma brucei brucei*
402 DdRpII RPB1 model A as a pharmacological targeting platform for advanced, *in silico* drug
403 design experiments using the novel findings of this study, both in the sequence and structural
404 level. The 3D models and sequence datasets that derived from this study will be made available
405 to the public, in an effort to pave the way for fellow scientists of multidiscipline backgrounds to
406 word in a synergic way towards the designing of novel anti-malarial agents with improved
407 biochemical and clinical characteristics in the future.
408
409

Abbreviations

DdRpII	DNA-directed RNA polymerase II
TBB	Trypanosoma brucei brucei
TBG	Trypanosoma brucei gambiense
TBR	Trypanosoma brucei rhodesiense
MOE	Molecular Operating Environment

Figures and Data legend

Figure 1: Phylogenetic reconstruction of *Trypanosoma brucei brucei* DdRpII RPB1 protein sequences. The tree was generated using the DdRpII family dataset (36 foul length protein sequences samples). The tree was constructed by Matlab Bioinformatics Toolbox utilizing Neighbour – Joining statistical method for 100 bootstrap replicates and visualized using MEGA cycle option. In the tree representation there are clearly separated in two monophyletic branches the RNA polymerases II subunits RPB1 (colored green) and RPB2 (colored blue). *Trypanosoma brucei* DdRpII RPB1 protein sequence was correctly classified and separated in the monophyletic sub-tree of the RPB1 group (highlight with red dots).

Figure 2: Representative conserved motifs for the DdRpII subunit RPB1. The nine suggested conserved motifs were extracted based on the multiple sequence alignment of the 18 protein sequences were classified and clearly separated in the DdRpII subunit RPB1 monophyletic sub-tree. The conserved motifs were identified through the consensus sequence and logo graph where generated using Jalview software.

Figure 3: Sequence alignment between the *Trypanosoma brucei brucei* DdRpII RPB1 and the corresponding sequence of the crystal structure of the *Schizosaccharomyces pombe* DdRpII RPB1. **(A)** Alignment of DdRpII RPB1 from *Trypanosoma brucei* DdRpII RPB1 (Labeled as "TB") with *Schizosaccharomyces pombe* DdRpII RPB1 (Labeled as "SB") was initially carried out with BLASTp and then manually adjusted. The nine suggested conserved motifs (Motifs 1a, 1b, 2, 3a, 3b, 3c, 4a, 4b, 4c) based on figure 2, domains and domain-like regions of *Trypanosoma brucei* DdRpII RPB1 represented in different colours. The amino acid residue numbers at the domain boundaries are indicated. Important structural elements and prominent regions involved in subunit interactions are also noted. Residues involved in the Zn and Mg coordination are highlighted in blue. **(B)** Domains and domain-like regions of the DdRpII subunit Rpb1. The amino acid residue numbers at the domain boundaries are indicated.

Figure 4: Model of the *Trypanosoma brucei brucei* DdRPII RPB1. **(A and B)** Ribbon representation of the produced *Trypanosoma brucei brucei* DdRPII RPB1 model (colored Orange) superposed with the corresponding *Schizosaccharomyces pombe* DdRpII RPB1 (in purple). **(C and D)** The nine suggested conserved motifs and the domains and domain-like regions of the *Trypanosoma brucei brucei* DdRPII RPB1. The motifs and RPB1 domains have been color-coded according to the Figures 2 and 3, and are shown in CPK format (Usual space filling). **(E and F)** Electrostatic surface potential for the *Trypanosoma brucei brucei* DdRPII RPB1. Represented with blue is the area of negative charge. Red is the area of positive charge and white is the un-charged region.

Figure 5: Structural superposition of the TBB DdRPII RPB1 models A and B. (A and B) Ribbon representation of the produced *Trypanosoma brucei brucei* DdRPII RPB1 model A (colored Orange) and model B (colored Blue) superposed with the corresponding *Schizosaccharomyces*

pombe DdRpII RPB1 (in Purple) and *Bos taurus* DdRpII RPB1 (in Grey). The four 3D structures are highly conserved in their active sites with few differences in the outer layer with overall RMSD 2.775 Å. **(C)** Ribbon representation of the produced *Trypanosoma brucei brucei* DdRPII RPB1 model A (colored Orange) superposed with the corresponding *Schizosaccharomyces pombe* DdRpII RPB1 (in purple). (RMSD = 1.242 Å). **(D)** Ribbon representation of the produced *Trypanosoma brucei brucei* DdRPII RPB1 model B (colored Blue) superposed with the *Bos taurus* DdRpII RPB1 (in Grey) respectively. (RMSD = 2.757 Å).

Figure 6: Zinc-finger formations in the *Trypanosoma brucei brucei* DdRpII RPB1 model. Ribbon representation of the produced *Trypanosoma brucei brucei* DdRPII RPB1 model. In the produced model were highlighted 3 main zing-finger domain formations (colored grey) were contained in the clam core, clam head and active site region. Domains and domain-like regions of the *Trypanosoma brucei brucei* DdRPII RPB1 have been color-coded according to conventions of Figures 3.

Figure 7: Molecular dynamics simulation charts for the *Trypanosoma brucei brucei* DdRpII RPB1 models. (A) The root mean square deviation (RMSD) of the model A during the time. **(B)** The root mean square fluctuation (RMSF) of the model A during the time. **(C)** The root mean square deviation (RMSD) of the model B during the time. **(D)** The root mean square fluctuation (RMSF) of the model B during the time.

Figure 8: Molecular dynamics simulation charts of the root mean square deviation (RMSD) for the *Trypanosoma brucei brucei* DdRpII RPB1 sub domains of the model A. The energy (Kcal/mol) vs time (ns) plot of the 100ns simulation trajectory of the TBB DdRpII RPBI model A. **(A)** Domain A RMSD. **(B)** Domain B RMSD. **(C)** Domain C RMSD. **(D)** Domain D RMSD. **(E)** Domain E RMSD.

Figure 9: Molecular dynamics simulation charts of the root mean square fluctuation (RMSF) for the *Trypanosoma brucei brucei* DdRpII RPB1 sub domains of the model A. (A) Domain A RMSF. **(B)** Domain B RMSF. **(C)** Domain C RMSF. **(D)** Domain D RMSF. **(E)** Domain E RMSF.

Figure 10: Molecular dynamics simulation charts of the root mean square deviation (RMSD) for the *Trypanosoma brucei brucei* DdRpII RPB1 sub domains of the model B. The energy (Kcal/mol) vs time (ns) plot of the 100ns simulation trajectory of the TBB DdRpII RPBI model B. **(A)** Domain A RMSD. **(B)** Domain B RMSD. **(C)** Domain C RMSD. **(D)** Domain D RMSD. **(E)** Domain E RMSD.

Figure 11: Molecular dynamics simulation charts of the root mean square fluctuation (RMSF) for the *Trypanosoma brucei brucei* DdRpII RPB1 sub domains of the model B. (A) Domain A RMSF. **(B)** Domain B RMSF. **(C)** Domain C RMSF. **(D)** Domain D RMSF. **(E)** Domain E RMSF.

Figure 12: The 3D pharmacophore model for the *Trypanosoma brucei brucei* DdRPII RPB1 model. In total 5 distinct pharmacophoric features were identified. An aromatic region (colored

orange), an electron donating region (colored green), two electron accepting regions (colored red) and a sulphur specific S-S interacting region (colored yellow).

Supplementary Figure 1: Phylogenetic reconstruction of *Trypanosoma brucei brucei* DdRPII RPB1 model DdRpII RPB1 protein sequences. The tree was generated using the DdRpII family dataset (36 foul length protein sequences samples) and the Jalview software. Tree was constructed using the average distance statistical method with PAM 250. In the tree representation there are clearly shown the two RNA polymerases II subunits RPB1 and RPB2 as two main monophyletic sub-trees. *Trypanosoma brucei* DdRpII RPB1 protein sequence was correctly classified in the monophyletic sub-tree of the RPB1 group.

Supplementary Figure 2: Multiple sequence alignment. The alignment was performed using the *Trypanosoma brucei brucei* DdRPII RPB1, the *Trypanosoma brucei gambiense* DdRpII RPB1, the crystal structure of *Schizosaccharomyces pombe* DdRpII RPB, the crystal structure of *Saccharomyces cerevisiae* DdRpII RPB1 and the electron microscopy structure of Human DdRpII DdRpII RPB1. **(A)** All nine suggested conserved motifs and major domains of DdRpII RPB1 have been marked (Motifs 1a, 1b, 2, 3a, 3b, 3c, 4a, 4b, 4c). Additionally, in the multiple sequence alignment were presented the major differences. **(B)** Domains and domainlike regions of the DdRpII subunit Rpb1. The amino acid residue numbers at the domain boundaries are indicated.

Supplementary Figure 3: Multiple sequence alignment. The alignment was performed using the *Trypanosoma brucei brucei* DdRPII RPB1, the crystal structure of *Schizosaccharomyces pombe* DdRpII RPB and the electron microscopy structure of *Bos taurus* DdRpII RPB1. All five sub-domains as referred in Pfam database have been marked with different colours.

Supplementary Data 1: DdRPII related proteins dataset.

Supplementary Data 2: MEGA software phylogenetic tree in newick format. The tree was constructed the Neighbour – Joining statistical method for 100 bootstrap replicates and the 36 extracted samples of the DpRpII.

Supplementary Data 3: Jalview software phylogenetic tree in newick format. The tree was constructed using the average distances statistical method and the 36 extracted samples of the DpRpII.

