APPLICATION OF STRUCTURAL BIOINFORMATICS IN AFRICAN SWINE FEVER VIRUS ANTIVIRAL DEVELOPMENT

DICKSON BENNET KINYANYI

B.Sc. (JKUAT); M.Sc. (UoN)

A Thesis Submitted in Partial Fulfillment of the Requirements for the Award of the Degree of Doctor of Philosophy in Bioinformatics

in the

School of Biological and Life Sciences

of

The Technical University of Kenya

(July 2022)

DECLARATION

This thesis is my original work and has not been presented in any other institution for a degree award or any other qualification.

Dickson Bennet Kinyanyi.
PHD/SSBX/06465P/2015

Signature _____ Date _____

Dr. Mark Wamalwa
Senior Lecturer of Bioinformatics,
Kenyatta University

Signature _____ Date _____

Dr. George Fredrick Opondo Obiero
Senior Lecturer of Bioinformatics,
The Technical University of Kenya

Signature _____ Date _____

Dr. George Ogello Obiero
Senior Lecturer of Biochemistry,
University of Nairobi

Signature _____ Date _____

Dr. Peris Wanza Amwayi
Senior Lecturer of Biochemistry,
The Technical University of Kenya

Signature _____ Date _____

DEDICATION

I dedicate this thesis to

Myself and YHWH

ACKNOWLEDGMENT

First and foremost, praises and thanks to God (YWHW), the Almighty for His showers and divine guidance throughout my research work. I also thank my parents for being my instructors and offering moral and financial support throughout my study period. I am indebted to Dr. Mark Wamalwa, Dr. George F. O. Obiero, Dr. George Obiero, and Dr. Peris W. Amwayi for being great supervisors who held my hand and walked with me through this journey right from the conceptualization of this work to completion. I am also thankful for their prized thoughts, sound logical advice, and co-authoring journals with me. Their advice has assisted me in aiming for eminence and helped me to improve as a researcher. Lastly, I am thankful to the Technical University of Kenya and its members (especially Professor Yole and Professor Lalah) for their continuous encouragement.

ABSTRACT

African swine fever (ASF) is a fatal hemorrhagic disease of domesticated pigs that is caused by the African swine fever virus (ASFV). This disease poses a threat to food security thus leading to economic losses. Presently, there are no reports of approved available vaccines for ASF. Despite the ASFV sequences having well-conserved promoter motifs, no protein with features able to bind onto the promoter TATA-like elements has been identified, nor an antiviral capable of targeting viral TATA-like elements involved in ASFV transcription. This study aimed at finding a TATA Binding Protein (TBP) and potential minor groove binders (MGBs) that can target conserved promoter motifs in ASFV. The study implemented sequence-based search methods followed by three-dimensional structure modeling. The posterior probability of fold family classification was calculated using TM-fold, and biological function was determined using TM-site, RaptorXBinding Site, Gene Ontology, and TM-align. Subsequently, congocidine congeners, Hoechst 33258, and tris-benzimidazole were selected as potential minor groove binders targeting synthetic DNA constructs containing TATA-like motifs mimicking conserved ASFV promoters. The binding scores and calculated binding energies of docked DNA-ligands complexes were evaluated. The ligand behaviour within the minor grooves was assessed using molecular dynamics (MD) simulation. The results of this study established that the three-dimensional structure of a previously uncharacterized protein pB263R had features similar to a TBP with a TM-score of 0.52, with > 95% posterior probability. Additionally, the selected minor groove binders were able to significantly dock on the AT-rich regions of the synthetic DNA constructs containing TATA-like motifs. Further, calculated binding energies revealed that less cytotoxic congocidine congeners and tris-benzimidazole were an improvement of cytotoxic congocidine. The MD simulation and molecular trajectory visualization revealed that the ligands remained embedded in the minor grooves of synthetic DNA constructs during the MD simulation time course. The findings of this study suggest that these ligands may be used as potential antivirals for ASF infection in abrogating ASFV transcription. Critical for control of several ASFV genotypes.

KEYWORDS: African Swine Fever Virus, antiviral, TATA-binding protein, congocidine congeners, Hoechst 33258, tris-benzimidazole

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ASF	African Swine Fever
ASFV	African Swine Fever Virus
ADME	Absorption, Distribution, Metabolism, Excretion
BA71V	Badajoz 1971 Vero-adapted
BLAST	Basic Local Alignment Search Tool
CASP	Critical Assessment of techniques for protein Structure Prediction
CATH	Class, Architecture, Topology, Homology
CM	Comparative Modeling
C-score(Cs)	Confidence score for I-TASSER
(CSt)	Confidence score for TM-SITE
FG-MD	Fragment Guided Molecular Dynamics simulation
GO	Gene Ontology
HSP	High-scoring Segment Pairs
HMM	Hidden Markov Model
ICM	Internal Coordinate Mechanism
I-TASSER	Iterative Threading ASSEmbly Refinement
LOMETS	Local Meta Threading Server
MCV	microencapsulation vesicle
MGB	Minor Groove Binder
NCBI	National Center for Biotechnology Information
NCLDV	Nucleocytoplasmic large DNA virus
NR	Non-Redundant
ORF	Open Reading Frame
PDB	Protein Data Bank
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool
RefSeq	Reference Sequence
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
RNAP	RNA polymerase
SCOP	Structural Classification of Protein
TBP	TATA binding protein
TM	Template Modeling
UniProtKB	Universal Protein Resource KnowledgeBase

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND INFORMATION

African swine fever (ASF) is an infectious hemorrhagic disease of domestic pigs caused by the African swine fever virus (ASFV) (Dixon *et al*., 2013). The disease is characterized by loss of appetite, hemorrhages from the ears, abdomen, legs, nose, rectum, and fever. Often, death occurs within ten days with a mortality rate of up to 100% (Montgomery, 1921). Pork accounts for more than 35% of the global meat intake (Utnik-Banaś *et al*., 2022). Consequently, ASF poses a risk to the global pork industry and food security (Utnik-Banaś *et al*., 2022). The ASF is prevalent and endemic in sub-Saharan Africa, it is also found in Asia, Europe, Oceania, and America (Sur, 2019). The natural hosts of ASFV include warthogs and bush pigs (both common in the African wild), wild swine (boar), and a tick vector of the genus *Ornithodoros*. Most primary hosts, such as the warthog, act as a natural reservoir of the ASFV without showing signs of disease. The virus is then taken up by the soft tick *Ornithodoros moubata* during a blood meal and passed on when the tick feeds on susceptible animals. Currently, there is no approved vaccine or antiviral that can control ASFV (Arabyan *et al*., 2019; Reis *et al*., 2016). However, various control strategies have been used; such as stringent import policies in ASF-free countries, control of soft tick vectors, rapid diagnosis and isolation of infected animals, and carcasses of infected animals being carefully disposed of (Afonso *et al*., 2004; Zsak *et al*., 1998).

The ASF pathogen, ASFV, is a double-stranded DNA virus with a large envelope. It is the only member of the Asfarviridae family (Dixon *et al*., 2004). The complete ASFV genome comprises

between 150 to 170 protein-coding open reading frames (ORFs), closely spaced on the coding strands(Dixon *et al*., 2013; Rodríguez & Salas, 2013). Despite the availability of a complete Badajoz 1971 Vero-adapted (BA71V) genome (a nonpathogenic, tissue-cultured strain of ASFV), about one-third of the ORFs lack a known function(s) (Yáñez *et al*., 1995).

Transcription of ASFV has been hypothesized to have at least two distinct phases: a pre-replicative (early genes) and a post-replicative (late genes phase). These phases are based on mRNA accumulation kinetics. After viral entry into the host and viral replication, viral early and late genes are expressed (Rodríguez & Salas, 2013). Promoter sequences from the ASFV genome have TATA-like motifs at or close to the transcription start sites of the genes. These include the early gene promoter motif TATA(N) (where N is any nucleotide) and the late gene promoter motif TATATA (Cackett, Matelska, *et al*., 2020). Another variant, TATTT motif, is also found in both early and late genes - as part of the bipartite motifs involved in transcription (García-Escudero & Viñuela, 2000; Oliveira *et al*., 2017, 2018; Rodríguez & Salas, 2013). Studies have also shown that promoter motif copy numbers and distribution within NCLDVs genome differ significantly, with the TATTT motif having higher numbers and distribution (Oliveira *et al*., 2018). Additionally, mutations in TATA-like sequences in BA71V inhibit the transcription of ASFV genes such as K78R ,B646L, A137R and EP402R (García-Escudero & Viñuela, 2000).

It is hypothetically possible that RNA polymerase involved in ASFV transcription gains temporal late gene expression reliant on the association of TATA Binding Protein (TBP) and Transcription Factor II B-like (TFIIB) (Cackett, Sýkora, *et al*., 2020). The TFIIB-like protein has been identified in ASFV (Iyer *et al*., 2006). However, there are no reports of a TBP in ASFV with features that

can support binding to the TATA-like sequences (TATATA and TATTT) (García-Escudero & Viñuela, 2000). The first part of this study scanned the entire ASFV-uncharacterized proteome using computational protein structure prediction approaches, to find a possible transcription factor with the features required for interaction with the TATA-like sequences.

DNA minor groove binders are molecules that preferentially bind to the DNA minor groove side either reversibly or irreversibly (Khan *et al*., 2012). The first group is markedly cytotoxic, while the latter can be modified to make a practical drug that can separate cytotoxicity from antiviral activity (Bialer *et al*., 1980). Currently, reversible MGBs have not been used or considered in abrogating the ASFV transcription. A rational study of ligands having antiviral-like will be tremendously handy in finding potential therapeutic or prophylactic drugs to combat ASF or design better antivirals that can target the TATA-like elements (TATATA and TATTT) involved in ASFV transcription (Cackett, Sýkora, *et al*., 2020; Inmaculada Galindo & Alonso, 2017; García-Escudero & Viñuela, 2000).

Antiviral agents that target viral genomes and transcription in other viruses are known (Beerman *et al*., 1991; Lown *et al*., 1989; Yakimovich *et al*., 2017). For example, antivirals targeting replicative elements of viral multiplication and the spread of classical swine fever (CSF) have been reported (Vrancken *et al*., 2009). The antiviral ligand (5-[(4-bromophenyl) methyl]-2-phenyl-5H-imidazo[4,5-c]pyridine (BPIP)) administered to pigs infected with CSF virus showed a 1000-fold viral reduction in blood and shortened the period of viremia and transmission by 74% and 85% respectively. Additionally, epidemiological modeling studies have shown that antiviral drugs were able to control outbreaks of CSF as effectively as emergency vaccination and preventative culling

(Backer *et al*., 2013; Vrancken *et al*., 2009). Additionally, Inhibiting TBP through promoter motifs by Hoechst 33258 (Yakimovich *et al*., 2017) and distamycin A (Bellorini *et al*., 1995; Broyles *et al*., 2004; Knutson *et al*., 2006) has been shown to effectively stop transcription in the Vaccinia virus a nucleocytoplasmic large DNA virus. Therefore, the development of ASFV antiviral drugs remains an important option in the absence of a vaccine (Cackett, Sýkora, *et al*., 2020; Inmaculada Galindo & Alonso, 2017).

Based on computational and structural studies of DNA, early computational studies have demonstrated the likelihood of targeting AT elements of the minor groove of B-DNA (Yoon *et al*., 1988) using small molecules such as netropsin (Coll *et al*., 1989), Hoechst 33258 (Carrondo *et al*., 1989; Harshman & Dervan, 1985) and distamycin (Kopka *et al*., 1985). However, targeting DNA A/T motifs is not straightforward because of biases in the binding affinities to various motif permutations (Abu-daya *et al*., 1995; Abu-daya & Fox, 1997). For instance, Hoechst 33258 can show a 50-fold change in binding affinity for the sites TATA and AATT (Abu-daya *et al*., 1995). Additionally, antivirals that bind the DNA minor groove such as congocidine and distamycin, exhibit high cytotoxicity, hence their use as practical drug candidates is limited (Bialer *et al*., 1980; Lombardi & Crisanti, 1997). Thus, the choice of a less cytotoxic or non-cytotoxic ligand that can separate cytotoxicity from the antiviral activity is of prime importance in targeting either TATATA or TATTT, which are conserved ASFV promoter motifs. Additionally, targeted drug delivery systems are important in reducing cytotoxicity (Cortesi *et al*., 2010; Hlaka *et al*., 2017; Kelly *et al*., 2011). This study analyzed results from docking, calculated binding energies and molecular dynamics simulations from synthetic DNA constructs having the aforementioned motifs to demonstrate that MGBs can be used in abrogating the ASFV gene transcription.

1.2 STATEMENT OF THE PROBLEM

Infections by ASFV cause high mortality rates in domesticated pigs, affecting the global pork industry and food security. The molecular elements in the ASFV genome have since been characterized including promoter motifs – TATATA and TATTT, crucial in the transcription of viral genes. Despite the knowledge that these motifs can potentially be targeted to stop the transcription of viral genes, no antiviral drug has been shown to target these promoters, and neither has any antiviral drug been approved for ASF treatment and management. Moreover, some DNA binding elements that drive transcription of the ASFV genes in swine hosts are unknown

1.3 JUSTIFICATION

Pork is one of the most consumed animal meat proteins in the world. The African swine fever virus (ASFV) causes a severe disease ASF in domestic pigs, resulting in the death of nearly all infected pigs, resulting in massive economic losses. There is presently no licensed antiviral therapy or vaccine for ASF treatment, and the mechanism of ASFV transcription remains poorly understood. Through ASFV protein structure prediction, the study may find a protein with features of a TBP in the viral proteome and improve our understanding of viral mechanisms. Since there are no drugs or vaccines approved for ASFV treatment and no report has identified any drug that can target the promoter motifs in the ASFV genome, this study provides the potential of using these motifs as drug targets. The outcomes of this study are crucial in providing a strategy for targeting multiple ASFV proteins as well as targeting heterologous ASFV strains which have proved to be a challenge using Live Attenuated Vaccines (LAVs).

1.4 MAIN OBJECTIVE

To find a TATA-like binding transcription factor from the ASFV-uncharacterized proteome and identify antiviral/s that can potentially target TATA-like motifs using *in silico* methods.

1.4.1 Specific Objectives

1. To identify the ASFV uncharacterized protein that has features to bind to the promoter region of TATA-like elements in the ASFV genome.
2. To validate *in silico,* whether or not antivirals can target the TATATA motifs similar to those seen in the ASFV promoters.
3. To validate *in silico,* whether or not antivirals can target the TATTT motifs similar to those seen in the ASFV promoters.

1.5 RESEARCH QUESTIONS

1. What characteristic features are required for a transcription factor to bind to ASFV TATA-like motifs?
2. How effective are congocidine congeners and tris-benzimidazole in targeting the TATATA motif?
3. How effective are congocidine congeners and Hoechst 33258 in targeting the TATTT motif?

CHAPTER TWO

LITERATURE REVIEW

2.1 AFRICAN SWINE FEVER (ASF)

African swine fever (ASF) is a severe infective disease of wild and domestic pigs caused by the African Swine Fever Virus (ASFV), with up to 100 % mortality (Schulz *et al*., 2017). In domestic pigs, the disease manifests in four different forms. The most common forms include peracute and acute although subacute and chronic forms may also occur (Sánchez-Vizcaíno *et al*., 2015; Schulz *et al*., 2017). In peracute and acute forms, the virulence is high, clinical symptoms include fever, loss of appetite, and lethargy with almost 100% mortality. Other symptoms may include diarrhea, vomiting, coughing, breathing difficulty, and occasional abortion. In subacute forms, the animals show mild clinical symptoms, with recurrent fever for one month with a high chance of recovery. The mortality rate ranges from 30-70% and recovered animals shed off the virus six weeks after infection. The chronic form of the disease manifest with few clinical signs such as occasional spells of fever, stunted growth, and thinness. The mortality rate is less than 30%. Other symptoms include the presence of necrotic skin patches and vulnerability to infection such as pneumonia and lameness (Schulz *et al*., 2017).