Supplementary Data 4: *Trypanosoma brucei brucei* DdRPII RPB1 model A in .pdb format.

Supplementary Data 5: *Trypanosoma brucei brucei* DdRPII RPB1 model B in .pdb format.

Supplementary Data 6: Protein structure report of the template.

Supplementary Data 7: Protein structure report of the model.

Supplementary Data 8: Protein structure report of the superposed models and templates.

548
549

References

Balatsos NA, Vlachakis D, Maragozidis P, Manta S, Anastasakis D, Kyritsis A, Vlassi M, Komiotis D, and Stathopoulos C. 2009. Competitive inhibition of human poly(A)-specific ribonuclease (PARN) by synthetic fluoro-pyranosyl nucleosides. *Biochemistry* 48:6044-6051. 10.1021/bi900236k

Bayele HK. 2009. Trypanosoma brucei: a putative RNA polymerase II promoter. *Exp Parasitol* 123:313-318. 10.1016/j.exppara.2009.08.007

S0014-4894(09)00233-1 [pii]

Berg JM, Tymoczko JL, and Stryer L. 2002. *Biochemistry*. New York: W.H. Freeman.

Bernecky C, Grob P, Ebmeier CC, Nogales E, and Taatjes DJ. 2011. Molecular architecture of the human Mediator-RNA polymerase II-TFIIF assembly. *PLoS Biol* 9:e1000603. 10.1371/journal.pbio.1000603

Bernecky C, Herzog F, Baumeister W, Plitzko JM, and Cramer P. 2016. Structure of transcribing mammalian RNA polymerase II. *Nature* 529:551-554. 10.1038/nature16482

nature16482 [pii]

Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, Bohme U, Hannick L, Aslett MA, Shallom J, Marcello L, Hou L, Wickstead B, Alsmark UC, Arrowsmith C, Atkin RJ, Barron AJ, Bringaud F, Brooks K, Carrington M, Cherevach I, Chillingworth TJ, Churcher C, Clark LN, Corton CH, Cronin A, Davies RM, Doggett J, Djikeng A, Feldblyum T, Field MC, Fraser A, Goodhead I, Hance Z, Harper D, Harris BR, Hauser H, Hostetler J, Ivens A, Jagels K, Johnson D, Johnson J, Jones K, Kerhornou AX, Koo H, Larke N, Landfear S, Larkin C, Leech V, Line A, Lord A, Macleod A, Mooney PJ, Moule S, Martin DM, Morgan GW, Mungall K, Norbertczak H, Ormond D, Pai G, Peacock CS, Peterson J, Quail MA, Rabbinowitsch E, Rajandream MA, Reitter C, Salzberg SL, Sanders M, Schobel S, Sharp S, Simmonds M, Simpson AJ, Tallon L, Turner CM, Tait A, Tivey AR, Van Aken S, Walker D, Wanless D, Wang S, White B, White O, Whitehead S, Woodward J, Wortman J, Adams MD, Embley TM, Gull K, Ullu E, Barry JD, Fairlamb AH, Opperdoes F, Barrell BG, Donelson JE, Hall N, Fraser CM, Melville SE, and El-Sayed NM. 2005. The genome of the African trypanosome Trypanosoma brucei. *Science* 309:416-422. 10.1126/science.1112642

Chakravorty DK, and Merz KM, Jr. 2014. Studying allosteric regulation in metal sensor proteins using computational methods. *Adv Protein Chem Struct Biol* 96:181-218. 10.1016/bs.apcsb.2014.06.009

S1876-1623(14)00010-8 [pii]

Chen HT, Warfield L, and Hahn S. 2007. The positions of TFIIF and TFIIE in the RNA polymerase II transcription preinitiation complex. *Nat Struct Mol Biol* 14:696-703. nsmb1272 [pii]

10.1038/nsmb1272

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, and Thompson JD. 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* 31:3497-3500.

Cheung AC, Sainsbury S, and Cramer P. 2011. Structural basis of initial RNA polymerase II transcription. *EMBO J* 30:4755-4763. 10.1038/emboj.2011.396

emboj2011396 [pii]

Chung HM, Lee MG, Dietrich P, Huang J, and Van der Ploeg LH. 1993. Disruption of largest subunit RNA polymerase II genes in Trypanosoma brucei. *Mol Cell Biol* 13:3734-3743.

Cramer P, Bushnell DA, and Kornberg RD. 2001. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* 292:1863-1876. 10.1126/science.1059493

1059493 [pii]

Darriba D, Taboada GL, Doallo R, and Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27:1164-1165. 10.1093/bioinformatics/btr088

598 btr088 [pii]
599 Das A, Li H, Liu T, and Bellofatto V. 2006. Biochemical characterization of Trypanosoma brucei RNA
600 polymerase II. *Mol Biochem Parasitol* 150:201-210. S0166-6851(06)00232-5 [pii]
601 10.1016/j.molbiopara.2006.08.002
602 Donaldson IM, and Friesen JD. 2000. Zinc stoichiometry of yeast RNA polymerase II and characterization
603 of mutations in the zinc-binding domain of the largest subunit. *J Biol Chem* 275:13780-13788.
604 275/18/13780 [pii]
605 Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic*
606 *Acids Res* 32:1792-1797. 10.1093/nar/gkh340
607 32/5/1792 [pii]
608 Evers R, Hammer A, Kock J, Jess W, Borst P, Memet S, and Cornelissen AW. 1989. Trypanosoma brucei
609 contains two RNA polymerase II largest subunit genes with an altered C-terminal domain. *Cell*
610 56:585-597. 0092-8674(89)90581-3 [pii]
611 Hahn S. 2004. Structure and mechanism of the RNA polymerase II transcription machinery. *Nat Struct*
612 *Mol Biol* 11:394-403. 10.1038/nsmb763
613 nsmb763 [pii]
614 Illergard K, Ardell DH, and Elofsson A. 2009. Structure is three to ten times more conserved than
615 sequence--a study of structural response in protein cores. *Proteins* 77:499-508.
616 10.1002/prot.22458
617 Jackson AP, Sanders M, Berry A, McQuillan J, Aslett MA, Quail MA, Chukualim B, Capewell P, MacLeod A,
618 Melville SE, Gibson W, Barry JD, Berriman M, and Hertz-Fowler C. 2010. The genome sequence
619 of Trypanosoma brucei gambiense, causative agent of chronic human african trypanosomiasis.
620 *PLoS Negl Trop Dis* 4:e658. 10.1371/journal.pntd.0000658
621 Koch H, Raabe M, Urlaub H, Bindereif A, and Preusser C. 2016. The polyadenylation complex of
622 Trypanosoma brucei: Characterization of the functional poly(A) polymerase. *RNA Biol* 13:221-
623 231. 10.1080/15476286.2015.1130208
624 Koonin EV, and Galperin MY. 2003. *Sequence - evolution - function : computational approaches in*
625 *comparative genomics*. Boston: Kluwer Academic.
626 Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, and Thornton JM. 1996. AQUA and PROCHECK-
627 NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR*
628 8:477-486.
629 Li W, Giles C, and Li S. 2014. Insights into how Spt5 functions in transcription elongation and repressing
630 transcription coupled DNA repair. *Nucleic Acids Res* 42:7069-7083. 10.1093/nar/gku333
631 gku333 [pii]
632 Loukatou S, Papageorgiou L, Fakourelis P, Filntisi A, Polychronidou E, Bassis I, Megalooikonomou V,
633 Makalowski W, Vlachakis D, and Kossida S. 2014. Molecular dynamics simulations through GPU
634 video games technologies. *J Mol Biochem* 3:64-71.
635 Malvy D, and Chappuis F. 2011. Sleeping sickness. *Clin Microbiol Infect* 17:986-995. 10.1111/j.1469-
636 0691.2011.03536.x
637 S1198-743X(14)61376-8 [pii]
638 Nayeem A, Sitkoff D, and Krystek S, Jr. 2006. A comparative study of available software for high-accuracy
639 homology modeling: from sequence alignments to structural models. *Protein Sci* 15:808-824.
640 10.1110/ps.051892906
641 Papageorgiou L, Loukatou S, Koumandou VL, Makalowski W, Megalooikonomou V, Vlachakis D, and
642 Kossida S. 2014. Structural models for the design of novel antiviral agents against Greek Goat
643 Encephalitis. *PeerJ* 2:e664. 10.7717/peerj.664

644 Papageorgiou L, Loukatou S, Sofia K, Maroulis D, and Vlachakis D. 2016. An updated evolutionary study
645 of Flaviviridae NS3 helicase and NS5 RNA-dependent RNA polymerase reveals novel invariable
646 motifs as potential pharmacological targets. *Mol Biosyst*. 10.1039/c5mb00706b
647 Ridley RG. 2002. Medical need, scientific opportunity and the drive for antimalarial drugs. *Nature*
648 415:686-693. 10.1038/415686a
649 415686a [pii]
650 Ross L, Lim ML, Liao Q, Wine B, Rodriguez AE, Weinberg W, and Shaefer M. 2007. Prevalence of
651 antiretroviral drug resistance and resistance-associated mutations in antiretroviral therapy-
652 naive HIV-infected individuals from 40 United States cities. *HIV Clin Trials* 8:1-8.
653 L61238775J512643 [pii]
654 10.1310/hct0801-1
655 Seeliger D, and de Groot BL. 2010. Ligand docking and binding site analysis with PyMOL and
656 Autodock/Vina. *J Comput Aided Mol Des* 24:417-422. 10.1007/s10822-010-9352-6
657 Sentenac A. 1985. Eukaryotic RNA polymerases. *CRC Crit Rev Biochem* 18:31-90.
658 Smith JL, Levin JR, Ingles CJ, and Agabian N. 1989. In trypanosomes the homolog of the largest subunit of
659 RNA polymerase II is encoded by two genes and has a highly unusual C-terminal domain
660 structure. *Cell* 56:815-827.
661 Spahr H, Calero G, Bushnell DA, and Kornberg RD. 2009. Schizosacharomyces pombe RNA polymerase II
662 at 3.6-A resolution. *Proc Natl Acad Sci U S A* 106:9185-9190. 10.1073/pnas.0903361106
663 Stecher G, Liu L, Sanderford M, Peterson D, Tamura K, and Kumar S. 2014. MEGA-MD: molecular
664 evolutionary genetics analysis software with mutational diagnosis of amino acid variation.
665 *Bioinformatics* 30:1305-1307. 10.1093/bioinformatics/btu018
666 btu018 [pii]
667 Suh H, Hazelbaker DZ, Soares LM, and Buratowski S. 2013. The C-terminal domain of Rpb1 functions on
668 other RNA polymerase II subunits. *Mol Cell* 51:850-858. 10.1016/j.molcel.2013.08.015
669 S1097-2765(13)00588-1 [pii]
670 Temiz AN, Benos PV, and Camacho CJ. 2010. Electrostatic hot spot on DNA-binding domains mediates
671 phosphate desolvation and the pre-organization of specificity determinant side chains. *Nucleic*
672 *Acids Res* 38:2134-2144. 10.1093/nar/gkp1132
673 gkp1132 [pii]
674 Thompson JD, Higgins DG, and Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive
675 multiple sequence alignment through sequence weighting, position-specific gap penalties and
676 weight matrix choice. *Nucleic Acids Res* 22:4673-4680.
677 Trouiller P, Olliaro P, Torreele E, Orbinski J, Laing R, and Ford N. 2002. Drug development for neglected
678 diseases: a deficient market and a public-health policy failure. *Lancet* 359:2188-2194. S0140-
679 6736(02)09096-7 [pii]
680 10.1016/S0140-6736(02)09096-7
681 Vilar S, Cozza G, and Moro S. 2008. Medicinal chemistry and the molecular operating environment
682 (MOE): application of QSAR and molecular docking to drug discovery. *Curr Top Med Chem*
683 8:1555-1572.
684 Vlachakis D. 2009. Theoretical study of the Usutu virus helicase 3D structure, by means of computer-
685 aided homology modelling. *Theor Biol Med Model* 6:9. 10.1186/1742-4682-6-9
686 1742-4682-6-9 [pii]
687 Vlachakis D, Bencurova E, Papangelopoulos N, and Kossida S. 2014a. Current state-of-the-art molecular
688 dynamics methods and applications. *Adv Protein Chem Struct Biol* 94:269-313. 10.1016/B978-0-
689 12-800168-4.00007-X
690 B978-0-12-800168-4.00007-X [pii]

691 Vlachakis D, Fakourelis P, Megalooikonomou V, Makris C, and Kossida S. 2015. DrugOn: a fully integrated
692 pharmacophore modeling and structure optimization toolkit. *PeerJ* 3:e725. 10.7717/peerj.725
693 725 [pii]
694 Vlachakis D, Kontopoulos DG, and Kossida S. 2013a. Space constrained homology modelling: the
695 paradigm of the RNA-dependent RNA polymerase of dengue (type II) virus. *Comput Math*
696 *Methods Med* 2013:108910. 10.1155/2013/108910
697 Vlachakis D, and Kossida S. 2013. Molecular modeling and pharmacophore elucidation study of the
698 Classical Swine Fever virus helicase as a promising pharmacological target. *PeerJ* 1:e85.
699 10.7717/peerj.85
700 Vlachakis D, Koumandou VL, and Kossida S. 2013b. A holistic evolutionary and structural study of
701 flaviviridae provides insights into the function and inhibition of HCV helicase. *PeerJ* 1:e74.
702 10.7717/peerj.74
703 Vlachakis D, Pavlopoulou A, Roubelakis MG, Feidakis C, Anagnou NP, and Kossida S. 2014b. 3D molecular
704 modeling and evolutionary study of the Trypanosoma brucei DNA Topoisomerase IB, as a new
705 emerging pharmacological target. *Genomics* 103:107-113. 10.1016/j.ygeno.2013.11.008
706 S0888-7543(13)00224-3 [pii]
707 Vlachakis D, Pavlopoulou A, Tsiliki G, Komiotis D, Stathopoulos C, Balatsos NA, and Kossida S. 2012. An
708 integrated in silico approach to design specific inhibitors targeting human poly(a)-specific
709 ribonuclease. *PLoS One* 7:e51113. 10.1371/journal.pone.0051113
710 PONE-D-12-20013 [pii]
711 Wang J, Cieplak P, and Kollman P. 2000. How well does a restrained electrostatic potential (resp) model
712 perform in calculating conformational energies of organic and biological molecules. *J Comp*
713 *Chem* 21:1049-1071.
714 Waterhouse AM, Procter JB, Martin DM, Clamp M, and Barton GJ. 2009. Jalview Version 2--a multiple
715 sequence alignment editor and analysis workbench. *Bioinformatics* 25:1189-1191.
716 10.1093/bioinformatics/btp033
717 btp033 [pii]
718 World Health Organization. 2015. Human African trypanosomiasis
719
720

Figure 1

Phylogenetic reconstruction of *Trypanosoma brucei brucei* DdRpII RPB1 protein sequences.

The tree was generated using the DdRpII family dataset (36 foul length protein sequences samples). The tree was constructed by Matlab Bioinformatics Toolbox utilizing Neighbour – Joining statistical method for 100 bootstrap replicates and visualized using MEGA cycle option. In the tree representation there are clearly separated in two monophyletic branches the RNA polymerases II subunits RPB1 (colored green) and RPB2 (colored blue). *Trypanosoma brucei* DdRpII RPB1 protein sequence was correctly classified and separated in the monophyletic sub-tree of the RPB1 group (highlight with red dots).

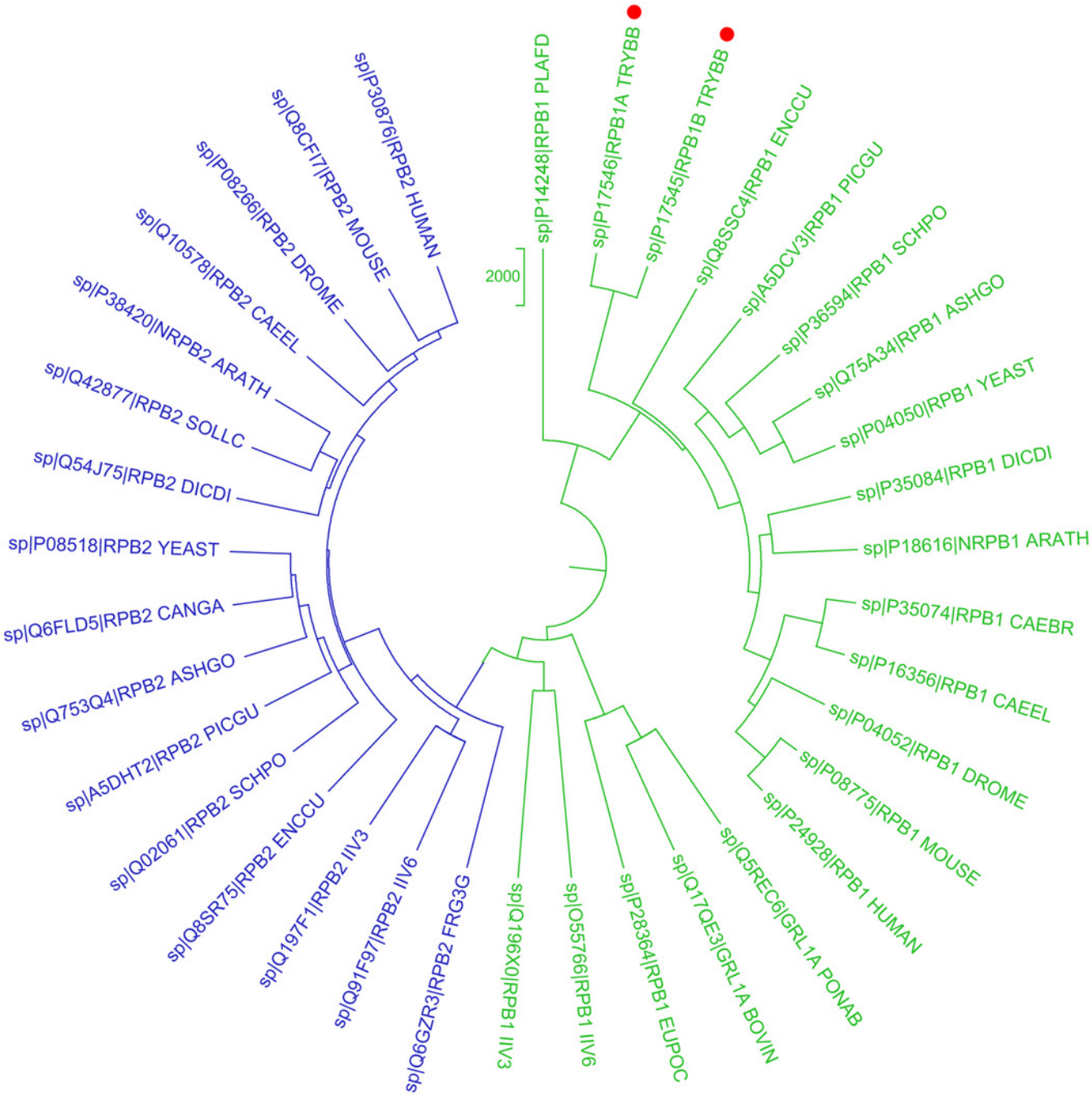
sp|P14248|RPB1 PLAFD
sp|P17546|RPB1A TRYBB
sp|P17545|RPB1B TRYBB
sp|Q8SSC4|RPB1 ENCCU
sp|A5DCV3|RPB1 PICGU
sp|P36594|RPB1 SCHPO
sp|Q75A34|RPB1 ASHGO
sp|P04050|RPB1 YEAST
sp|P35084|RPB1 DICDI
sp|P18616|NRPB1 ARATH
sp|P35074|RPB1 CAEBR
sp|P16356|RPB1 CAEEL
sp|P04052|RPB1 DROME
sp|P08775|RPB1 MOUSE
sp|P24928|RPB1 HUMAN
sp|Q5REC6|GRL1A PONAB
sp|Q17QE3|GRL1A BOVIN
sp|P28364|RPB1 EUPOC
sp|O55766|RPB1 IIV6
sp|Q196X0|RPB1 IIV3
sp|Q6GZR3|RPB2 FRG3G
sp|Q91F97|RPB2 IIV6
sp|Q197F1|RPB2 IIV3
sp|Q8SR75|RPB2 ENCCU
sp|Q02061|RPB2 SCHPO
sp|A5DHT2|RPB2 PICGU
sp|Q753Q4|RPB2 ASHGO
sp|Q6FLD5|RPB2 CANGA
sp|P08518|RPB2 YEAST
sp|Q54J75|RPB2 DICDI
sp|Q42877|RPB2 SOLLC
sp|P38420|NRPB2 ARATH
sp|Q10578|RPB2 CAEEL
sp|P08266|RPB2 DROME
sp|Q8CFI7|RPB2 MOUSE
sp|P30876|RPB2 HUMAN
2000