ASF is of primary importance because pork is one of the primary sources of animal proteins. Pork accounts for more than 35% of the global meat intake (Utnik-Banaś *et al*., 2022). Hence, this disease poses a serious problem to food security worldwide. The disease burden is felt mostly in pork-consuming continents for example in Asia (China), where the disease outbreak led to at least 50% of the national pig herd being lost. China produces at least 700 million pigs annually accounting for almost 38.6 % of the world pork supply, in an industry worth over $128 billion

(Utnik-Banaś *et al*., 2022). An outbreak in Vietnam imposed significant consequences on the pig industry, over six million pigs were put to death, accounting for more than 20% of the country's pig population. (Nguyen-Thi *et al*., 2021). Additionally, 20 percent of the world's total production of pork is produced in Europe (Zu Ermgassen *et al*., 2016). While North America makes up roughly 11.4% of the total (Utnik-Bana *et al*., 2022). According to estimates, the first year of an ASF outbreak might cost $15 billion US Dollars (Daniel Rock, 2021). Calculating the full economic impact of ASF in Africa is challenging. Despite this, Sub-Saharan Africa continues to be the primary reservoir for the disease, with at least all 24 of the currently existing genotypes circulating (Njau *et al*., 2021). A large-scale free-range pig breeding program is becoming increasingly crucial for village agricultural subsistence in Sub-Saharan African nations (Kagira *et al*., 2010).

2.2 AFRICAN SWINE FEVER VIRUS (ASFV)

ASFV is a double-stranded DNA (dsDNA) virus. The virus to the family Asfarviridae and genus *Asfivirus* (Sánchez-Vizcaíno *et al*., 2019). ASFV is the only insect-borne DNA virus. The virus has an icosahedral shape and is around 200 nm in diameter (Dixon *et al*., 2013). The virus is made up of an envelope, a capsid, an inner capsule membrane, a core shell, and an inner core (Wang *et al*., 2021). Depending on the virus strain, the linear viral DNA varies between 170-190 kb (Dixon *et al*., 2013). Studies of replication strategy and the genome revealed similarities with the Poxviridae, although the viruses differ structurally (Dixon *et al*., 2013; Rodríguez & Salas, 2013). Comparative genomic studies have shown other viruses that are similar to ASFV, such as Faustovirus (Reteno *et al*., 2015), kaumoebavirus (Bajrai *et al*., 2016), and pacmanvirus (Andreani *et al*., 2017).

2.3 ASFV TRANSMISSION CYCLES

The hosts of ASFV include bush pigs and warthogs. Ticks in the genus *Ornithodoros* act as biological reservoirs and vectors of ASFV (Sánchez-Vizcaíno *et al*., 2015), having the ability to spread the ASFV after years of absence from the viremic hosts (Boinas *et al*., 2011). ASFV transmission cycles involve a sylvatic cycle that comprises warthogs and soft ticks mostly in African regions (Jori & Bastos, 2009; Sánchez-Vizcaíno *et al*., 2012), and a domestic transmission cycle that circulates amongst infected domestic pigs or among wild boar without tick involvement (Jori & Bastos, 2009; Sánchez-Vizcaíno *et al*., 2015). The dissemination of ASFV can also occur due to unregulated movements of pigs and their processed products, swill feeding, and lack of biosecurity steps (Costard *et al*., 2009).

2.4 ASFV IN KENYA AND ITS NEIGHBOURS

The ASF is a transboundary animal disease (Beltran-Alcrudo *et al*., 2019). Kenya shares a common border with Uganda, Tanzania, Somalia, Southern Sudan, and Ethiopia. Outbreaks of African swine fever in domestic pigs are required to be reported to the World Organization for Animal Health (OIE). To date, Africa remains an endemic zone. All the 24 genotypes (I–XXIV) that have been classified based on the nucleotide sequence of the B646L gene coding the main capsid protein p72 exist in Africa (Njau *et al*., 2021) Nonetheless, disease outbreak reporting in Africa is described as random and inadequate, with the major causes being weak communication networks, fear of culling, and quarantine, which may result in small-scale farmers incurring substantial losses without compensation (Onzere *et al*., 2018).

However, the total number of ASF outbreaks reported to OIE by Kenya from 2005 to 2019 was 18, the Genotypes I used in this study, IX and X have been found in Kenya (Bishop *et al*., 2015; Hakizimana *et al*., 2021; Njau *et al*., 2021). In Uganda, the total outbreaks reported were 31, and predominant genotypes were IX and X, while in Tanzania, 89 outbreaks have been reported, and genotypes II, X, XVI, and XV were dominant over the same period of 2005-2019 (Hakizimana *et al*., 2021). Recently a new variant genotype XXIII has been found in Ethiopia that shares a common ancestry with genotypes IX and X (Achenbach *et al*., 2017).

2.5 THE BA71V GENOME AS A MODEL FOR GENOMIC AND PROTEOMIC STUDIES

The scope of this study focused primarily on the analysis of the BA71V genome (genotype I). The BA71V is a non-virulent ASFV strain that originated from a virulent BA71 strain that was isolated from an infected pig in 1971 in Badajoz, Spain. The two isolates are almost identical but differ in virulence and slight differences in the genome. For instance, BA71V encodes 152 open reading frames (ORFs) while BA71 isolate encodes 161 major ORFs out of which 149 are common to both genomes (Chapman *et al.,* 2008; Rodríguez *et al.,* 2015). Numerous research groups have preserved the BA71V genome as a template model. This is because in the last 40 years BA71V has been well-studied, with large data output regarding viral molecular biology, gene expression, promoter studies, transcription, replication, repair of DNA and morphogenesis (Cackett, Matelska, *et al*., 2020; García-Escudero & Viñuela, 2000; Oliveira *et al*., 2017, 2018; Rodríguez *et al*., 1996; Rodríguez & Salas, 2013).

2.6 ASFV TARGETS PORCINE MACROPHAGES

ASFV has constrained cellular tropism, mainly targeting the porcine macrophage (Basta *et al*., 2010; Dixon *et al*., 2019; Sánchez-Cordón *et al*., 2017). Macrophages are essential cells that activate and coordinate immune responses to infection. The pathology of ASFV is attributed to its ability to infect and predominantly multiply in macrophages. The virus also infects other cell types such as reticular cells, fibroblasts, and endothelial cells. Nonetheless, infection in these cells occurs in the late stages when the typical symptoms of ASFV have taken place (Rodriguez *et al*., 1996). Therefore, this implies that these cells are not crucial in ASFV disease pathology (León *et al*., 2012).

ASFV destruction of macrophages is known to impair hemostasis by cytokines, complement factors, and arachidonic metabolite release (Pikalo *et al*., 2019). Apoptosis of lymphocytes in infected pigs causes lymphocytopenia, while prominent hemorrhagic lesions are caused by intravascular coagulation and the release of cytokines by infected macrophages (Pikalo *et al*., 2019.

ASFV displays a largely cytoplasmic replication cycle. Viral entry into the host immune cells is by both clathrin-mediated endocytosis and constitutive macropinocytosis. However, the main mechanism of entry is receptor-mediated endocytosis (Hernáez *et al*., 2016). The cellular receptor for binding and entry of ASFV remains unknown (Hernáez *et al*., 2016). After internalization into the cell, the viral envelope fuses with endosomes and in the presence of favorable conditions such as a low pH, triggers the viral core release into the host cell cytosol (Hernáez *et al*., 2016; Sánchez, Pérez-Núñez, *et al*., 2017).

2.7 TRANSCRIPTION OF THE ASFV GENES

ASFV initiates transcription upon entry into the host cytoplasm. The host RNA polymerases are not used at this stage (Rodríguez & Salas, 2013). Transcription in ASFV is similar to that of vaccinia virus system, a NCLDV in the family Poxviridae (Rodríguez & Salas, 2013). Based on a vaccinia virus model, ASFV gene expression has been postulated to include immediate-early, early, intermediate and late genes (Almazan *et al*., 1993; Rodríguez & Salas, 2013). Experimental evidence for the postulated four separate ASFV gene expression phases has yet to be proved (Cackett, Matelska, *et al*., 2020). However, two separate subsets of transcription initiation factors strongly support two stages, early and late which are representative of pre-replicative and post-replicative phases of the ASFV life cycle (Cackett, Matelska, *et al*., 2020). At beginning of an infection, a division of early viral proteins is essential for host cell manipulation and DNA replication. Towards the end of infection, transcription of late genes occurs, leading to the synthesis of structural proteins, enzymes, and non-structural proteins present in the mature particle that are necessary for the assembly of the virus (Dixon *et al*., 2013; Rodríguez & Salas, 2013). Following the transcription of these late structural proteins, conversion of the asters of vimentin into cages occurs that act as virus assembly factories. Late virus gene expression is mandatory for cage formation or vimentin rearrangement (Stefanovic *et al*., 2005).

ASFV is independent in terms of viral mRNA modification and transcription within the eukaryotic host cytoplasm. The virus codes an mRNA capping enzyme, a poly-A polymerase, and RNA polymerase (RNAP) (Rodríguez & Salas, 2013). Notably, extracts from mature viral particles are capable of transcription (Alejo *et al.*, 2018; Keßler *et al.*, 2018).

The ASFV basal transcription machinery is similar to the eukaryotic RNAPII system, made up of 8-subunits RBP1, RBP2, RBP3, RBP5, RPB6, RBP7, RBP9, and RBP10 in ASFV-RNAP (Rodríguez *et al.*, 2015; Rodríguez & Salas, 2013), the TFIIB and TFIIS (Rodríguez & Salas, 2013; Yáñez *et al*., 1995). ASFV also codes a histone-like DNA binding protein pA104R (Frouco *et al*., 2017) and ASFV topoisomerase II (pP1192R) (Coelho *et al*., 2016) which cooperate to generate DNA-binding and supercoiling activity (Frouco *et al*., 2017).

Of particular interest are promoter motifs through which the ASFV-RNAP may access promoters in pre-replicative and post-replicative stages of gene expression. ASFV has short DNA conserved regions located upstream of the transcription start site of the ASFV ORFs. These conserved regions act as DNA binding areas for transcriptional factors. The short DNA conserved regions constitute the promoter DNA elements (also promoter motifs). In the ASFV genome, these promoter elements include The TATA-like elements, TATATA, TATTT, TATA(N) where (N) represents any nucleotide (Cackett, Matelska, *et al*., 2020; García-Escudero & Viñuela, 2000; Oliveira *et al*., 2018; Rodríguez & Salas, 2013). Well-known promoter motifs are commonly associated with tissue-specific gene transcription. Promoter motifs are highly conserved DNA elements that aid transcription factors with recognition sites thus, increasing transcription efficiency. The transcription initiation sites of ASFV late genes indicate that transcription often starts around the A or T residues (Cackett, Matelska, *et al*., 2020; García-Escudero & Viñuela, 2000). From the observed maximized identities, TATA-like motifs are located at or near the starting point of the late genes (Figure 2.1).

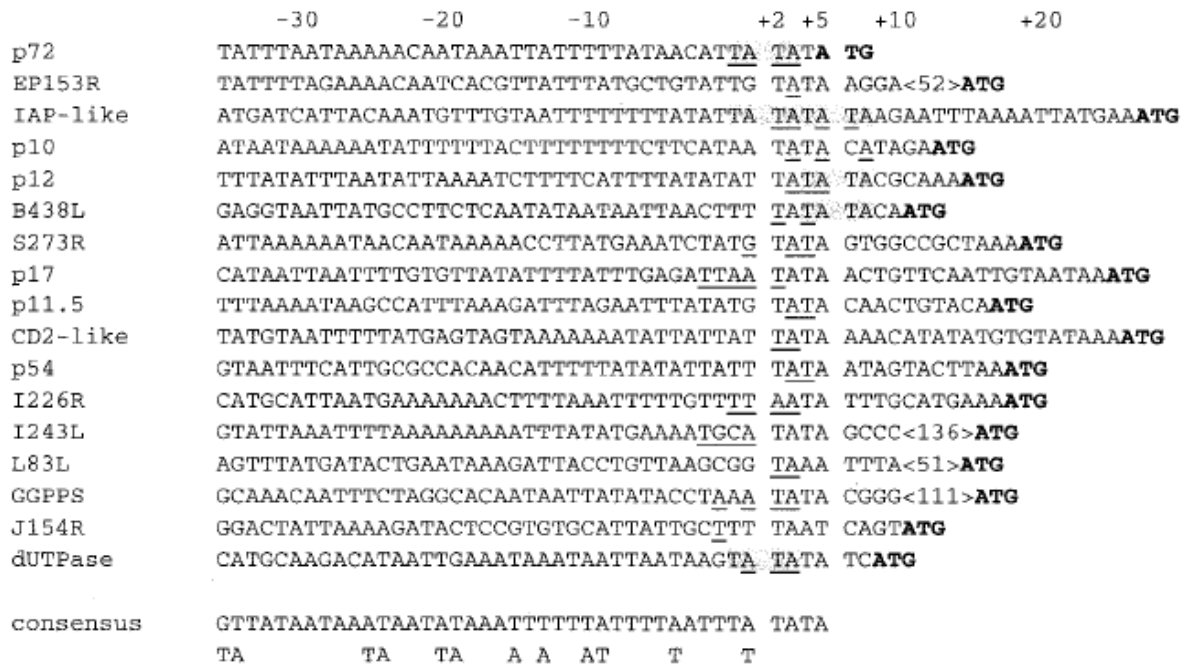


Figure 2.1 Alignment of 5'flanking sequences of the maximized identities of late ASFV genes

The translation initiation codons are indicated in bold: p72 (Liu *et al*., 2019), EP153R (Immaculada Galindo *et al*., 2000), IAP-like (Chacón *et al*., 1995), p12 (Almazán *et al*., 1993), B438L (Inmaculada Galindo *et al*., 2000), S273R (Alejo *et al*., 2003), L83L (Borca *et al*., 2018), p17 (Alcami *et al*., 1993), p10 and p11.5 (Alejo *et al*., 2018), CD2-like (Rodríguez *et al.*, 1993), p54 (Rodríguez *et al.*, 2004), I226R and I243L (Rodríguez *et al.*, 1996), geranyl pyrophosphate synthase (GPPS) (Alejo *et al*., 1997), J154R (Almazán *et al*., 1995), and dUTPase (Oliveros *et al*., 1999). Courtesy (García-Escudero & Viñuela, 2000).

Various studies on the promoter motifs of the ASFV BA71V p72 gene (coding the major capsid protein) have shown promoter activity of the p72 gene occurs at −1 to +5 (TATATA) and −15 to −11 (TATTT) bases upstream of the transcription start site. Replacement of the TATTT motif by a GCGCG motif obliterated the activity of the promoter. Additionally, single-nucleotide replacements within the TATATA motif of the B646L promoter have shown that position −1 is intolerant to any nonconformity from the wild-type motif. The TATA-like sequence positioned between −2 and +2 is essential for B646L promoter activity. More significantly, the total

replacement of a corresponding 'TATA' motif on the late genes EP402R, K78R, and A137R by the 'GCGC' motif stagnated transcription, signifying the A/T rich motifs are essential for late promoter function (García-Escudero & Viñuela, 2000).