Figure 2(on next page)

Representative conserved motifs for the DdRpIIsubunit RPB1.

The nine suggested conserved motifs were extracted based on the multiple sequence alignment of the 18 protein sequences were classified and clearly separated in the DdRpII subunit RPB1 monophyletic sub-tree. The conserved motifs were identified through the consensus sequence and logo graph where generated using Jalview software.

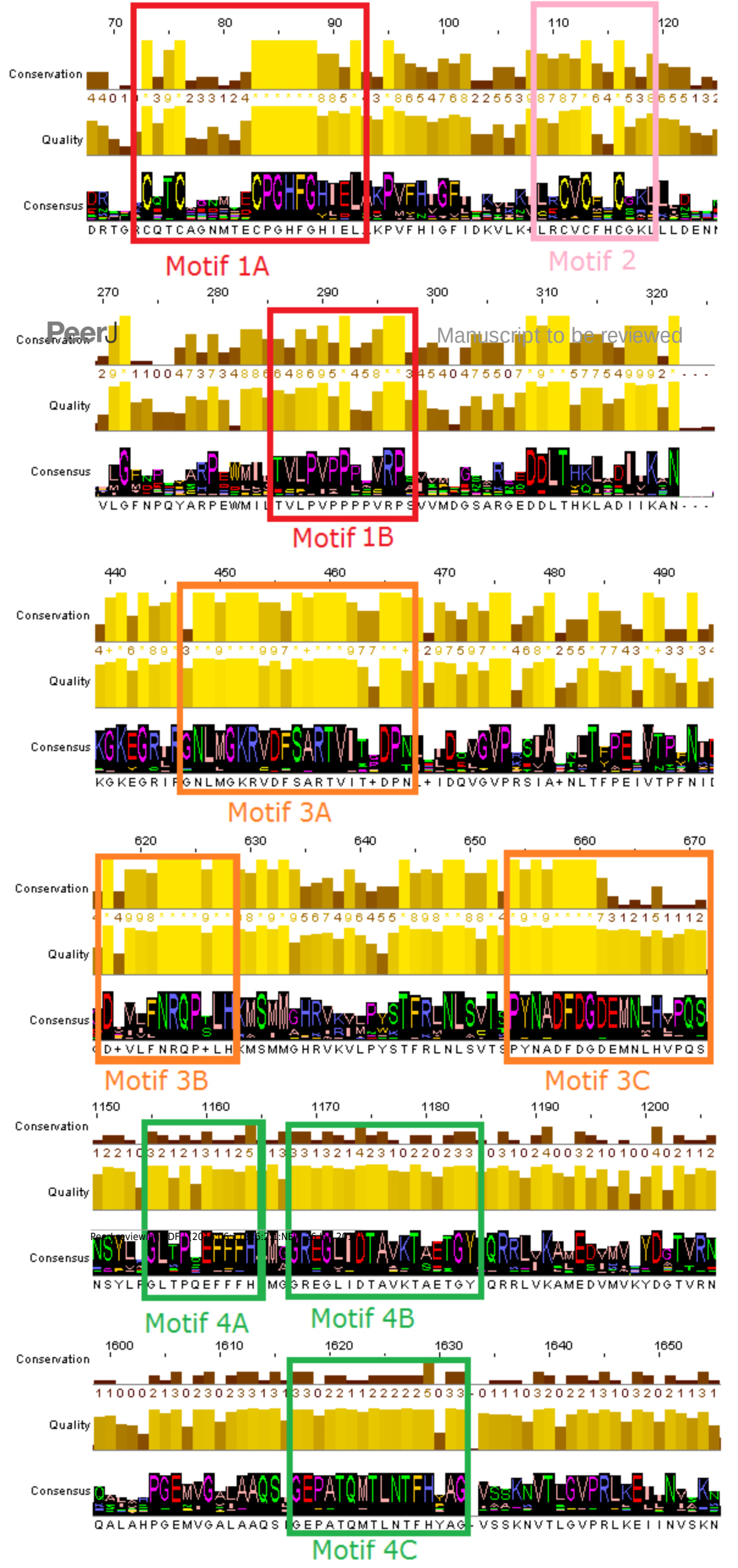

Figure 3

Sequence alignment between the *Trypanosoma brucei brucei* DdRpII RPB1 and the corresponding sequence of the crystal structure of the *Schizosaccharomyces pombe* DdRpII RPB1.

(A) Alignment of DdRpII RPB1 from *Trypanosoma brucei* DdRpII RPB1 (Labeled as "TB") with *Schizosaccharomyces pombe* DdRpII RPB1 (Labeled as "SB") was initially carried out with BLASTp and then manually adjusted. The nine suggested conserved motifs (Motifs 1a, 1b, 2, 3a, 3b, 3c, 4a, 4b, 4c) based on figure 2, domains and domain-like regions of *Trypanosoma brucei* DdRpII RPB1 represented in different colours. The amino acid residue numbers at the domain boundaries are indicated. Important structural elements and prominent regions involved in subunit interactions are also noted. Residues involved in the Zn and Mg coordination are highlighted in blue. **(B)** Domains and domain-like regions of the DdRpII subunit Rpb1. The amino acid residue numbers at the domain boundaries are indicated.