Apart from ASFV, the abundance of TATTT motifs has been reported in faustovirus and kaumoebavirus genomes (Oliveira *et al*., 2018). The TATTT motifs can exist as a monomer or as repeats (Oliveira *et al*., 2018). Copy numbers of TATTT motifs are significantly higher than TATATA (Oliveira *et al*., 2018). Besides, the TATTT motif can pair up as TATTT and TATATA as seen in intergenic regions of faustovirus, kaumoebavirus, and asfarvirus (Oliveira *et al*., 2018). The split structure of TATTT/TATATA observed in late promoters of ASFV is analogous to intermediate and late promoters in poxviruses that contain an initiator and a core region (Hänggi *et al*., 1986; Rodríguez & Salas, 2013).

Despite the confirmed presence of ASFV promoters (Cackett, Matelska, *et al*., 2020; García-Escudero & Viñuela, 2000; Rodríguez & Salas, 2013), a transcription factor with features capable of binding to TATA-like motifs (TATATA and TATTT) has remained elusive (García-Escudero & Viñuela, 2000). And the option of targeting TATA-like sequences in ASFV responsible for the temporal control of transcription in the late replicative stages of ASFV using antivirals has not been studied (Cackett, Sýkora, *et al*., 2020). Therefore, the study searched for a transcription factor that can bind to in the first objective and small molecules that can target these TATA-like motifs in the second and third objectives.

2.8 THE SEARCH FOR A VACCINE AGAINST ASFV

Pigs that survive ASFV infection tend to develop a strong and protective immunity, demonstrating the possibility of a promising working active vaccine against ASFV (Pérez-núñez *et al*., 2019). Nonetheless, efforts to develop vaccines via recombinant proteins, DNAs, or recent combinations of DNAs plus viral proteins and inactivated viruses are unproductive (Arias *et al*., 2017; Elena *et al*., 2019; Revilla *et al*., 2018; Daniel Rock, 2017). Additionally, other methods that use adenovirus as a viral vector and subunit vaccines against ASFV antigens have been unsuccessful (Arias *et al*., 2017; Elena *et al*., 2019). Due to the failure of various vaccine options to elicit a strong immune response, the use of the gene-deleted live attenuated vaccine (LAV) has emerged as an alternative control of ASFV (Gladue & Borca, 2022; Daniel Rock, 2021). Live attenuated ASFV vaccine candidates are known to produce several side effects such as joint swelling and persistent fever (King *et al*., 2011). There are also safety concerns with LAVs based on observations that live attenuated vaccines are known to revert to virulence (Gladue & Borca, 2022). Experiments have also shown that whether viruses are genetically modified or occurring naturally, it is difficult to grow them in cell culture, which is crucial for making vaccines in large quantities, as macrophages do not last very well in culture. (Sánchez, Riera, *et al*., 2017). Moreover, LAVs protect against a homologous ASFV challenge in most cases. Heterologous protection remains a challenge in all 24 genotypes existing in Africa (Turlewicz-Podbielska *et al*., 2021). These challenges have impaired research and the commercial production of LAV candidates (Revilla *et al*., 2018), which has led to the search for alternative small molecules to stop the rapid spread of the disease (Arabyan *et al*., 2019; Cackett, Sýkora, *et al*., 2020; Inmaculada Galindo & Alonso, 2017).

2.9 PROSPECTS OF ASFV ANTIVIRAL IN VACCINE ABSENCE

Unlike vaccines, which are biological preparations providing active acquired immunity to a particular infectious disease, antivirals, are small molecules that can inhibit viral multiplication. The antiviral drug approach might be feasible in disrupting the ASFV transcription because the ASFV genome is replete with conserved TATA-like motifs that are druggable targets essential for gene transcription (Cackett, Matelska, *et al*., 2020; García-Escudero & Viñuela, 2000; Oliveira *et al*., 2017, 2018; Rodríguez & Salas, 2013).

To date, no selective anti-viral drug has been authorized for the prevention or treatment of ASF. Therefore, there is a need for more detailed studies on antivirals that can mitigate ASFV (Arabyan *et al*., 2019). Previous studies have shown that certain compounds can inhibit ASFV infection *in vitro*. These compounds include microalgae (Fabregas *et al*., 1999), lauryl gallate (Hurtado *et al*., 2008), small peptide inhibitors (Hernáez *et al*., 2010), resveratrol and oxyresveratrol (Galindo *et al*., 2011), fluoroquinolones (Coelho *et al*., 2016; Freitas *et al*., 2016; Mottola *et al*., 2013), genkwanin (Hakobyan *et al*., 2019), apigenin (Hakobyan *et al*., 2016), polyvalent two-dimensional entry inhibitors (Ziem *et al*., 2017) and sodium phenylbutyrate (Frouco *et al*., 2017). These antivirals show promise, nonetheless, most of them cannot target multiple genotypes and proteins once, however, a single point mutation in a protein is enough to change binding site properties. The option of targeting conserved promoter elements in ASFV and transcription factors has not been exploited (Cackett, Sýkora, *et al*., 2020). Targeting the viral transcription may be an effective strategy that leads to sustained inhibition of transcription of the virus by making it unable to multiply, hence, reducing the emergence of drug resistance mutations.

2.10 STRUCTURAL BIOINFORMATICS

Structural bioinformatics is associated with the exploration of the three-dimensional structure of biological macromolecules like proteins, RNA, DNA, and their complexes. Structural bioinformatics is involved in macromolecular comparisons of folds, substructures of folds, binding sites, principles of docking, molecular dynamics (MD) simulations, and structure/function relationships. Structural bioinformatics works from experimentally solved structures and computational models available in databases.

2.11 DATABASES FOR PROTEIN AND DNA ANALYSIS USED IN THIS STUDY

Over the years, the field of biology has become a science that is rich in data, consequently, to store the data collected from laboratory data, a broad range of databases have been created from experiments and simulations. This information has been organized into various repositories in the form of databases for easier access and to minimize redundancy the databases used in this study are summarized below.

2.11.1 Protein Data Bank.

Protein Data Bank (PDB) is a database for 3D structural information of proteins and nucleic acids (Berman *et al*., 2000). As of June 2020, the PDB contained over 165,000 protein structures.

2.11.2 Nucleic acid Database

Information on experimentally-determined nucleic acids and complex assemblies is included in the Nucleic Acid Database (NDB) (Narayanan *et al*., 2014). The goal of the NDB is to archive and share structural information about nucleic acid details. As of June 2020, the NDB contained more than 10800 nucleic acid complexes.

2.11.3 Reference Sequence Database

The reference sequence database (NCBI RefSeq) provides an annotated and curated collection of sequences of nucleotides and protein sequences used as standard reference sequences for organisms (Pruitt *et al*., 2007). The NCBI RefSeq provides a crucial platform for the integration of sequence, genetic and functional knowledge, and is used internationally as a genome annotation standard.

2.11.4 UniProtKB

Universal protein resource (UniProtKB) is an organized protein sequence repository that provides a high degree of annotation such as protein structure, protein function, protein domains, protein structures, and post-translational modifications. (Bateman *et al*., 2017). The study retrieved all the uncharacterized proteins from UniProtKB.

2.11.5 BioLiP

BioLip is a semi-manually organized databank for biologically important interactions, which contains more than 204,223 entries, including information on protein-protein, nucleic acid-protein, lipid-protein, and small molecule–protein interactions. BioLip is used for binding pocket and ligand binding site prediction (Jianyi *et al*., 2013a).

2.11.6 Class, Architecture, Topology, Homology (CATH)

The CATH database contains proteins split into structural domains and allocated a hierarchical classification of homologous superfamilies. The classification uses a four-tier structural annotation of Class, Architecture, Topology, and Homology. CATH uses a mixture of manual and automated

techniques which include, statistical and empirical evidence, literature review, expert analysis, and computational algorithms (Knudsen & Wiuf, 2010). The study used CATH for fold family classification.

2.11.7 Structural Classification of Proteins (SCOP)

Structural Classification of Proteins (SCOP) uses a classification centered on full proteins instead of domains. The hierarchy offered by SCOP has four tiers: class, fold, superfamily and family. SCOP has been updated to SCOP2, to focus on knowledge-based expert analysis and fold family classification of proteins that are structurally characterized and deposited in the Protein Data Bank. The proteins are organized according to their structural and evolutionary relationships in this database (Chandonia *et al*., 2019). SCOP was used for fold family classification.

2.11.8 Gene Ontology (GO)

The Gene Ontology database uses structured, regulated vocabulary and classifications, to provide annotations of genes, gene products, and sequences (Gene Ontology Consortium, 2004). The Gene Ontology database uses molecular functions, biological processes, and cellular components in assigning function/s.

2.11.9 PubChem

PubChem is an organized data repository found in the National Center of Biotechnology Information (NCBI). PubChem contains mostly small molecules, though, larger molecules such as carbohydrates, peptides nucleotides, carbohydrates, lipids, and chemical-modified macromolecules can also be found (Kim *et al*., 2016).

2.12 EVALUATING SIMILARITY BETWEEN PROTEINS

The similarity between two or more structures or sequences can be obtained using the dynamic programming of Needleman Wunsch (Needleman & Wunsch, 1970) or Smith-Waterman (Smith & Waterman, 1981). By comparing two proteins, similarities can be detected which allows inference of the function and structure of unknown proteins using already characterized proteins. Dynamic programming has been applied in this study for pairwise sequence comparisons, profile hidden Markov models, and protein structural alignment. The general case of the dynamic programming algorithm requires three components (i) Input sequences, (ii) A substitution matrix for scoring, and (iii) A matrix from which a traceback path can be calculated to find optimal alignments. The dynamic programming algorithm has an (i) Initialization step, (ii) matrix-filling step, and (iii) trace-back step. In the first step, for any two structures or sequences, S1 and S2, of lengths x and y, a matrix (M) of size (x+1, y+1) is constructed. The matrix M is initialized to M (0, 0) = 0. M (0, j) is then scored in the horizontal direction arithmetically calculated as $F(i,j-1)+d$, where d is a gap penalty resulting from the alignment of S1. Similarly, M (i, 0) is initialized in the vertical direction arithmetically calculated as $F(i-1,j)+d$, resulting from the alignment of S2. For every cell in the matrix M with $i, j > 0$, the M (i, j) represents a diagonal match score $F(i-1,j-1)+s(x_i,y_j)$ signifying how favorably the residue or the nucleotide matched in i and j are scored, otherwise, an insertion or deletion arises. The matrix M is then filled from the top-right to the bottom-left, using the formula:

$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i,j-1) + d \\ F(i-1,j) + d \end{cases}$$ Global alignment

Here: (1) F defines the values in the rows and columns *i,j* respectively, and the score *s* (x_i, y_j) is a similarity score of the substitution matrix, for structural comparisons or emission probabilities for profile hidden Markov models. Throughout the matrix filling process, for each cell coordinate, a maximal value is kept for trace-back. For a global alignment also known as the Needleman Wunsch algorithm, these choices are traced back in the last step, starting from the bottom-right cell of the calculated matrix M to the top-left in an end-to-end alignment.

For a local alignment, also known as the Smith-Waterman algorithm, a modification is introduced such that a negative scoring cell is set to 0 during the matrix filling process, hence aiding in the identification of cells that positively score rendering the alignments local. In the local alignment, there is no end-to-end alignment, the trace-back begins from a cell with the maximum value until a cell with a value less than 0 is encountered. The local alignment is defined below using the formula.

$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i, j-1) + d \\ F(i-1, j) + d \\ 0 \end{cases} \quad \text{Local alignment}$$

2.13 MONTE CARLO METHODS IN CONFORMATIONAL SAMPLING

Protein structure prediction from its amino acid sequence requires a powerful optimization method able to find the minimum energy structure in a possible amount of computational time. In thermodynamic terms, the minimum energy structure corresponds to the native state (Anfinsen, 1973). This global minimum low energy state or ligand conformation depends on an energy function that can be found using the Metropolis-Hastings (Hastings, 1970) criterion where the probability (P) is exponentially enhanced and compared with a random walk. Each trial movement can be accepted or rejected based on the Metropolis-Hastings criterion, that is, if the ratio is greater than or equal to 1, the next state is accepted. If not, then a random number is generated (0<R<1), if the exponentially enhanced (P) is less than the random number (0-1) then a move is accepted to the next state, otherwise, if the random number is greater than P it is returned to the previous state (Figure 2.2). The Metropolis-Hastings algorithm is run for a considerably long period before the sampling starts.

Probability (P) is evaluated by: $P = e^{-\Delta G/kT}$

Where; delta G is the change in energy, T is the temperature and k is the Boltzmann constant.

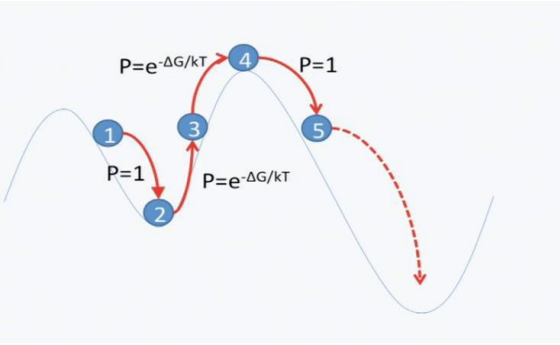


Figure 2.2 Depiction of the sampling process in the energy landscape

The blue circles represent the possible states (1, 2, 3, 4, and 5) while the P represents the probability of acceptance.

The energy landscape of the protein is characterized by several local minima separated by energy barriers. Therefore, the Metropolis-Hastings algorithm can get trapped in these local minima energy landscapes. As a result, only minor portions of the whole space are explored in practice, making it slow for protein structure sampling. To bypass this local minimum trapping, replica sampling is employed, where the simulations of several replicas are applied at varying temperatures (Kihara & Skolnick, 2003). By swapping the states at different temperatures, the higher temperature structures can assist the lower-temperature structures to cross the energy barriers between various basins and attain ergodicity. Each swap and trial movement is rejected or accepted according to a probability represented below.

$$p = \min\left(1, \frac{\exp\left(-\frac{E_j}{kT_i} - \frac{E_i}{kT_j}\right)}{\exp\left(-\frac{E_i}{kT_i} - \frac{E_j}{kT_j}\right)}\right) = \min\left(1, e^{(E_i - E_j)\left(\frac{1}{kT_i} - \frac{1}{kT_j}\right)}\right)$$

Where; $1/kT$ is equal to the general inverse temperature in i and j respectively.

Further modifications involve smoothing the local high-energy barriers found on the irregular energy landscape, thus simulations at different temperatures can be applied on a reasonably smooth landscape.

2.14 SEQUENCE-BASED ALIGNMENTS

2.14.1 Basic Local Alignment Search Tool (BLAST) Heuristic Database Search

Although searching a large database using dynamic programming methods, such as the Needleman–Wunsch algorithm gives accurate and reliable results, it is too slow and impracticable when computational resources are limited. Word methods are thus applied, an example being the Basic Local Alignment Search Tool (BLAST). The BLAST algorithm is a heuristic program (Altschul *et al*., 1990), and uses computational strategies to find an empirical or near-optimal solution using rules of thumb. BLAST identifies local similarity regions of sequences. The algorithm compares protein and nucleotide sequences to sequences in the database and computes the statistical significance of the matches. BLAST infers evolutionary and functional relationships between sequences. The BLAST algorithm involves three key steps. During the first step, the query sequence is divided into 'k-mers', where the default value of a 'k-mer' is 3 for a protein sequence. All possible combinations of k-mers i.e (20^3) are then scored using a substitution matrix (BLOSUM 62) against the query terms for proteins. During the second step, words higher than the threshold cut-off value of T are marked as hits. The chosen hits are used to search for the query sequence within the database. Using un-gapped alignment, database matches are extended in both directions until the alignment score falls under a cut-off threshold score S. In the final step, high-scoring HSPs are further extended by gapped alignment and the highest-scoring hits are returned with the final alignment in terms of e-values or p-values.

2.14.2 PSI-BLAST Heuristic Database Search

The selection of the Block Substitution Matrix (BLOSUM) matrix used in heuristic algorithms like BLAST for alignment scores decides how accurately the system can differentiate random hits

from true hits in the case of a database search. Thus, by selecting a matrix that best describes the features of the query protein family, the sensitivity of these heuristic algorithms is increased. This idea forms the basis of the Position-Specific Iterative BLAST (PSI-BLAST) algorithm, which uses the query profile to rank database sequences (Altschul *et al*., 1997). A Position-Specific Scoring Matrix (PSSM) or profile is a $20 \times L$ dimension matrix, where L is the query sequence length, and the 20 columns comprise substitution scores for the standard 20 amino acids (Henikoff & Henikoff, 1992). A repetitive search technique is implemented in PSI-BLAST, where the initial search is the same as running BLAST. Statistically, essential hits (those with an E-value lower than a user-specified threshold) are chosen and converted to a Position-Specific Scoring Matrix (PSSM) from which a multiple sequence alignment is built. The PSSM includes details about both the matrix substitutions and the query sequences. The PSSM contains processed data about the query sequence and the substitution matrix. These are incorporated in the ensuing round of database searches.

2.14.3 Profile Hidden Markov Models (pHMM)

When a lack of a statistically significant match with proteins of known structure or function from BLAST, PSI-BLAST, or other database search methods is not found, it is possible to use more comprehensive methods that include the details found in "match", "insert" and "delete" probabilities in a statistical framework. With a detailed statistical summary of the underlying alignment, a profile HMM is created. An HMM column covers the probabilities of each of the 20 amino acids created for each column in the various alignments that have a residue in the query sequence. These insert/delete and match probabilities are converted into position-specific gap penalties when an HMM is aligned to another HMM or sequence. The query HMM is then matched

with each HMM in the selected database. The database HMMs have been pre-calculated and also contain secondary structure processed data projected by secondary structure predictors. The database search can be performed using the optimal log odd alignment Viterbi algorithm, the consequent likelihood can be compared with the random model (Eddy, 1998; Zimmermann *et al*., 2018).

2.15 STRUCTURE-BASED ALIGNMENTS

Structural alignment aims to find homology between two or more structures based on their three-dimensional structures. Structural alignment is usually applicable to tertiary DNA, RNA, ligands, and protein structures. Structural alignment does not require prior knowledge of equivalent residues of the two structures (Zhang Yang & Skolnick, 2005b). The reason for structural alignment is that nature is conserved and has evolved by small progressive modifications instead of major modifications that can alter the protein function. Consequently, by searching for remotely homologous proteins, similarities can be detected. Protein structural alignment has an advantage over sequence-based alignment because amino acid residues sequences of proteins positioned far apart in the primary sequence can be closely juxtaposed in the three-dimensional structure. Thus modeling protein structure using tools such as iterative- threading assembly refinement (I-TASSER) is important in bringing sequences far apart into close juxtaposition (Jianyi & Yang, 2015; Roy *et al*., 2010; Zhang Yang, 2008). In addition, the structure is considered to be more preserved than the sequence, since different sequences can potentially fold into the same three-dimensional structure despite having different amino acids. Moreover, the three-dimensional structure of proteins can find the specific protein fragments essential to the functionality of a protein compared to the primary sequence.

2.16 SIMILARITY METRICS USED FOR 3D PROTEIN COMPARISONS

Given a scoring scheme and two proteins, the goal of any structure alignment method is to find the maximal score of a range of comparable structures or substructures within two proteins. Two scoring schemes have been used in this study; The Root Mean Square Deviation (RMSD) and the Template Modeling-score (TM score).

2.16.1 Root Mean Square Deviation (RMSD)

The RMSD is a commonly used structural scoring similarity metric. To calculate the RMSD, an optimal rotation and translation is needed to superimpose the complete structure or substructures of biological structures and vice versa. A formula based on the Kabsch algorithm (Kabsch, 1976) is used to calculate this optimal superposition of RMSD of the superposed structures. The RMSD is defined as:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

Where δ_i is the distance between atom *i* and a reference structure. The RMSD is often calculated for the backbone heavy atoms C, N, O, and C-alpha (Cα) or sometimes just the Cα atoms in proteins.

2.16.2 Template Modeling-Score (TM-score)

The TM score is a metric for calculating the similarity between two protein structures (Zhang & Skolnick, 2004). The main advantage of the TM-score over the RMSD in the measurement of structural similarity lies in its ability to account for both the structural similarity in the aligned

regions and the alignment coverage using a single parameter. The TM-score is more robust to the overall topology of the protein structures than RMSD since it weights smaller distances more than larger ones between aligned C-alpha (Cα) pairs. For RMSD comparisons, all distances are weighted equally as a result of local errors in the model (such as an incorrectly oriented tail) which can result in a large RMSD value even if the global topology of the two structures is similar. The TM score is defined as:

$$\text{TM-score} = \max \left[\frac{1}{L_{\text{target}}} \sum_{i}^{L_{\text{aligned}}} \frac{1}{1 + \left(\frac{d_i}{d_0(L_{\text{target}})} \right)^2} \right]$$

Where L_{target} *is* the length of the target protein.

L_{aligned} is the number of equivalent residues in two proteins.

d_i is the distance of the *i*-th pair of the equivalent residues between the two structures, which depends on the superposition matrix;

'max' means the procedure to find the optimal superposition matrix that maximizes the sum.

$d_0(L_{\text{target}}) = 1.24\sqrt[3]{L_{\text{target}} - 15} - 1.8$ is defined to normalize the TM-score in a way that the magnitude of the average TM-score for random protein structures does not depend on the size of the proteins. The TM-score range is between [0-1] a higher value indicates a stronger similarity.

2.17 TM-ALIGN

TM-align is a structural alignment algorithm for the comparison of proteins whose sequences may be different (Zhang Yang & Skolnick, 2005b). TM-align initially generates optimized residue-to-residue superposition starting from a range of initial alignments, which are then modified based

on heuristic iterations of the Needleman-Wunsch dynamic programming, the similarity matrix used is defined as:

$$S(i,j) = \frac{1}{1 + d_{ij}^2 / d_0(L_{\min})^2}$$

Where; S (i,j) is defined such that d_{ij} is the distance of the i^{th} residue in structure 1 and the j^{th} residue in structure 2 under the TM-score superposition.

$d_0(L_{min}) = 1.24\sqrt[3]{L-15} - 1.8$, L_{min} is the length of the smaller protein.

The optimal superposition of the two structures is based on the detected alignment, as well as the returned TM-score value that defines the structural similarity.

2.18 TM–FOLD

The RMSD and TM scores give measures of similarities between protein structures. These score metrics cannot calculate the statistical significance of the protein structure alignments which is critical in various statistical studies. As a result, a quantitative relation between the conventional fold classifications and the score measures ends up lacking (Levitt & Gerstein, 1998). Although distance-matrix alignment (DALI) (Holm & Rosenström, 2010) and matching molecular models obtained from theory (MAMMOTH) (Ortiz *et al.*, 2009) provide *Z*-score or p-value as a measurement of the significance of sequence-independent structural alignments, they do not qualitatively describe the probability of two protein structures sharing the same or different topologies/folds in known databases. The TM-fold algorithm used in this study is designed for automated protein structure classifications. TM-fold calculates the structural similarity for a given

pair of protein structures using TM-align and outputs the posterior probability for the structures belonging to the same CATH and SCOP fold family.

2.19 NEAR COMPLETENESS OF THE PDB

Even though over 150 million protein sequences have been deposited in the UniProtKB databases as of May 2020, this increase does not reflect biological knowledge since most sequences lack functional annotation. The functional annotation is important for a better understanding of physiological processes and biological roles of the genes/sequences. However, a critical task in modern cell and molecular biology is accurate functional annotation. In line with this, Nuclear Magnetic Resonance (NMR) and X-ray crystallography are the most commonly used methods for determining the atomic resolution of protein structures. Knowledge from these two methods may be used to infer function. Nonetheless, predictions of protein structure from these two methods are much too slow and expensive. Additionally, other experimental approaches to determine protein function (like gene knockout, targeted mutation, and inhibitions of gene expression studies), in nature are low throughput. As such, the capacity to generate sequence data far exceeds the rate at which structural and functional knowledge can be generated. Thus, the limitations of experiments for the prediction of protein structures have stimulated structural bioinformatics which improves the efficacy of the prediction of experimental structures as a means of inferring molecular function. From various studies in structural bioinformatics (Kihara & Skolnick, 2003; Zhang Yang *et al*., 2006; Zhang Yang & Skolnick, 2005a), it has been observed that the percentage of new fold entries in the current PDB library has continued to decrease. For example, the percentage of novel folds was 27% in 1995, but 5% in 2001 (Zhang Yang *et al*., 2006). These studies in retrospect have elucidated the dynamics of protein structure space and implications for protein structure prediction.

Thereby highlighting that there is a likely limited collection of protein topologies, such that the library of solved protein structures in the PDB is nearly complete (Kihara & Skolnick, 2003; Zhang Yang *et al*., 2006; Zhang Yang & Skolnick, 2005a). For instance, a representative sampled benchmark set of proteins containing 200 residues or smaller showed that 93% can be structurally aligned to a single template structure within the PDB with a mean sequence identity of 9.8 having an RMSD less than 4 Å and a protein mean coverage of 79%. On the other hand, for proteins up to 320 residues long, a 90% coverage below an RMSD of 3.5 Å can be achieved (Kihara & Skolnick, 2003). Therefore, the uncharacterized protein structure prediction can be reduced to a structural/topological classification problem for preliminary annotation based on the near completeness of the PDB.

2.20 COMPUTATIONAL APPROACHES FOR PROTEIN TERTIARY STRUCTURE PREDICTION

The aim of predicting the three-dimensional structure of a protein is to evaluate the spatial position of every atom of a protein in a three-dimensional space. The methods used to achieve this can be classified into three categories: homology modeling (Sali & Blundell, 1993), fold recognition or threading (Bowie *et al*., 1991), and *ab initio* modeling (Bradley *et al*., 2005).

Comparative modeling, also known as homology modeling, involves the construction of an atomic-resolution model of the target protein amino acid sequence from an experimentally preexisting three-dimensional structure/s of a related homologous protein/s known as a template. The amino acid residue equivalence between the template and query is achieved by aligning

sequence profiles or sequences. Fold recognition is centered on two simple observations: that there is an almost finite repertoire of different folds in nature (approximately > 1300); and that 90% of the new structures submitted to the PDB in recent years have similar structural folds to those that are in the PDB (Kihara & Skolnick, 2003; Zhang Yang *et al*., 2006). Fold recognition approaches align the unknown protein sequence directly to three-dimensional structures of experimentally predetermined protein structures to identify similar protein folds using scoring functions, which may lack clear evidence of an evolutionary connection with the query protein sequence. When no good template cannot be detected by threading, the last resort for predicting the protein structure uses *ab initio* protein structure modeling. The *ab initio* protein structure prediction assumes the native structure of a protein matches its global free energy minimum (Anfinsen, 1973), and the conformational space is sampled to find this native structure as directed by accurate energy force fields.

Although these predictive techniques have come a long way in recent years, substantial efforts in finding a definitive solution may be achieved. Research in *ab initio* modeling can be roughly divided into fragment assembly and first principle-based approaches. Fragment assembly approaches attempt to assign a fragment with a known structure to a section of the unknown query sequence, the fragments of the query are then compiled and scored using the potential energy scoring functions to infer similarities with the experimentally solved structures, inter-residue contacts inferred from co-evolutionary signals in sequence homologs can significantly facilitate the fragment assembly based PSP (Jianyi *et al*., 2014; Jianyi & Yang, 2015; Rohl *et al*., 2004; Roy *et al*., 2010; Xu & Yang, 2012, 2013; Zhang Yang, 2008). The first principle approaches start with an unfolded conformation, usually surrounded by solvent, and allow simulated physical forces to fold the protein as would normally happen *in vivo* (Alder & Wainwright, 1959, 1960).

Nonetheless, the boundaries between the traditional categories of comparative modeling, fold recognition, and *ab initio* methods have turned out to be fuzzy. For example, both homology modeling and fold recognition use profile-profile and sequence-profile alignments for finding templates. Likewise, up-to-date *ab initio* approaches like I-TASSER use evolutionary information to establish spatial constraints or to find local structural secondary structure elements. Recently, Critical Assessment of Structure Prediction (CASP) studies have revealed the important benefits of using combined methods in predictions of protein structure which have won numerous CASP competitions. (Moult *et al*., 2018).

2.21 CRITICAL ASSESSMENT OF TECHNIQUES FOR PROTEIN STRUCTURE PREDICTION (CASP)

The Critical Assessment of techniques for protein Structure Prediction (CASP) competition provides a standardized podium to judge modern algorithms for the structure and function prediction of proteins. The predicted models from the algorithms are then compared to experimental results obtained to ascertain the accuracy of the algorithms. Iterative-Threading ASSEmbly Refinement (I-TASSER), a three-dimensional structure prediction server that predicts both the structure and function of proteins that share low sequence identity with other identified structures has consistently been ranked first in CASP7 (Battey *et al*., 2007), CASP8 (Kryshtafovych *et al*., 2009) CASP9 (Mariani *et al*., 2011), CASP10 (Huang *et al*., 2014) CASP11(Kinch *et al*., 2016) and CASP12 (Moult *et al*., 2018) community-wide double-blind experiments to benchmark protein fold recognition. Other well-performing 3D structural prediction programs include raptor (Ma *et al*., 2013) and the recent deep-learning alpha fold based on CASP rankings (Senior *et al.,* 2019, 2020).

2.22 LOCAL META-THREADING-SERVER (LOMETS)

Local Meta-Threading-Server (LOMETS) is a meta-server for PSP that links an array of multiple threading servers (Wu & Yang, 2007). From individual threading algorithms templates are ranked by several structure-based scores and sequence-based scores. The top protein hits are then picked for further consideration in I-TASSER for protein structure prediction. Based on a normalized Z-score described below, the template alignment quality and the expected complexity of target modeling are evaluated:

$$Norm.\,Z\text{-score} = \frac{Z\text{-}score}{Z_0}$$

A Z-score cutoff, Z_0, is allocated to each algorithm to determine if the query is bad or good based on a large benchmarked test (Zhang Yang, 2008). If there are more than eight templates with Z_i greater than Z_0 (i.e. the mean count of good templates is >1 per threading program) the query is classified as an "Easy", which normally matches a close homologous protein. If there is no good template with Z_i greater than Z_0, then the query is classified as "Hard", which usually corresponds to a non-homologous protein. Otherwise, a query is defined as "Medium".

2.23 ITERATIVE-THREADING ASSEMBLY REFINEMENT (I-TASSER)

After the LOMETS threading process, continuous fragments are cut out from PDB template structures and used by I-TASSER to recreate a complete model that joins the secondary structure fragments (Jianyi & Yang, 2015; Roy *et al*., 2010; Zhang Yang, 2008). I-TASSER algorithm uses a reduced model to denote the amino acid chain, the amino acid residue is represented by its C-alpha (Cα) atom and its center of mass side-chain. This reduced model improves the effectiveness of the conformational search. Regions that are not aligned, such as loops and tails, regularly have lower modeling precision during the threading process. Therefore, structural modeling in these regions is limited to a grid size of 0.87 Å lattice. The aligned regions from the LOMETS threading normally have high precision. Modeling in the aligned regions from threading is kept off-lattice and the fragments from the templates are kept invariant throughout the modeling. This is important in preserving the high-resolution structures in aligned regions. Thereafter, modeling, the conformational space is searched using parallel hyperbolic Monte Carlo sampling, (Zhang Yang, 2008). The overall simulation is directed by a hybrid knowledge-based force field described below.