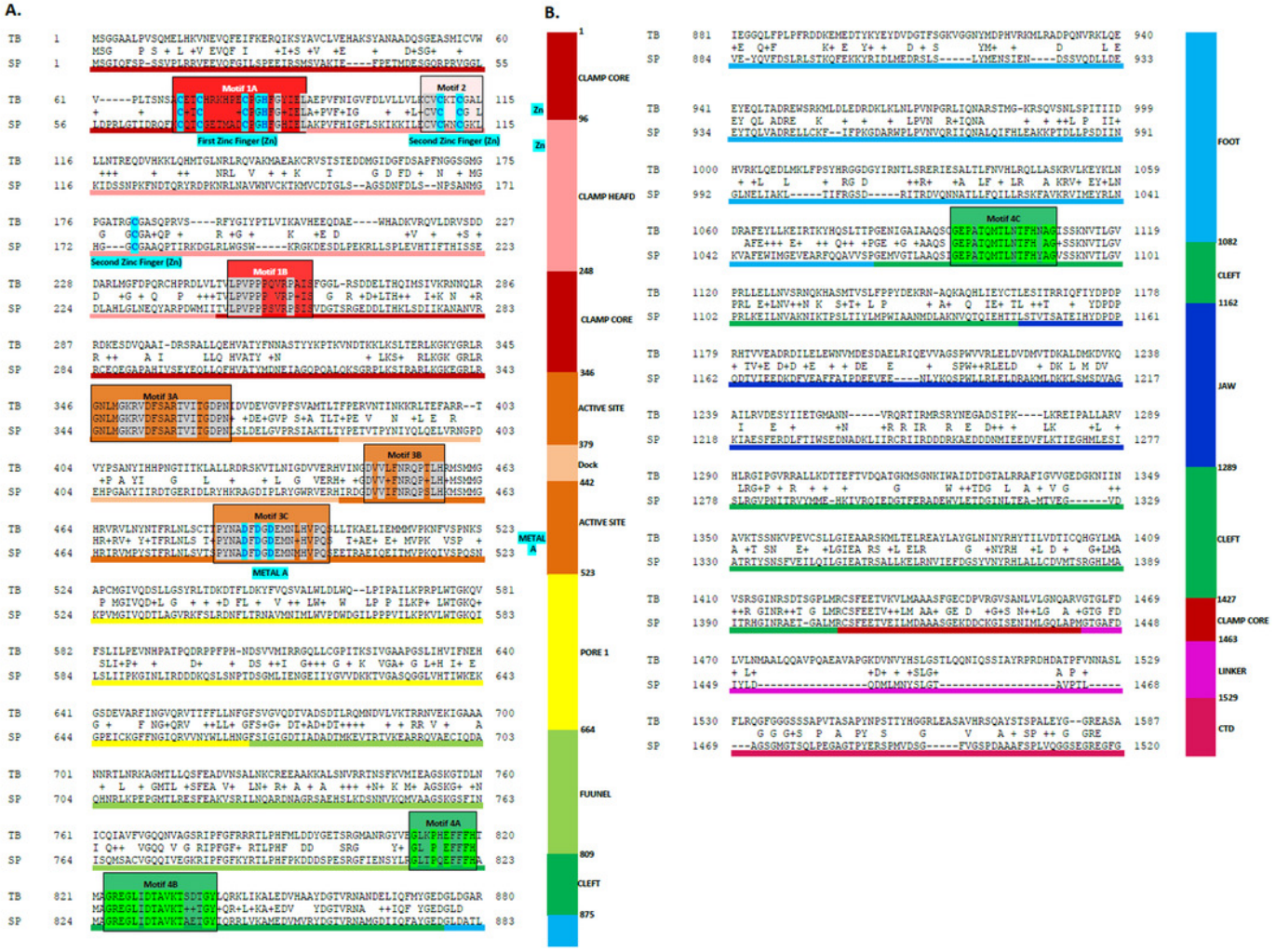
A.
B.
Motif 1A
Motif 2
First Zinc Finger (Zn)
Second Zinc Finger (Zn)
Motif 1B
Second Zinc Finger (Zn)
Motif 3A
Motif 3B
Motif 3C
METAL A
Motif 4A
Motif 4B
Motif 4C
CLAMP CORE
CLAMP HEAD
ACTIVE SITE
Dock
ACTIVE SITE
PORE 1
FUUNEL
CLEFT
Zn
Zn
METAL A
FOOT
CLEFT
JAW
CLEFT
CLAMP CORE
LINKER
CTD

Figure 4

Model of the *Trypanosoma brucei brucei* DdRPII RPB1 .

(A and B) Ribbon representation of the produced *Trypanosoma brucei brucei* DdRPII RPB1 model (colored Orange) superposed with the corresponding *Schizosaccharomyces pombe* DdRpII RPB1 (in purple). **(C and D)** The nine suggested conserved motifs and the domains and domain-like regions of the *Trypanosoma brucei brucei* DdRPII RPB1 . The motifs and RPB1 domains have been color-coded according to the Figures 2 and 3, and are shown in CPK format (Usual space filling). **(E and F)** Electrostatic surface potential for the *Trypanosoma brucei brucei* DdRPII RPB1 . Represented with blue is the area of negative charge. Red is the area of positive charge and white is the un-charged region.

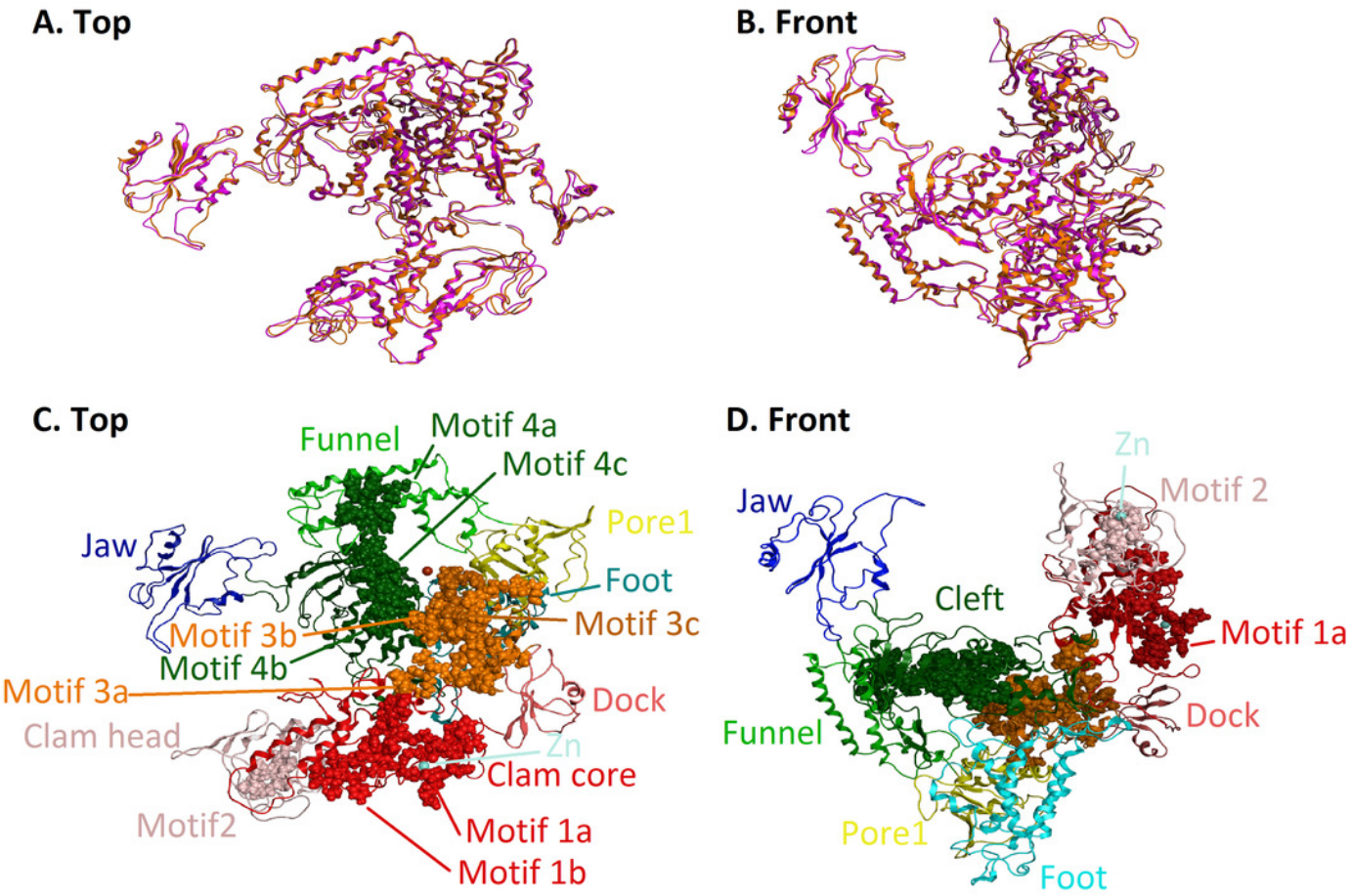
A. Top
B. Front
C. Top
Funnel
Motif 4a
Motif 4c
Pore1
Jaw
Foot
Motif 3c
Motif 3b
Motif 4b
Motif 3a
Dock
Clam head
Zn
Clam core
Motif2
Motif 1a
Motif 1b
D. Front
Zn
Motif 2
Jaw
Cleft
Motif 1a
Dock
Funnel
Pore1
Foot
E. Top
F. Front

Figure 5

Structural superposition ofthe TBB DdRPII RPB1 models A and B

(A and B) Ribbon representation of the produced *Trypanosoma brucei brucei* DdRPII RPB1 model A (colored Orange) and model B (colored Blue) superposed with the corresponding *Schizosaccharomyces pombe* DdRpII RPB1 (in Purple) and *Bos taurus* DdRpII RPB1 (in Grey) . The four 3D structures are highly conserved in their active sites with few differences in the outer layer with overall RMSD 2.775 Å . **(C)** Ribbon representation of the produced *Trypanosoma brucei brucei* DdRPII RPB1 model A (colored Orange) superposed with the corresponding *Schizosaccharomyces pombe* DdRpII RPB1 (in purple). (RMSD = 1.242 Å). **(D)** Ribbon representation of the produced *Trypanosoma brucei brucei* DdRPII RPB1 model B (colored Blue) superposed with the *Bos taurus* DdRpII RPB1 (in Grey) respectively. (RMSD = 2.757 Å).