2.23.1 The I-TASSER Force Field

The I-TASSER simulations are directed by a knowledge-based force field that comprises, (1) statistical terms resulting from the PDB (C α and side-chain correlations, hydrophobicity, and hydrogen bonds). (2) An array of statistical long-range and short-range correlation terms derived from LOMETS (Wu & Yang, 2007). And (3) sequence-based support vector machine contact predictions from SVMSEQ (Wu & Yang, 2008).

2.23.2 Iterative Strategy by SPICKER

Following the first fragment assembly, to find low free energy state conformations, the low-temperature replicas are clustered by SPICKER (Zhang Yang & Skolnick, 2004). The I-TASSER Monte Carlo simulation begins from the carefully chosen SPICKER cluster centroids and is repeated, with in-built I-TASSER potential remaining unaffected in the subsequent run. Exterior restraints derived from assembling the initial high-confident restraints from threading alignments are then added. The contact and the distance restraints resulting from a combination of the centroid structures and PDB structures are detected by TM-align (Zhang Yang & Skolnick, 2005b).

2.23.3 Full Atomic Model Construction

The outputs of I-TASSER Monte-Carlo simulations and SPICKER clustering are reduced models, and each amino acid residue is denoted by the Cα atom and side-chain center of mass. To increase their biological meaningfulness, full-atomic models are constructed by the Repotology Modeling algorithm, (REMO) from these cluster centroids (Li & Zhang, 2009). The process optimizes the hydrogen bonding network, where the hydrogen bonding list is pre-constructed based on secondary structure predictions and the 3-D backbone model. Frequently, the models have distortions in side-chain atoms and structure. To improve the hydrogen bonding and local geometry, and to decrease backbone and side-chain steric clashes of the models generated by REMO. The models are forwarded to Fragment Guided-Molecular Dynamics (FG-MD) algorithm (Jian Zhang *et al*., 2011). The FG-MD simulations are executed in a vacuum in a Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) package (Plimpton, 1995). The force field comprises energy terms from Amber 99 (Cornell *et al*., 1995), statistical hydrogen-bonding potential, Cα repulsive potential, and distance restraints acquired from both the structural fragments and

templates searched by the TM-align algorithm. The original model is kept as a probe. Distance restraints are created by a combination of distance maps from the initial model. The FG-MD refinement simulation is the final step of the structure prediction. The function of the modeled query protein is inferred by structurally matching the predicted 3D model against the proteins of known structure and function in the PDB using TM align (Zhang Yang & Skolnick, 2005b).

2.23.4 Assessment of the Accuracy of Predicted Model

The quality of assessment of a protein model is critical because it finally decides whether drug designers or biologists can use the predicted protein model in their study. The confidence score (C -score) is used to evaluate the accuracy of protein structure predictions. Mathematically expressed as:

$$\mathrm{C-score} = \ln\left[\frac{M}{M_{tot}} \times \frac{1}{\langle RMSD \rangle} \times \sum_{i=1}^{N} Norm.\,Z-score(i)\right]$$

Where *M* is a group of structure decoys in the structural clusters identified by SPICKER (Zhang Yang & Skolnick, 2004); M_{tot} is the total count of decoys from I-TASSER simulations. The RMSD refers to the average RMSD of clustered decoys to the centroid cluster; *Norm. Z-Score* (*i*)/ ($Zi/Z_{cut,}$) is the normalized Z-score of the highest-scoring threading alignment from the i^{th} threading server in LOMETS, and *N is* the number of servers used in LOMETS. C−score values range from −5 to 2. A higher score represents a protein model of better quality.

2.24 DNA STRUCTURE

The double-stranded Deoxyribonucleic acid (DNA) is comprised of, guanine (G), cytosine (C), adenine (A), and thymine (T). In a single strand, these nucleotides are linked via a phosphodiester bond. The two single DNA strands are held together by hydrogen bonds, G and C have three hydrogen bonds while A and T have two hydrogen bonds. The major and minor grooves arise from the antiparallel arrangement of the two backbone strands opposite each other (Figure 2.3). The grooves (minor and major) are structural features of the DNA molecule which allow the binding of proteins and small molecules that recognize specific DNA sequences leading to conformational changes during gene expression.

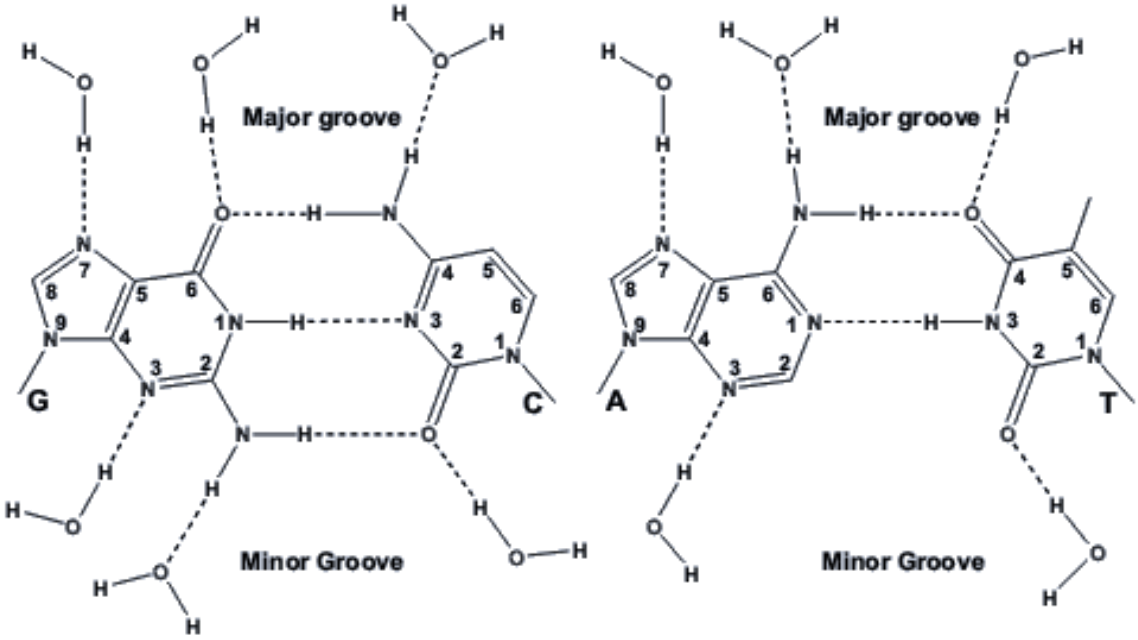


Figure 2.3 Illustration of the minor and major groove sides of DNA

2.25 DNA-MINOR GROOVE BINDERS

DNA minor groove binders are small molecules that selectively bind to the minor groove side of DNA. Binding to DNA is usually controlled by a combination of forces comprising; hydrogen bonding, van der Waal interactions, and electrostatic interactions (Kopka *et al*., 1985). DNA minor groove binders can be categorized into two functional ligand groups, those that cause irreversible DNA damage through covalent modification and those that react with DNA non-covalently and induce reversible inhibition of DNA- dependent function. The first group is markedly cytotoxic while the latter is marginally cytotoxic and can be modified to make practical drug candidates. Congocidine and distamycin A have served as models for minor groove binders (MGBs) for reversible inhibition, and are described to have antiviral properties that can inhibit the replication of some DNA- viruses in cell cultures both *in vitro* and animals *in vivo* (Becker *et al*., 1972). DNA minor groove binders, though not studied in ASFV, constitute an important class of compounds potentially capable of inhibiting ASFV due to their ability to bind the minor grooves of DNA motifs. These DNA minor groove binders selectively bind to motifs with ranging binding affinities (Abu-daya *et al*., 1995; Abu-daya & Fox, 1997; Portugal & Waring, 1987).

2.25.1 Distamycin A as a Model for a Minor Groove Binding Antiviral

Distamycin A (Figure 2.4) is a fermentation product of *Streptomyces distallicus* (Figure 2.4). Distamycin A can also be chemically synthesized in the laboratory (Bialer *et al*., 1978, 1980). Distamycin A contains three N-methyl-pyrrole units and can bind to AT-rich DNA sequences (Kopka *et al*., 1985). Distamycin A has been shown to inhibit transcription in vaccinia (Becker *et al*., 1972; Broyles *et al*., 2004). However, distamycin exhibits high cytotoxicity, hence its use as a practical drug candidate capable of separating antiviral activity from cytotoxicity is limited (Bialer *et al*., 1980; Broyles *et al*., 2004; Lombardi & Crisanti, 1997)

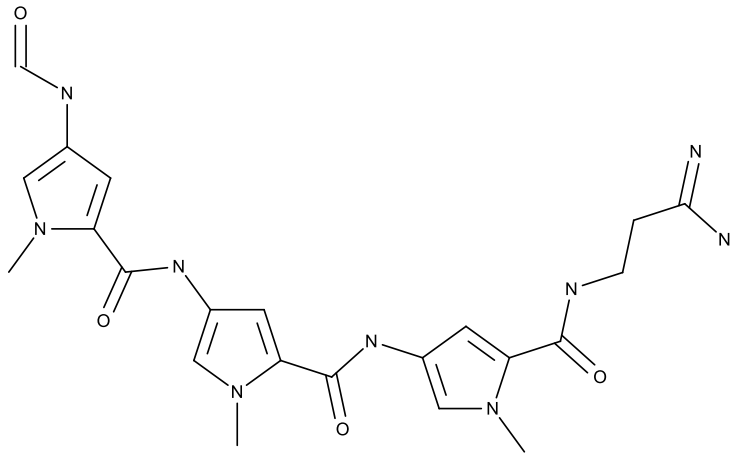


Figure 2.4 Structure of Distamycin A

3-((4-((4-((4-formylamino-1-methyl-1H-pyrrol-2-yl)-formylamino)-1-methyl-1H-pyrrol-2-yl)-formylamino)-1-methyl-1H-pyrrol-2-yl)-formylamino)-propanimidamide

2.25.2 Congocidine as a Model for Minor Groove Binding Antiviral

Congocidine is an oligopeptide antibiotic isolated as a fermentation product of *Streptomyces netropsis* (Figure 2.5). Congocidine is reported to possess antiviral properties that inhibit the replication of deoxyribonucleic acid (DNA)-containing viruses in cell cultures *in vitro* (Becker *et al*., 1972). Like distamycin A, congocidine is highly cytotoxic (Bialer *et al*., 1980). Studies on the mechanism of action of congocidine have demonstrated that the antibiotic molecule interacts with (A-T) rich regions in DNA (Abu-daya *et al*., 1995). The DNA binding pattern of congocidine is substantially similar to that observed in distamycin A with an influence of the electrostatic binding component due to the presence of two cationic ends (Kopka *et al*., 1985).

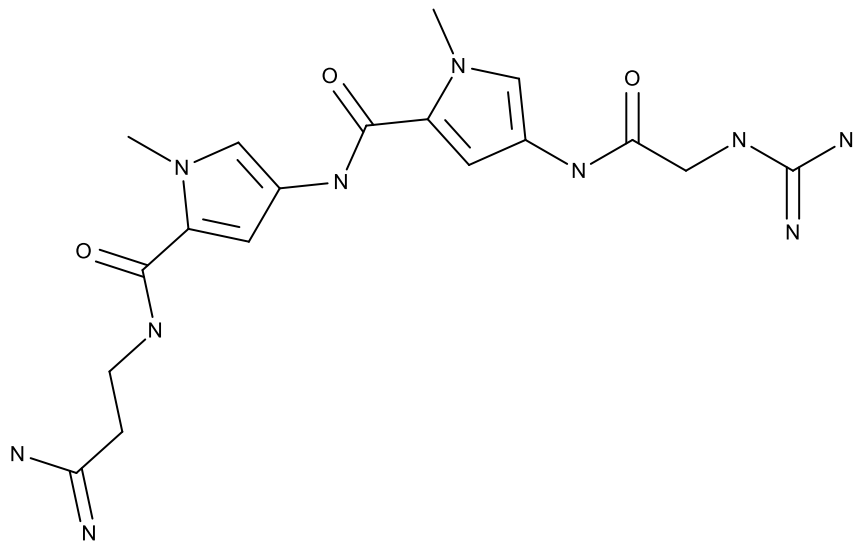


Figure 2.5 Structure of congocidine

3-((4-((4-(2-(amino-imino-methylamino)-acetylamino)-1-methyl-1H-pyrrol-2-yl)-formylamino)-1-methyl-1H-pyrrol-2-yl)-formylamino)-propanimidamide

2.25.3 Congocidine Congeners as Antivirals

Congocidine congeners (congocidine 2 and congocidine 3) used herein refers to compounds that are similar to congocidine in their chemical structures but have a different number of N-methylpyrrole rings. Congocidine has two N-methylpyrrole rings (Figure 2.5) while the congocidine congeners used in this study have three N-methylpyrrole rings (Figure 2.6). Additionally, the terminal end for congocidine 2 is an ethanimidamide moiety, while congocidine 3 has a butanenitrile moiety. The guanidinium ends of these compounds are also identical. Both congocidine 2 and congocidine 3 have been tested for cytotoxicity (Bialer *et al*., 1980). The butanenitrile-containing derivative demonstrates non-cytotoxicity while the ethanimidamide-containing derivative is mildly cytotoxic to cultured cells. The toxic effect is regarded as cell death within 2 days of incubation of cultured cells at 37 degrees celsius (Bialer *et al*., 1980).

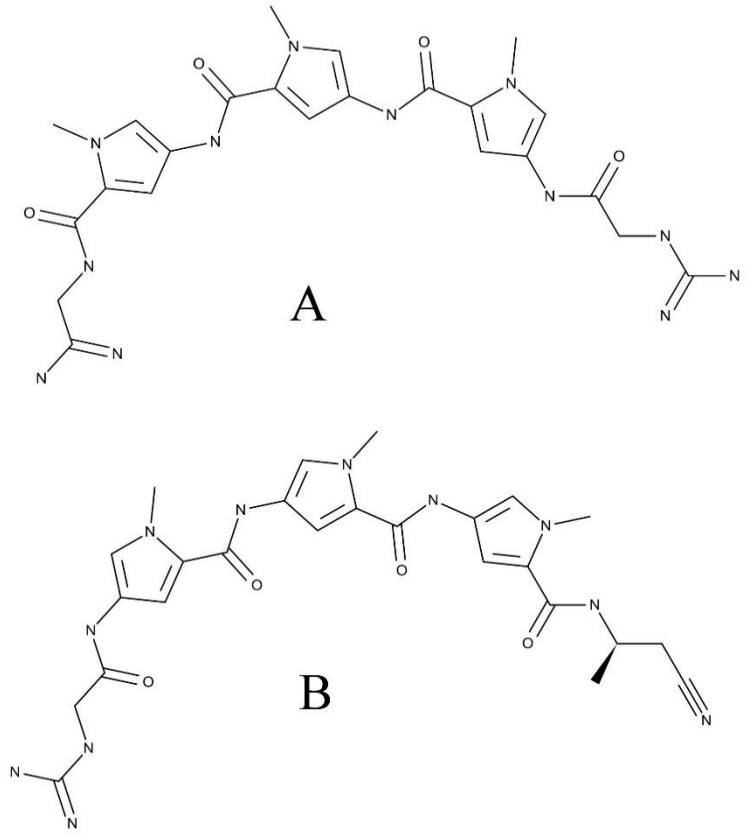


Figure 2.6 Structures of Congocidine 2 (A) and Congocidine 3 (B)

A: Congocidine 2 (2-[(4-[(4-[(4-[2-(amino-imino-methylamino)-acetylamino]-1-methyl-1H-pyrrol-2-yl)-formylamino]-1-methyl-1H-pyrrol-2-yl)-formylamino]-1-methyl-1H-pyrrol-2-yl)-formylamino]-ethanimidamide)

B: Congocidine 3 (3-[(4-[(4-[(4-[2-(amino-imino-methylamino)-acetylamino]-1-methyl-1H-pyrrol-2-yl)-formylamino]-1-methyl-1H-pyrrol-2-yl)-formylamino]-1-methyl-1H-pyrrol-2-yl)-formylamino]-butanenitrile)

2.25.4 Bioisosteres

In pharmaceutical chemistry, bioisosteres are compounds resulting from the substitution of a group of atoms or an atom with another having similar chemical or physical properties that produce roughly analogous biological properties to another chemical compound. Bioisosterism can be used to minimize toxicity, modify bioavailability, or alter the activity or metabolism of a lead compound. The main bioisosteres used in this study are tris-benzimidazole and Hoechst 33258 (Figure 2.7). Tris-benzimidazole was chosen because of the shape resemblance to congocidine congeners, while Hoechst 33258 was chosen because of its iso-geometrical nature to congocidine (Stockert, 2014). Additionally, Hoechst 33258 has previously been tested in a phase I-phase II advanced pancreatic carcinoma study in humans, where it was notably well-tolerated in both phases (Patel *et al*., 1991).