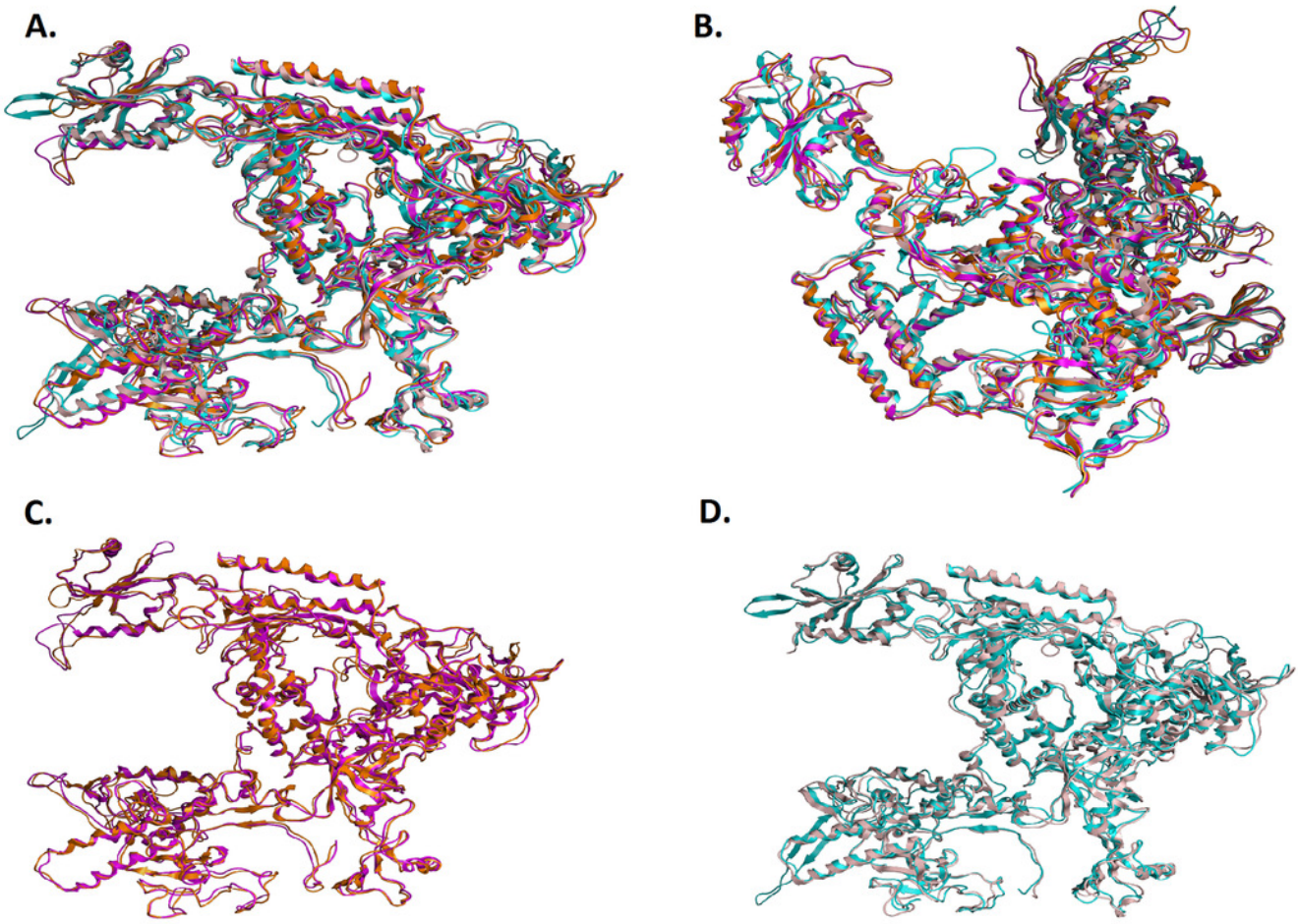
A.
B.
C.
D.

Figure 6

Zinc-finger formationsin the *Trypanosoma brucei brucei* DdRpII RPB1 model.

Ribbon representation of the produced *Trypanosoma brucei brucei* DdRPII RPB1 model. In the produced model were highlighted 3 main zing-finger domain formations (colored grey) were contained in the clam core, clam head and active site region. Domains and domain-like regions of the *Trypanosoma brucei brucei* DdRPII RPB1 have been color-coded according to conventions of Figures 3.

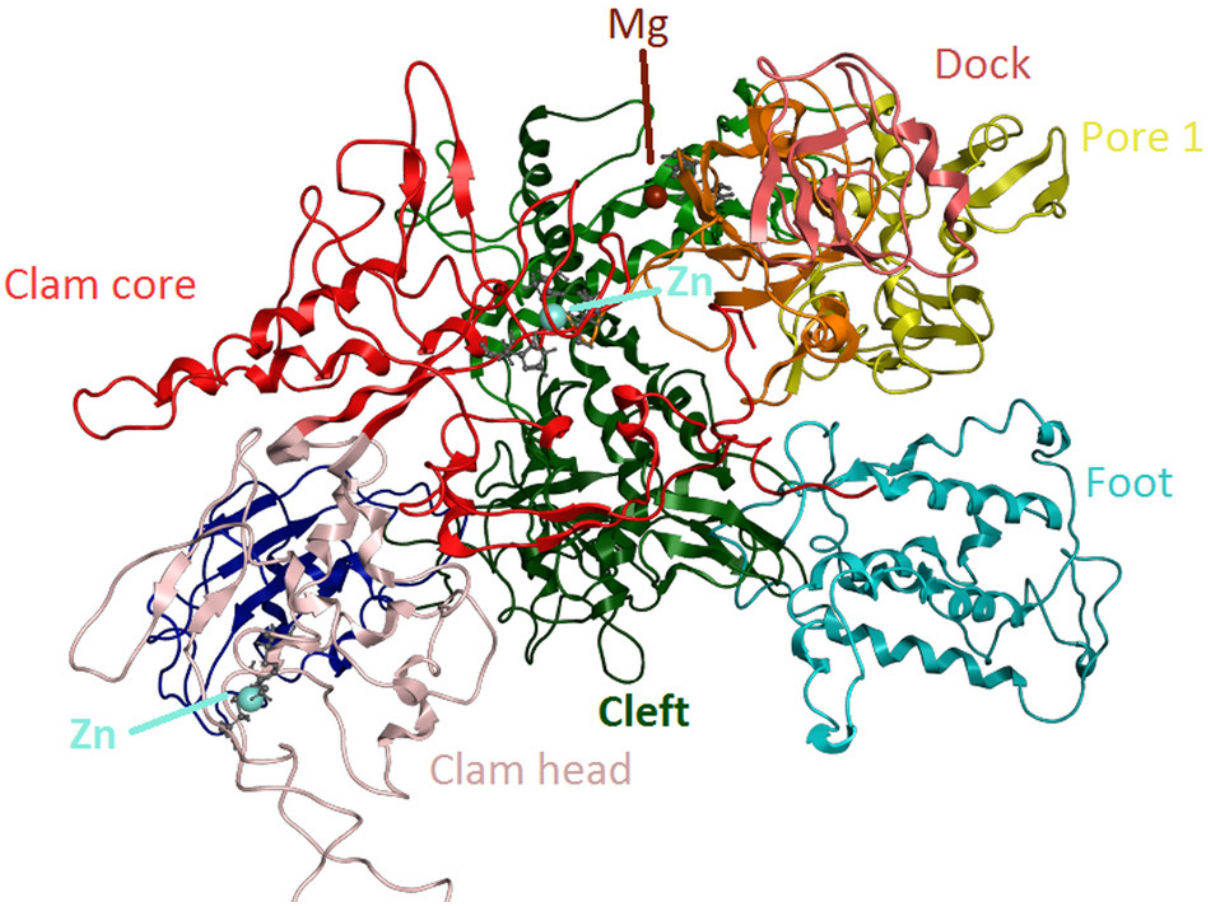

Figure 7

Molecular dynamics simulationcharts for the *Trypanosoma brucei brucei* DdRpII RPB1 models.

(A) The root mean square deviation (RMSD) of the model A during the time. **(B)** The root mean square fluctuation (RMSF) of the model A during the time. **(C)** The root mean square deviation (RMSD) of the model B during the time. **(D)** The root mean square fluctuation (RMSF) of the model B during the time.