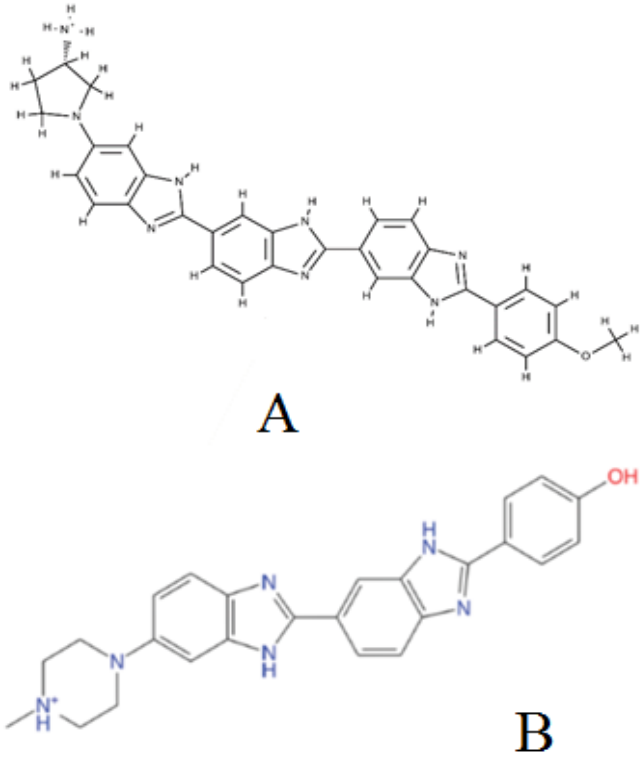


Figure 2.7 Structures of tris-benzimidazole (A) and Hoechst 33258 (B)

2.26 DOCKING STUDIES USING INTERNAL COORDINATE MECHANISM

Internal Coordinate Mechanism (ICM) docking algorithm is a global optimization process exploring the conformational space of the ligand in the locality of a receptor (protein or nucleic acid) based on Monte Carlo minimization (Abagyan *et al*., 1994). The receptor is kept rigid and is described by pre-calculated grid potential maps, defining the interaction energy between the receptor and the probe at each grid point. Five grid potentials are normally calculated, within a grid spacing of 0.5 Å, hydrogen bond grid potential, van der Waals grid potential for hydrogen probes, van der Waals grid potential for non-hydrogen probes, electrostatic grid potential, and hydrophobic grid potential (Neves *et al*., 2012). The ICM algorithms consist of the following steps: 1) Random orientation and conformational change of the ligand, where the ligand is placed and rotated around its center of gravity by pseudo-Brownian random movements. 2) Local energy minimization of analytically differentiable terms using the conjugate gradient minimization process. 3) Determination of non-differentiable terms and the addition of the terms to the local energy minimum obtained and 4) Applying of the Metropolis-Hastings selection criteria to determine if to reject or accept the proposed conformation. If the determined probability is higher than the previously obtained probability it is accepted and used as the new energy criterion. A move to the next state can also be accepted if it is less than a randomly generated probability, otherwise, it's rejected. The probability (P) is calculated as follows:

$P = e^{-\Delta G/kT}$

Where ΔG, is the change in energy of the two conformations

k is the Boltzmann constant

T is the temperature

Global optimization is performed from multiple starting points (Neves *et al*., 2012). Ligand-receptor docking conformations are ranked by a scoring function based on the most energetically favorable. The quality of the ligand binding complexes is ranked using the ICM scoring function. The ICM scoring function is a summation of the energy changes when the ligand docks to a receptor, given by:

$$\Delta S_{bind} = \Delta E_{IntFF} + T\Delta S_{Tor} + \alpha_1 \Delta E_{HBond} + \alpha_2 E_{HBDesol} + \alpha_3 E_{SolEl} + \alpha_4 E_{Hphob} + \alpha_5 Q_{Size}$$

Here ΔE_{IntFF} represents the van der Waals interaction change due to ligand-receptor binding and the internal force-field energy of the ligand, $T\Delta S_{Tor}$ is the free energy change of the conformational entropy. The overall ΔS_{bind} is weighted by (α_1–α_5). The ΔE_{HBond} represents a hydrogen bond term. Further, $\Delta E_{HBDesol}$ represents the disturbance of hydrogen bonds with solvent on docking, ΔE_{SolEl} represents the solvation electrostatic energy alteration upon binding, ΔE_{HPhob} represents hydrophobic free energy gain, and Q_{size} is the ligand size correction term. The ICM is used in this study to predict the docking pose of ligands to the DNA receptor.

2.27 FREE ENERGY CALCULATION OF DNA-DRUG BINDING

DNA-drug binding may be described, as illustrated in Figure 2.8 below;

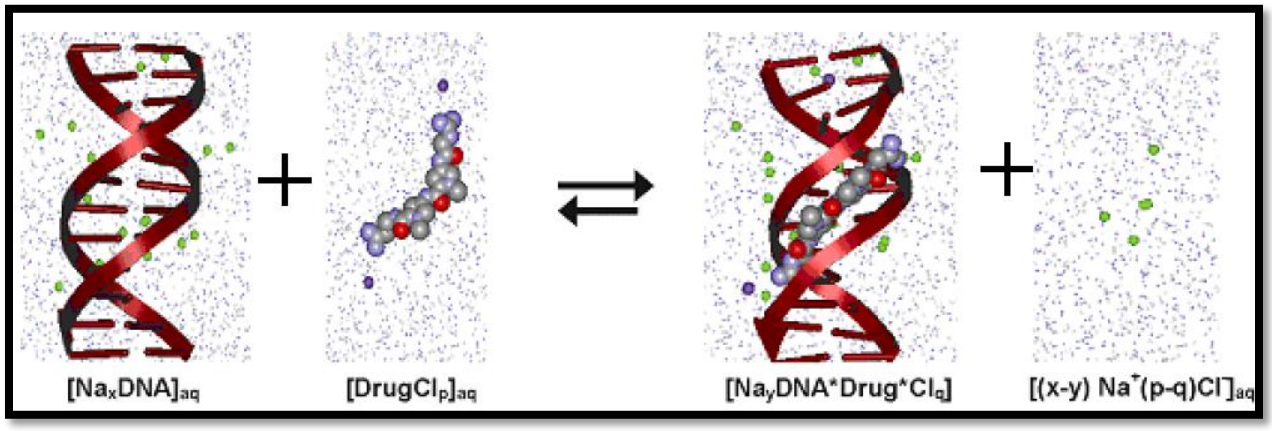


Figure 2.8 A representative scheme of ligand binding reactions in an aqueous environment within the minor groove of DNA. Image courtesy of (Shaikh *et al*., 2004)

In an aqueous setting, the ligand is often charged and DNA is polyanionic. The associated counter ions are close to the charged groups and are partially solvated. When ligand-docking occurs, the net result is the solvent displacement from the binding site on both the ligand and DNA due to the partial compensation of opposite charges between the ligand and DNA. Additionally, the docking process is linked to structural deformation and adaptation of the DNA and the ligand to suit each other. All these processes are linked with some energetic gains and losses that could give a crucial relation between the structure and function of molecular systems. Correlating binding affinities with structural features of drug-ligands complexes allows the design of improved minor groove binders. Therefore, an algorithm that can estimate the binding affinity with high precision is very important in drug design activities. Studies have shown novel ligands are being designed to target

specific DNA motifs, such as alkylfurans (Trent *et al*., 1996), bis-benzimidazole derivatives (Wang *et al*., 2014), and polyamides (Bignon *et al*., 2017). Concomitant with the increase in ligand numbers, numerous energy scoring functions based on knowledge-based methods, empirical methods, and force fields have been described in the literature for protein-DNA, protein-ligand, and protein-protein interactions (Schulz-Gasch & Stahl, 2004). These have been widely used for virtual ligand screening, docking, and binding affinity prediction. However, efforts for DNA-ligand binding affinity predictions have not been exploited fully (Shaikh & Jayaram, 2007). Approaches like Molecular Mechanics Poisson–Boltzmann Surface Area (MMPBSA) (Genheden & Ryde, 2015) and Molecular Mechanics/Generalized Born Surface Area (MMGBSA) (Genheden & Ryde, 2015) have been used to calculate and predict binding free energy of biomolecular systems. Nonetheless, these methods show limited success for ligand-DNA complexes in mimicking experimentally verified binding affinities (Spacková *et al*., 2003). Consequently, for binding affinity predictions of ligand-DNA complexes, the choice of a reproducible and accurate algorithm is critical.

The preddicta algorithm, a computational protocol for free energy calculation and specific for reversible minor groove binders has been tested in terms of prediction accuracy on DNA-ligand complexes. Free energy calculation results from the preddicta algorithm have shown that the calculated binding energies accurately approximate experimental binding free energies (Shaikh & Jayaram, 2007). The energy function utilized by preddicta calculates the electrostatics, van der Waals, rotational and translational entropy, and hydration-free energy change for the DNA-ligand complex. These are then summed up to produce the calculated binding energy ($\Delta G°_{cbe}$) (Shaikh & Jayaram, 2007). The preddicta protocol correlates binding affinities with structural features of

DNA-ligand complexes (single structures) to test if ligands have enhanced binding affinities compared to known binding affinities, this is usually enough. Nonetheless, molecular dynamics simulations can be further used to assess test ligand behavior and patterns while docked to the DNA-duplex in a dynamic environment. Thus, the two techniques can complement each other. The preddicta protocol used here to predict binding affinities for reversible minor groove binders has been benchmarked on a large dataset and found to have accuracy levels above 90% (Shaikh & Jayaram, 2007). The study thus found it reasonable to use this approach because of its accuracy and its high reproducibility for the quantitative estimation of binding affinities, specifically for reversible minor groove binders.

2.27.1 Vander Waal Term ΔH°_{vdw}

Vander Waal term ΔH°_{vdw} calculates direct van der Waals contributions from ligand -DNA interactions and desolvation van der Waals due to the interaction of solvent with ligand, complex and free DNA. For a single time-averaged crystal minimized model-built structure,

$\Delta H^{\circ}_{vdw} = \Delta E^{\circ}_{vdw}$ and is calculated as, $\Delta E^{\circ}_{vdw} = E_{vdwdir} + E_{vdwdsl}$.

Direct van der Waals interactions are modeled using a Lennard - Jones potential between the atom i of DNA and j of the ligand as follows:

$$E_{\text{vdwdir}}\,(\text{kcal/mol}) = \sum_i \sum_j \left[\frac{C_{12}^{ij}}{r_{ij}^{12}} - \frac{C_6^{ij}}{r_{ij}^{6}} \right]$$

The desolvation van der Waals component is calculated as:

$$E_{vdwdsl}\,(\text{kcal/mol}) = -\,0.0398 \times \Delta A$$

Where ΔA is the accessible surface area loss (in $Å^2$) of the ligand and DNA upon binding.

2.27.2 Electrostatic Term $\Delta H°_{el}$

The electrostatic term, $\Delta H°_{el}$ accounts for interactions between partial atomic charges on the ligand and DNA calculated from Coulomb's law using a sigmoidal dielectric function for solvent screening and scaled phosphate charges based on the manning theory (Manning, 1978). For a single time-averaged minimized model-built structure, $\Delta H°_{el} \approx \Delta E°_{el}$. For atom i of DNA and j of the ligand are determined using the equation below:

$$E_{el}\,(\text{kcal/mol}) = \frac{\sum_i \sum_j 332 q_i q_j}{D(r) r_{ij}}$$

Where q_i and q_j are the partial atomic charges on atoms i and j, r_{ij} is the i - j distance,

D(r) is a sigmoidal dielectric function.

2.27.3 Rotational and Translational Entropy $T\Delta S°_{rt}$

Rotational and translational entropy $T\Delta S°_{rt}$ represents the rotational and translational entropy changes for complex formation (Shaikh & Jayaram, 2007). The translational partition function is calculated as; $Q_{trans} = (V/h^3)\,(2\pi m/\beta)^{3/2}$

Where V is the volume, $\beta=1/kT$, where k is the Boltzmann constant, T is the temperature, and m is the mass. The rotational partition function is computed as $Q_{rot} = \sigma^{-1}\,(1/\beta hc)^{3/2}\,(\pi/ABC)^{1/2}$

Where σ is the symmetry number and A, B, and C are the rotational constants of the molecule.

The Sackur-Tetrode equation is employed for $\Delta S°_{rt}$ estimation

$S^o = k \ln Q - (1/T)\,(\partial \ln Q/\partial a\beta)_V$.

2.27.4 Hydration Term $\Delta G°_w$

The hydration term $\Delta G°_w$ is responsible for the rearrangement of waters surrounding ligand and DNA upon binding, the hydration term is calculated from the change in hydration heat capacity (ΔCp) as shown in the equations below:

$$\Delta G_w \text{ (kcal/mol)} \approx T\Delta S = 0.080 \times \Delta Cp \text{ (cal mol}^{-1}\text{K}^{-1}\text{)}$$

$$\Delta Cp \text{ (cal mol}^{-1}\text{ K}^{-1}\text{)} = 0.382 \sum_{np} \Delta A_{np} - 0.121 \sum_{p} \Delta A_{p}$$

Where ΔA_{np} and ΔA_{p} represent the changes in the accessible molecular surface area (ASA, in Å units) for nonpolar (C, P, and H attached to C) and polar (N, O, and H attached to N and O) atoms, respectively. The above relation yields a high correlation between calculated and experimental ΔCp for DNA - ligand systems (Ren *et al*., 2000; Spolar *et al*., 1989).

2.28 MOLECULAR DYNAMICS SIMULATIONS IN A NUTSHELL

Molecular dynamics (MD) evaluates the motions of particles owing to their interactions. Many models describe these interactions including molecular mechanics, quantum mechanics, coarse-grained models, and more. Through MD simulations, researchers can study and experiment with a molecular model to explain biological processes and simulate macroscopic properties by observing atomic movements and their interactions (Bowers *et al*., 2006). The MD simulations contain explicit treatment of water and intrinsic receptor flexibility, capturing the dynamic nature of macromolecules. The MD simulations normally have non-bonded atoms interacting and bonded atoms interacting. Non-bonded atoms interact due to electrostatic and van der Waal interactions. Bonded atoms interact as a result of dihedral torsions, angle bending, and bond stretching. Interactions can be described by the potential function between the particles and the force resulting from the potential changes in the positions of particles. The movement of molecules is governed by Newton's second law of motion commonly known as F = ma where F is the force, m is the mass of the atom, and a is the acceleration of the atom. Datum analysis of MD simulation requires considerable technical skills besides the information of the molecular system being studied combined with analyzing and animating the trajectories. The information generated from simulation information is useful in creating an interaction hypothesis or pattern between a drug-like molecule and a macromolecule (Shahbaaz *et al*., 2019).

2.28.1 Ligand-RMSD and Ligand-RMSF

Movements at an atomic level can be captured using the simulation interaction diagram tool used to analyze interactions after an MD simulation is complete (Bowers *et al*., 2006). The results of MD simulation can then be organized in plots and figures for easy analysis in form of L-Root

Mean Square Deviation (L-RMSD) and L-Root Mean Square Fluctuation (L-RMSF). The ligand-RMSD is used to measure the mean change in the displacement of all of the atoms in a particular frame to a reference frame (Bowers *et al*., 2006). Generally, RMSD is calculated for all frames in the trajectory and represented as a pictorial graph to show fluctuations of the entire molecule. The RMSD for a given frame x is defined as follows:

$$RMSD_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i'(t_x)) - r_i(t_{ref}))^2}$$

Where N is the number of atoms in the atom selection,

t_{ref} is the reference time (typically, the first frame is used as the reference and is regarded as time t = 0),

r' is the position of the selected atoms in frame x after superimposing on the reference frame, where frame x is recorded at time t_x. This procedure is repeated for every frame in the simulation trajectory. The L-RMSD can indicate how stable the ligand is with respect to the receptor, as well as the evolution of its internal conformation. The L-RMSF is used in the visualization of segments /heavy atoms along the macromolecule that fluctuates the most during the simulation. The metric characterizes changes in the ligand atom positions. The L-RMSF for atom i is:

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(r_i'(t)) - r_i(t_{ref}))^2},$$

Where T is the trajectory time over which the RMSF is calculated

t_{ref} is the reference time (usually for the first frame, and is regarded as *zero* at time t = 0)

r is the position of atom i in the reference at time t_{ref} ,

r' is the position of atom i at time t after superposition on the reference frame.

2.29 MACROPHAGE TARGETING USING LIPOSOMES

The ability of ASFV to infect, evade the phagocytic system of the host, replicate, and create pathogen reservoirs in macrophages that disseminate throughout the body, emphasizes the need for the development of therapeutics to target macrophages. As the function of monocytes and macrophages becomes better understood, strategies for targeting these types of cells are increasingly important. Liposomes naturally target macrophages (Kelly *et al*., 2011). Therefore, liposomes can be harnessed in drug delivery to target macrophages where ASFV predominantly replicates.

The natural targeting of macrophages can be enhanced by controlling the physicochemical properties of the liposome such as charge size and lipid composition. A variety of ligand-mediated liposome strategies designed to target mononuclear phagocytic system cells have been tested. Some include antibodies, peptides, and lectin coating directly attached to drug-loaded liposomes to target many of the several receptor forms expressed in macrophages and monocytes. Furthermore, it has been shown that negatively charged lipids have enhanced macrophage internalization (Ahsan *et al*., 2002). Liposomes are also known to reduce cytotoxicity and increase specificity (Staroverov *et al*., 2008; Tang *et al*., 2014). There are significant liposomal packaging efforts involving minor groove binders like distamycin A, a class of compounds displaying antiviral activity. However, their main disadvantage is leakage of the highly soluble forms (Cortesi *et al*., 2010).

The preparation of two liposomal formulations containing distamycin A and C16-Dist, an alkyl derivative has been described (Cortesi *et al*., 2010). Comparative analysis of the *in vitro* performance of the liposomal form (C16-Dist) and the free form (native distamycin A) showed that the encapsulation efficiency was 99.8% and 19.0%, respectively. Additionally, antiproliferative activity for native and the alkyl derivative of distamycin A was 11-fold and 8-fold on human leukemic K562 cells respectively. This activity was higher than corresponding drug-free liposomal formulations (Cortesi *et al*., 2004). The microencapsulation vesicle (MCV) method involves the preparation of liposomes of high encapsulation efficiency with homogeneous particle sizes. Liposomes that encapsulate water-soluble drugs and lipophilic drugs can be prepared using the (MCV) technique. The encapsulation efficiency of drugs (5-fluorouracil and ibuprofen) has been examined and the encapsulation efficiency revealed that the encapsulation efficiency of water-soluble 5-fluorouracil was 12–15% while lipophilic ibuprofen was around 90% (Nii & Ishii, 2005). The encapsulation efficiency was found to strongly correlate to the LogP (octanol/water), and the LogP (chloroform/water) (Nii & Ishii, 2005). This study used *in silico* LogP prediction as a means of predicting the ability of the ligands to be encapsulated into liposomes and thus predict a possible accumulation in macrophages for a specific release and minimize potential toxicity.

CHAPTER THREE

IN SILICO IDENTIFICATION OF AN ASFV UNCHARACTERIZED PROTEIN THAT HAS FEATURES THAT CAN BIND TO THE PROMOTER REGION OF TATA-LIKE ELEMENTS IN THE ASFV GENOME

3.1 INTRODUCTION

The ASFV is a DNA virus that is double-stranded containing a large envelope. This virus is a member of the *Asfivirus* genus, Asfarviridae family, and order Asfuvirales (Sánchez-Vizcaíno *et al*., 2019). Warthogs, bush pigs, and soft ticks of the genus *Ornithodoros* are the natural hosts of ASFV (Carrillo *et al*., 1994). In spite of extensive research on ASFV and its devastating effects, there are no licensed vaccines or treatments for controlling ASF infection, which poses a significant danger to the global pig industry (O'Donnell *et al.*, 2016; Reis *et al.*, 2016).

The complete ASFV genome encodes between 150 and 170 open reading frames (ORFs) along both strands of the viral genome (Dixon *et al*., 2013; Yáñez *et al*., 1995). The functions of about one-third of these ORFs remain unknown. Most of which are classified as uncharacterized proteins (Cackett, Matelska, *et al*., 2020; Dixon *et al*., 2013; Rodríguez & Salas, 2013; Yáñez *et al*., 1995). Moreover, a major limitation for developing effective antiviral treatments for ASFV is our limited knowledge of the ORF functions which is a significant barrier to understanding viral molecular mechanisms (Cackett, Sýkora, *et al*., 2020; Rodríguez & Salas, 2013).

Of particular interest in this study are late promoters which are enriched in adenine (A) and thymine (T) bases (Cackett, Matelska, *et al*., 2020; Rodríguez & Salas, 2013). These late promoter have TATA-like motifs and exist at or near the transcription start site (Cackett, Matelska, *et al*., 2020; García-Escudero & Viñuela, 2000; Oliveira *et al*., 2017, 2018; Rodríguez & Salas, 2013). The motifs TATTT and TATATA initiator region (Inr) have been identified.Additionally, putative ASFV RNA Polymerase subunits (pNP1450L, pEP1242L, pH359L, pD205R, pC147L, pD339L , C105R, and pCP80R) and transcription factors (pD1133L,pQ706L,pG1340L,pB962L, pC315R,pNP868R, and pI243L) involved in transcription have been reviewed (Rodríguez & Salas, 2013). However, a transcription factor that binds to TATA-like sequences has not been reported (García-Escudero & Viñuela, 2000). Therefore, the first objective of this study attempted to search for a protein with features of a TBP using *in silico* approaches.

3.2 MATERIALS AND METHODS

3.2.1 Sequence-Based Similarity Search for a Potential TBP

The proteome of ASFV strain BA71V containing uncharacterized proteins was obtained from UniProt using proteome ID UP000000624 (https://www.uniprot.org/proteomes/UP000000624) (Bateman *et al*., 2017). To identify a protein with characteristic features of binding to TATA-like sequences, the ASFV uncharacterized proteome consisting of 51 proteins was used to conduct an initial search against the NCBI database using the PSI-BLAST with default parameters (Altschul *et al*., 1997). Because PSI-BLAST was not able to identify experimentally solved template sequences, HHpred, a more sensitive profile sequence search algorithm (Alva *et al*., 2016; Söding *et al*., 2005) was used to search homologous protein families against Pfam 04 jul16 (Finn *et al*., 2016), SCOPe 2.06 (Fox *et al*., 2014), and PDB v 3.3 (Berman *et al*., 2000) databases. The recommended iterations were three from maximum multiple sequence alignment generation used to search homologs three times, and a cut-off e-value of 0.0001 (Söding *et al*., 2005).

3.2.2 Structural Modeling and Structural Classification of pB263R

LOMETS was used to identify protein templates. The ten best protein templates of pB263R (for both full-length and truncated sequences) were used to model the 3D structure in I-TASSER (Roy *et al*., 2010; Yang & Zhang, 2015). For the full-length sequence, the topology of pB263R was also predicted using RaptorX (Källberg *et al*., 2012). The pB263R model 1a analogs were searched using the TM-align structural alignment tool (Roy *et al*., 2012; Zhang Yang & Skolnick, 2005b) and the best templates were visualized with the ICM browser version 3.8.5 (Raush *et al*., 2009). The fold family classification of the pB263R model 1a was determined using the TM fold tool (Jinrui & Yang, 2010). The TM fold was installed in Fedora 17 and a Perl script (Appendix II) was

used to calculate p-values using the extreme value distribution cumulative distribution function and posterior probabilities for proteins assigned a particular TM score using the formula (Jinrui & Yang, 2010);

$$P\text{-value}(x) = \int_x^1 f(x|\mu,\sigma)\mathrm{d}x = 1 - \exp\left[-\exp\left(\frac{\mu - x}{\sigma}\right)\right]$$

The Perl line code implemented was;

$p-value = 1-exp (- exp (($μ-$tms)/$sigma)),

Where the parameter definitions were $μ = 0.1512, $sigma = 0.0242 and $p-value = $printf ("%.15f", $p-value).The posterior probabilities of the SCOP/CATH family were calculated using a sigmoidal Boltzmann function using the formula (Jinrui & Yang, 2010);

$$y = \frac{A_1 - A_2}{1 + e^{(x - x_0)/dx}} + A_2$$

The accompanying code and its parameters were implemented in a Perl script to calculate the posterior probabilities using a sigmoidal Boltzmann function (Y):

$a1_scop (min) = -0.0071735; $a2_scop (max) = 0.99803;

(Tm_scop $x0_scop) = 0.57917 $ δx _scop = 0.048934;

$a1_cath (min) = -0.004103; $a2_cath = 0.99217(max);

(Tm_cath$x0_cath) = 0.52695; $ δx _cath = 0.04688;

Y= max+ (min-max)/ (1+exp ((TM_n – Tm_scop)/ $ δx _scop)).

Y= max+ (min-max)/ (1+exp ((TM_n – Tm_cath)/ $ δx _cath)).

Here, min is the initial value (right horizontal asymptote), max is the final value (left horizontal asymptote), Tm_scop and Tm_cath are the center points of inflection, and δx is the change in X corresponding to a significant change in Y.

The stereochemical qualities of the predicted structure were assessed using protein structure analysis (ProSA) (Wiederstein & Sippl, 2007) and Rampage (Lovell *et al*., 2003) algorithms. And hydropathy profiles were plotted using Expasy (Gasteiger *et al*., 2003).

3.2.3 Functional and Ligand Binding Site Annotation of pB263R model 1a

The TM-site algorithm was used to compare the structure of the sub-sequence from the first to last binding residues of the queried 3D structure against the BioLip database (Jianyi *et al*., 2013b, 2013a). TM-site also identified potential ligand-binding site residues in the predicted pB263R model 1a structure. In the TM-site predictions, the binding pockets were selected and ranked based on the confidence score (CS_t). A CS_t prediction of 0.35 has an average false-positive rate below 0.16 and false-negative rates below 0.13 (Jianyi *et al*., 2013a). The ligand binding site residues of the pB263R were independently predicted using the RaptorX binding site prediction server (Källberg *et al*., 2012). Further, a search was performed to identify the 10 top-ranked structural analogs of pPB263R model 1a using the RaptorX structure alignment server (http://raptorx.uchicago.edu/DeepAlign/submit/) the results of the structural alignment were mapped back to a multiple sequence alignment and observed using JalView v.2.10.1 (Waterhouse *et al*., 2009) to identify conserved residues. The pB263R model 1a functional annotation was deduced with GO terms (Huntley *et al*., 2015).

3.2.4 Phylogenetic Classification of pB263R

Various protein sequences used for phylogenetic classification of TBPs and TBP-like factors were searched for by HHblits (Appendix III), pB263R, and TBPs from representative members of the NCLDVs were obtained from the UNIPROT database. The protein sequences were aligned using Multiple Sequence Comparison by Log-Expectation (MUSCLE) (Edgar, 2004) in Molecular Evolutionary Genetic Analysis tool version 6 (MEGA6) (Tamura *et al*., 2013). A pairwise genetic distance matrix and a phylogenetic tree were generated using the maximum likelihood method. Phylogeny bootstrap test was carried out based on a preliminary neighbor-joining inferred tree and a total of 1,000 bootstrap replicates were used to generate a final consensus tree. A summary of all prediction procedures for protein structure is shown in Figure 3.1.

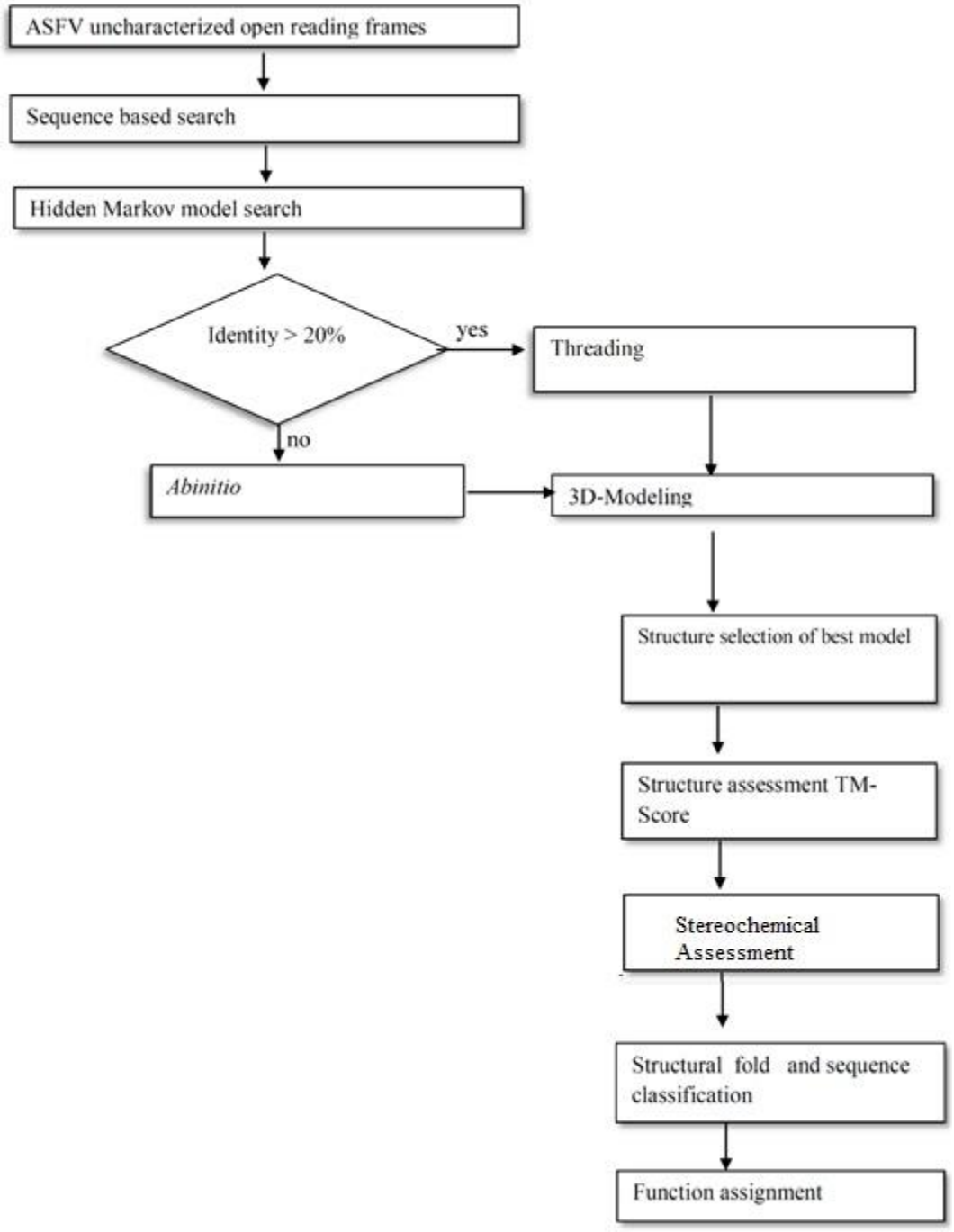


Figure 3.1 Flow chart showing protein structure prediction process.