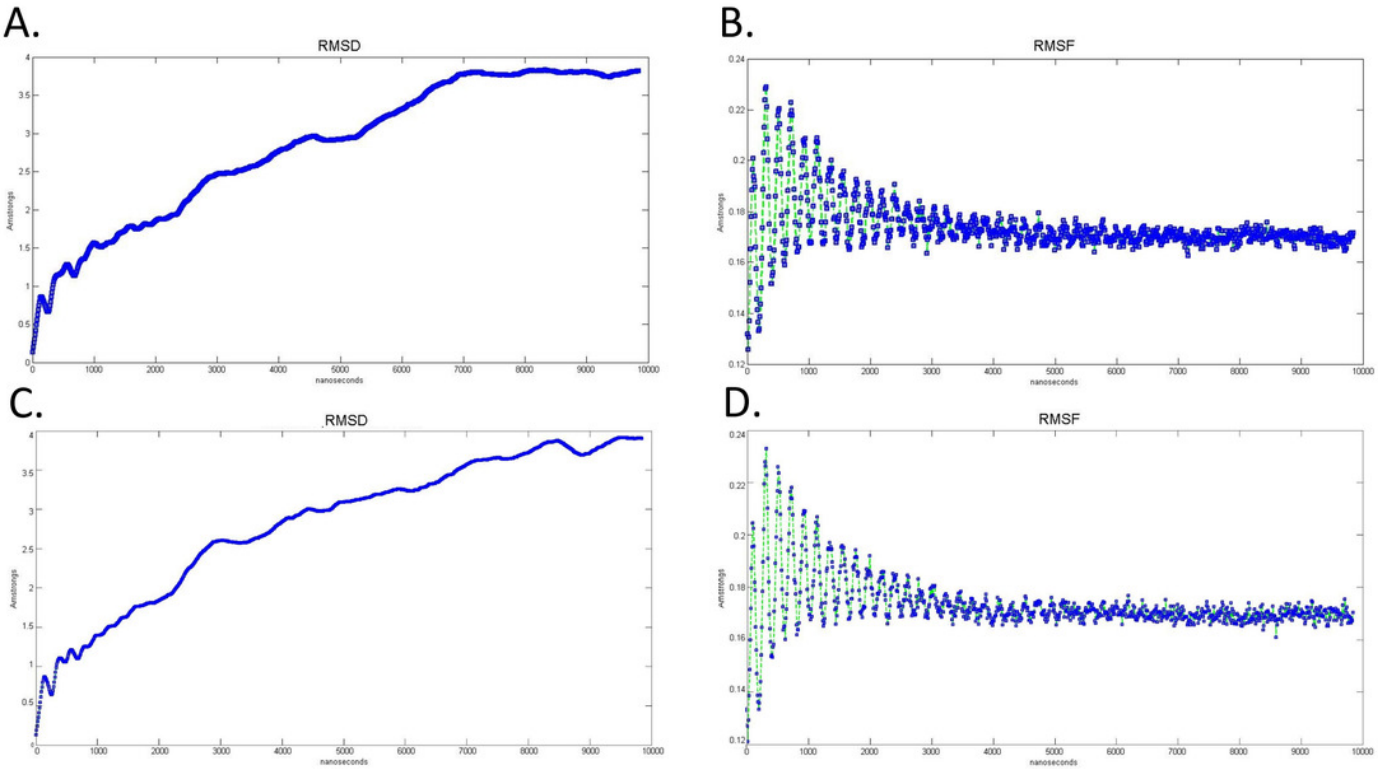

Figure 8

Molecular dynamics simulationcharts of the root mean square deviation (RMSD) for the *Trypanosoma brucei brucei* DdRpII RPB1 sub domainsof the model A.

The energy (Kcal/mol) vs time (ns) plot of the 100ns simulation trajectory of the TBB DdRpII RPBI model A. **(A)** Domain A RMSD. **(B)** Domain B RMSD. **(C)** Domain C RMSD. **(D)** Domain D RMSD. **(E)** Domain E RMSD.

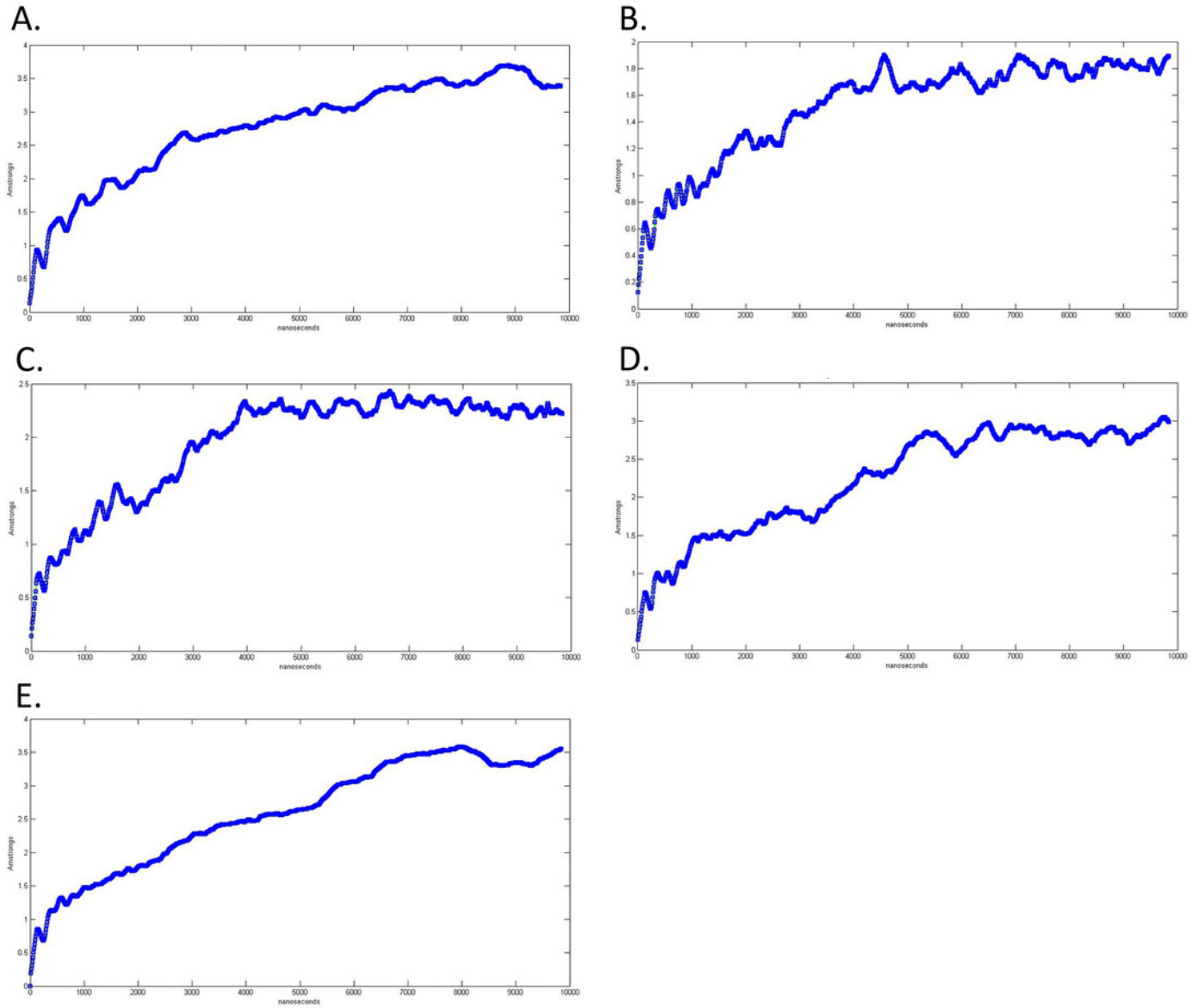

Figure 9

Molecular dynamics simulation charts of the root mean square fluctuation (RMSF) for the *Trypanosoma brucei brucei* DdRpII RPB1 sub domains of the model A.

(A) Domain A RMSF. **(B)** Domain B RMSF. **(C)** Domain C RMSF. **(D)** Domain D RMSF. **(E)** Domain E RMSF.

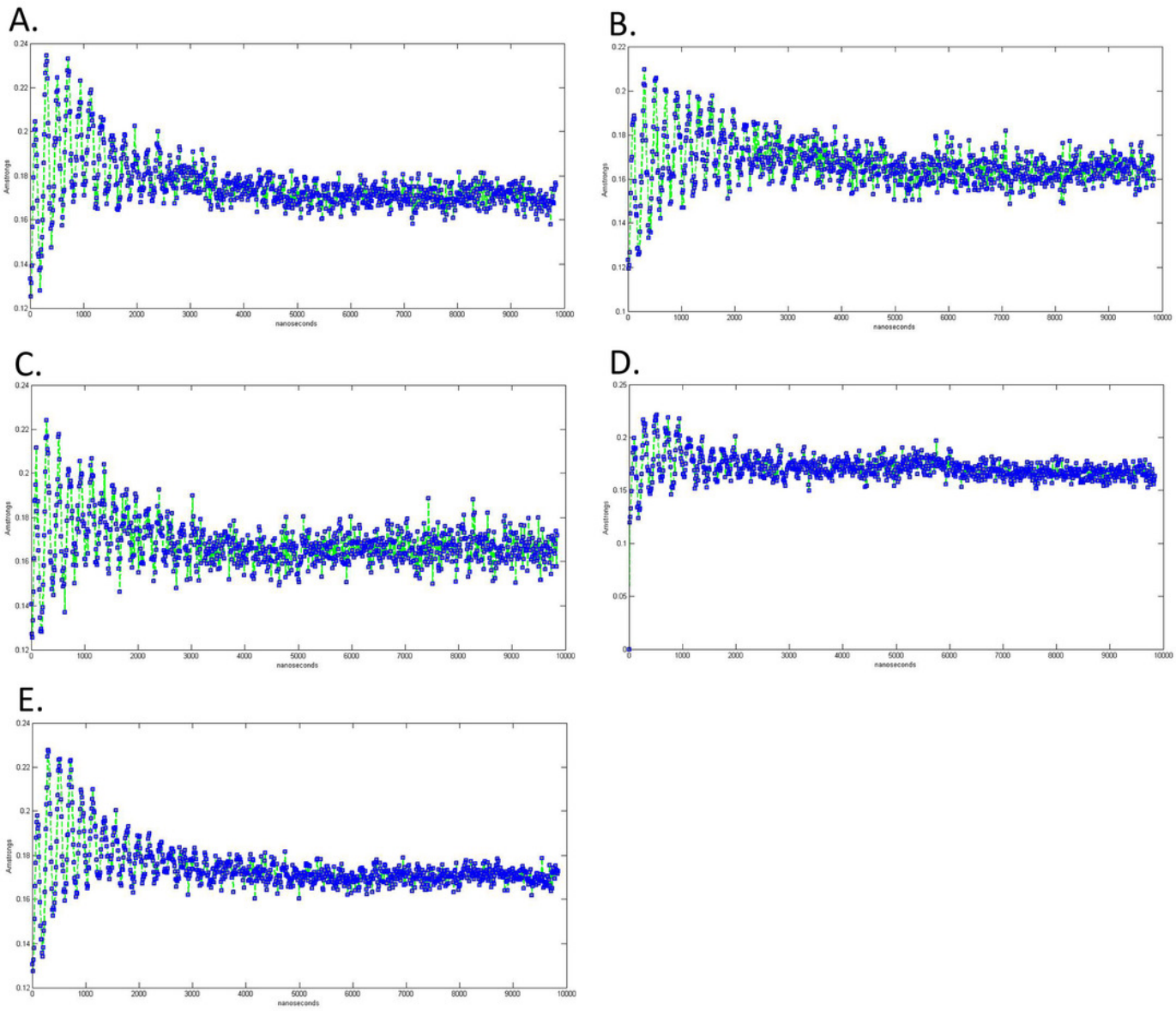

Figure 10

Molecular dynamics simulation charts of the root mean square deviation (RMSD) for the *Trypanosoma brucei brucei* DdRpII RPB1 sub domainsof the model B.

The energy (Kcal/mol) vs time (ns) plot of the 100ns simulation trajectory of the TBB DdRpII RPBI model B. **(A)** Domain A RMSD. **(B)** Domain B RMSD. **(C)** Domain C RMSD. **(D)** Domain D RMSD. **(E)** Domain E RMSD.

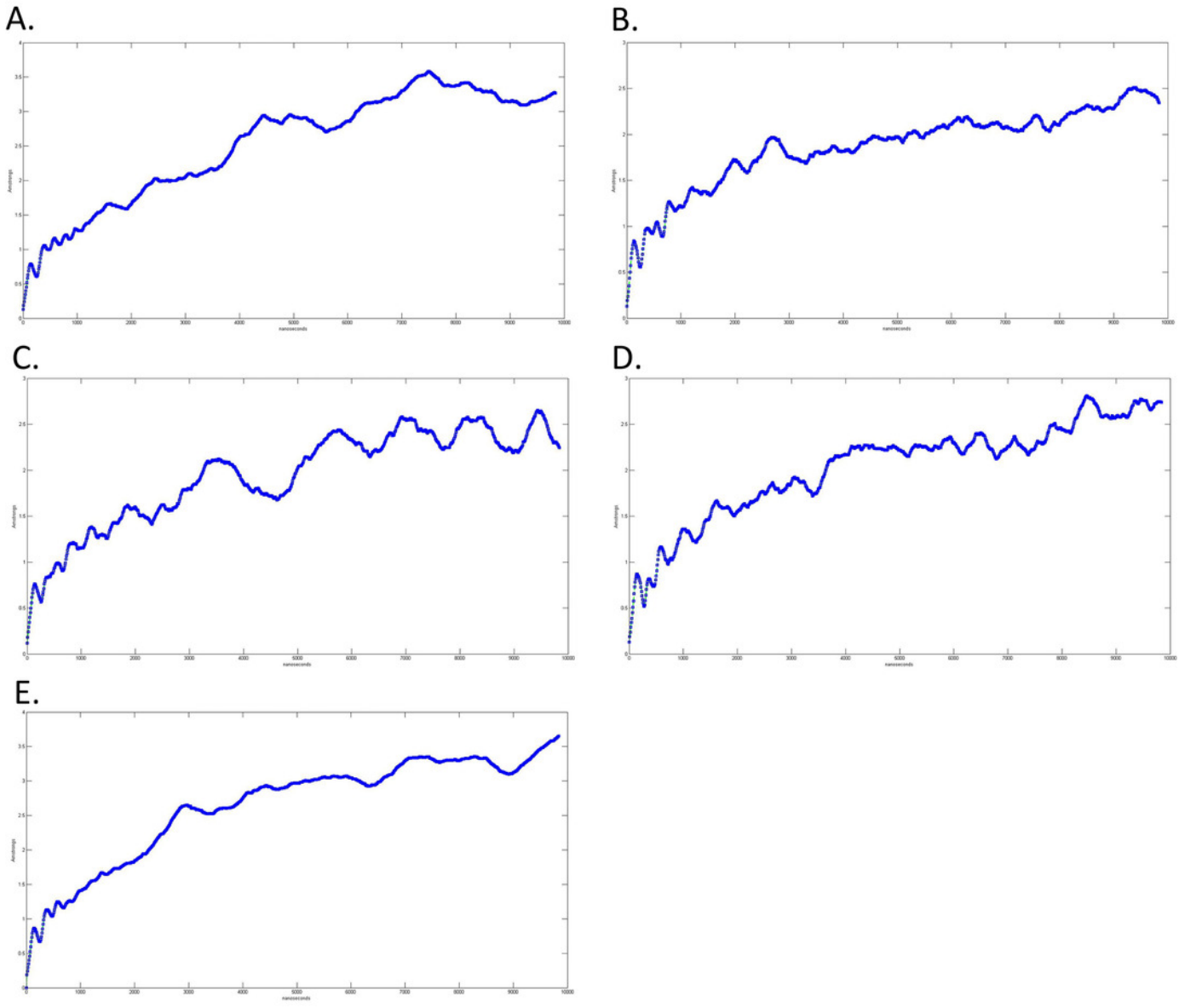

Figure 11

Molecular dynamics simulation charts of the root mean square fluctuation (RMSF) for the *Trypanosoma brucei brucei* DdRpII RPB1 sub domains of the model B.

(A) Domain A RMSF. **(B)** Domain B RMSF. **(C)** Domain C RMSF. **(D)** Domain D RMSF. **(E)** Domain E RMSF.

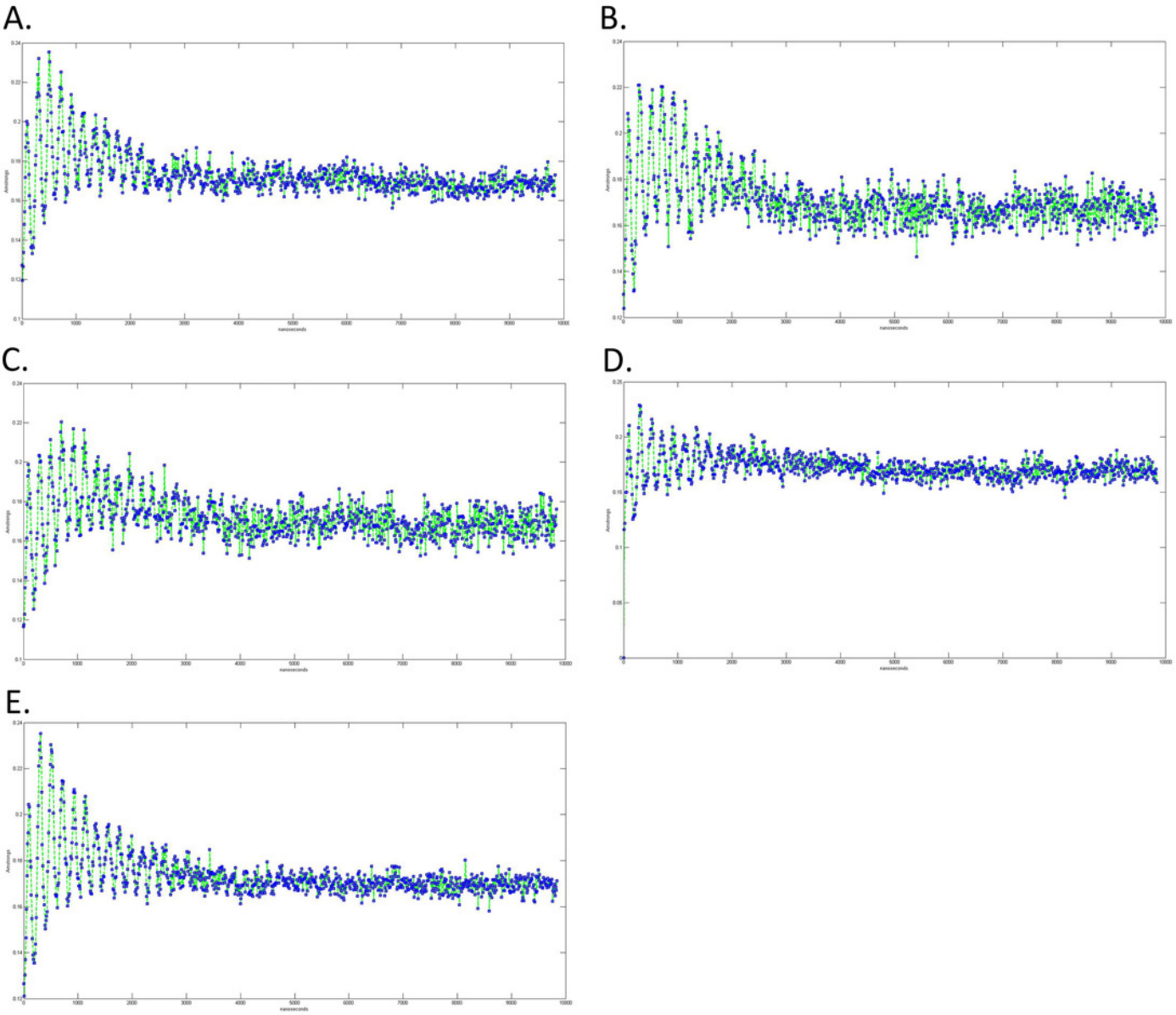

Figure 12

The 3D pharmacophore model for the *Trypanosomabrucei brucei* DdRPII RPB1 model.

In total 5 distinct pharmacophoric features were identified. An aromatic region (colored orange), an electron donating region (colored green), two electron accepting regions (colored red) and a sulphur specific S-S interacting region (colored yellow).