A protein structure prediction flowchart showing processes used in sequence structure to function assignment of pB263R

3.3 RESULTS

3.3.1 Sequence-Based Analysis Reveals that pB263R is Homologous to a TATA Binding Protein

In the present study, the existence of an ORF that has features that can bind to promoter TATA-like motifs was evaluated. No functional homologs were identified upon querying ORFs of uncharacterized sequences of the proteome (https://www.uniprot.org/proteomes/UP000000624) on the NCBI non-redundant protein database using BLAST and PSI-BLAST protein sequence search method. A majority of the sequences identified were similar to putative proteins conserved among ASFV, faustovirus, kaumoebavirus, and pacmanvirus. Further analysis of the uncharacterized sequences was performed on a more sensitive sequence-based search method to detect remote homologs. The results indicated that from the entire uncharacterized proteome, only pB263R protein had experimentally solved functional homologs that showed low sequence identity ranging from 12%–15%. These homologs were identified as TATA-binding domains of *Methanocaldococcus jannaschii*, *Pyrococcus woesei*, *Saccharomyces cerevisiae*, *Sulfolobus acidocaldarius*, and *Encephalitozoon cuniculi*. The cut-off level for the e-values was less than 0.0001 with a 100% probability of being correctly identified (Table 3.1). The HMM queried regions showed that there was no good sequence similarity approximately between amino acid residues (1 – 44).

Table 3.1 **pB263R homologs identified by HHpred**

Rank	PDB ID	Organism	Protein domain	Probab.[1]	e-value	Identity[2]	Query HMM region	Columns [3]
1	2z8uA	*M. jannaschii*	TBP	100%	7.6×10^{-39}	15%	44–253	176
2	1aisA	*P. woesei*	TBP	100%	1.3×10^{-38}	15%	40–251	178
3	1ytbA	*S. cerevisiae*	TBP	100%	2.1×10^{-38}	15%	44–253	175
4	1mp9A	*S.acidocaldarius*	TBP	100%	3.7×10^{-38}	12%	44–263	185
5	3eikA	*E. cuniculi*	TBP	100%	1.0×10^{-37}	13%	34–253	185

Note: PDB ID: protein data bank identification; TBP: TATA-binding protein; HMM: hidden Markov model; [1]probability of the template is a true positive; [2] percent sequence identity between the template and query sequences; [3] number of aligned HMM-HMM match-match columns.

3.3.2 Local Meta-Threading Server (LOMETS) Identified Good Templates for 3D Modeling

The quality of template alignments was assessed based on a normalized Z score by LOMETS independently. About 6 out of 10 threading templates had a normalized Z score of >1 all of which turned out to be TBPs (Table 3.2). A normalized Z score of > 1 in an experimentally solved protein suggests highly accurate primary and secondary alignment (Roy *et al*., 2010). Based on results from different organisms, other structures with good alignment coverage but not TBPs were identified and used in modeling to highlight the independence of I-TASSER (Table 3.2).

Table 3.2 Top ten analogs identified by LOMETS

Rank	Protein Domain	Organism	PDB ID	Iden1[1]	Iden2[2]	Cov[3]	Norm.Z-score[4]
1	TBP	*Encephalitozoon cuniculi*	3eikA	14	17	0.67	4.01
2	TBP	*Sulfobolus Acidocaldarius*	1mp9A	24	17	0.67	1.11
3	TBP	*S. cerevisiae*	4b0Aa	16	20	0.74	3.98
4	TBP	*Methanocaldococcus jannaschii*	2z8Ua	16	13	0.63	3.87
5	TBP	*Pyrococcus Woisei*	1aisA	16	15	0.68	3.87
6	TBP	*Sulfobolus Acidocaldarius*	1mp9A	12	17	0.73	3.64
7	CPSF-100	*Saccharomyces cerevisiae*	2i7Xa	15	16	0.90	0.53
8	TBP	*Homo sapiens*	1jfiA	13	17	0.65	0.67
9	GS	*Chromohalobacter Salexigens*	5dm3A	9	15	0.88	0.61
10	TBP	*Saccharomyces cerevisiae*	1ytbA	27	17	0.67	0.62

Note:[1]Iden1 is the percentage sequence identity of the templates in the threading aligned region with the query sequence; [2]Iden2 is the percentage sequence identity of whole template chains with the query sequence; [3]Cov. Represents the coverage of the threading alignment and is equal to the number of aligned residues divided by the length of the query protein; [4]Norm. Z-score is the normalized Z-score of the threading alignments; alignments with a normalized Z-score > 1 mean a good alignment and vice versa.

3.3.3 Predicted pB236R Structure is a TBP-like Domain

Following the excision of the threaded templates and iterative simulations by I-TASSER, the first model (pB263R model 1a) out of the five models predicted as a full-length simulation from the pB263R sequence was obtained as a three-dimensional coordinate file. The predicted pB263R model 1a structure, revealed a typical pseudo-dyad symmetric alpha/beta (α/β) saddle-like structure respectively, with bifurcating stirrups. Each subdomain of the C-terminal domain (CTD) was populated with α helices (H1'/H2' and H1/H2) and β strands (S1'-S5'and S1-S5) (Figure 3.2).

The β strands formed the curved underside of the structure for binding to the minor groove of probably the TATA element in the DNA sequence.

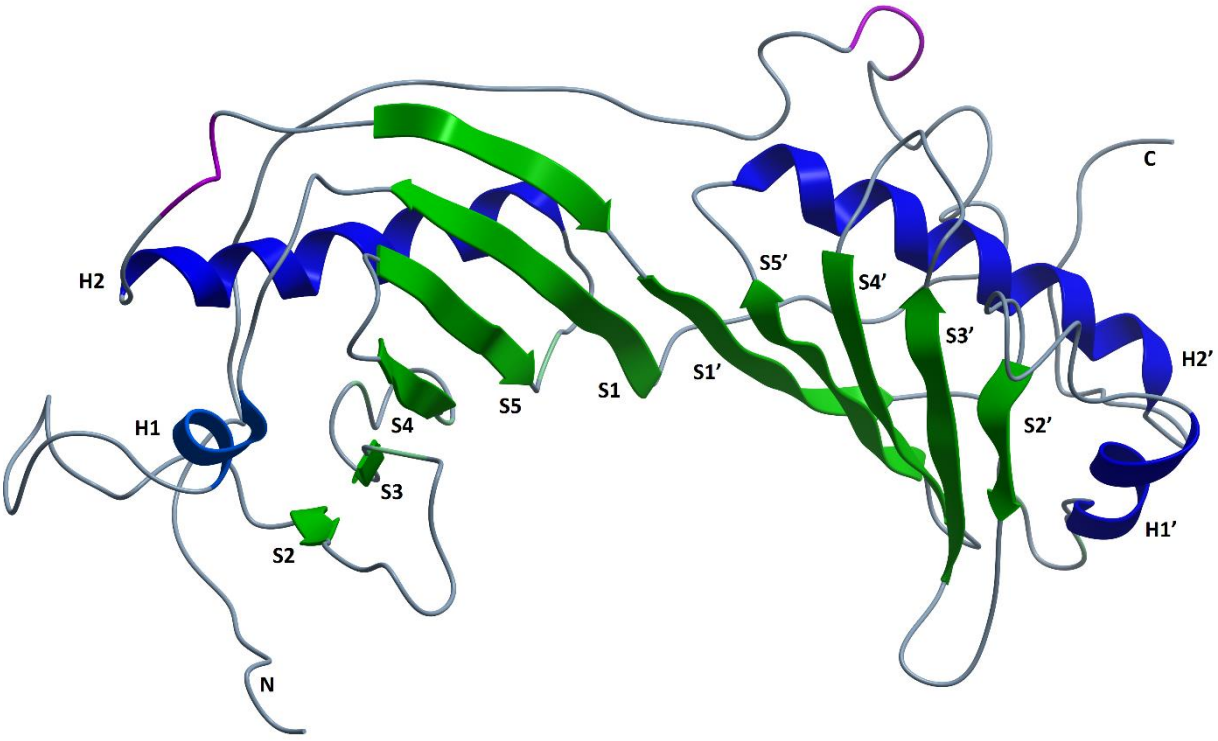


Figure 3.2 I-TASSER modeled topology of pB263R protein sequence

A ribbon-style image of pB263R model 1a, with a TM score of 0.52 and Cs of −1.61. The β strands in green and α helices in the blue form a pseudo-dyad structure comprising two similar subdomains, each with two α helices (H1/H2 and H1'/H2') and five β strands (S1–S5 on subdomain 1 and S1'–S5'on subdomain 2). The C and N indicate the C and N termini respectively.

Moreover, the protein prediction quality was defined by two parameters the template modeling score (TM score) and confidence score (Cs). The TM-score indicates the level of protein similarity, where a TM-score of < 0.17 indicates a randomly selected protein pair while a TM score > 0.5 demonstrates a significant fold. TM score ranges between zero and one. In the present study, the predicted pB263R model 1a had a TM-score of 0.52. This implied that the obtained TM score was statistically significant and had a correct fold topology. However, the Cs of the pB263R model 1a was −1.61 slightly below the −1.5 threshold. The low Cs was due to the mismatching regions in the N-terminal opening loop from the sequence region between 1–44 and not the core region. A significant improvement in TM-score from 0.52 to 0.57 and Cs from −1.61 to −1.15 were obtained from modeling pB263R sequence residues from 45 to 263 to emphasize the structural integrity (Figure 3.3).

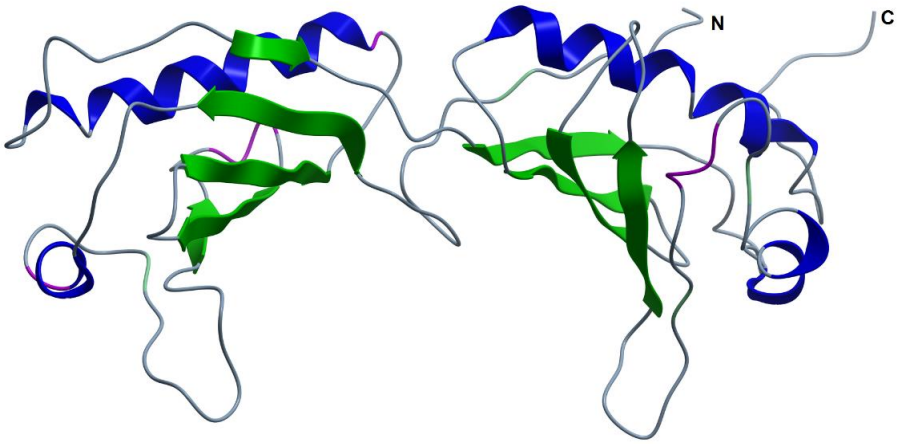


Figure 3.3 Model of pB263R from residues 45 (A1) to 263 (D219) predicted by I-TASSER

The structure had a typical TBP topology but with an asymmetric number of anti-parallel β sheets. The structure had a high TM score of 0.57 and Cs −1.15.

RaptorX prediction of the full-length pB263R sequence again indicated the presence of a long N-terminal coil with amino residues at positions 1–44 (Figure 3.4).

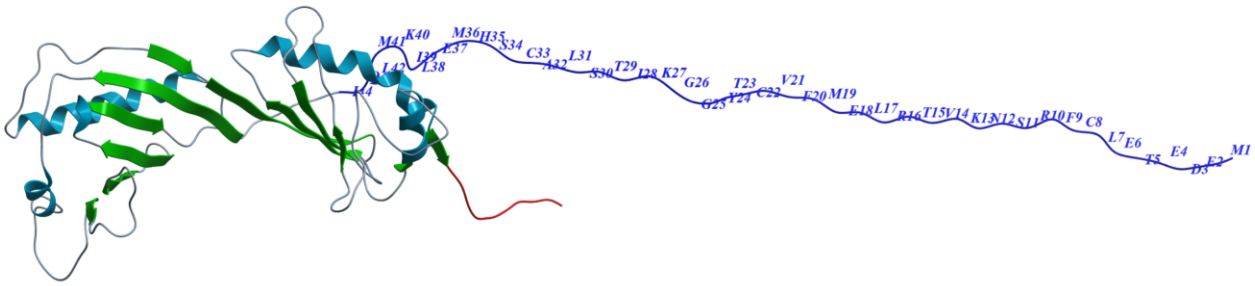


Figure 3.4 A pB263R model independently predicted by RaptorX indicates a long-disordered coil spanning residue 1 (Met) to Ala-44 (dark blue region)

The remaining residues fold into the two symmetric domains adopting an analogous fold structure similar to typical TBPs; light blue indicates alpha-helices, light green indicates beta-sheets, and grey indicates coiled regions.

3.3.4 Fold Family Posterior Probability Classification Revealed that pB263R Model 1a has a Similar Topology to 4b0aA, a TBP

The posterior probability is the probability that two protein structures have the same fold family assignment in either SCOP or CATH structural classification databases for a given TM score. For pB263R model 1a the best 8 out of 10 experimentally solved structures identified by TM align were TBPs (Table 3.3) and had posterior probabilities greater than 97.5% in SCOP and 90% in CATH respectively of sharing a fold with pB263R model 1a (Table 3.3).

The 4b0aA was ranked as the highest structural analog (TM-score of 0.711). The posterior probabilities of fold family sharing with pB263R model 1a were 98.32% in CATH and 98.81% in

SCOP databases respectively. These findings suggest that 4b0aA and pB263R model 1a belonged to the same fold family. However, 4b0aA had not yet been classified in SCOPe 2.06 (Table 3.3). Therefore, fold family classification was considered for the second-ranked structure 1vokA, a TBP from *Arabidopsis thaliana* (Nikolov & Burley, 1994). The posterior probabilities of pB263R model 1a sharing the same fold family with 1vokA were 99.18% in SCOP and 99.66% in CATH respectively. From this finding, the two structures can be inferred to belong to the same fold families in CATH and SCOPe 2.06, the CATH code is 3.30.310.10 for class, alpha-beta, architecture, two-layer sandwich, topology, TBP, homologous superfamily, TATA-binding protein. While in SCOPe 2.06, 1vokA had the classification of d.129.1.1 for class d: α and β proteins, fold d.129: TBP-like, superfamily d.129.1, and TBP-like, family d.129.1.1 and TBP like, C-terminal domain. (Table 3.3). These results indicate that pB263R model 1a can be structurally classified as a TBP. Additionally, the posterior probabilities of 4b0aA and 1vokA structures sharing the same fold family were 99.18% in SCOP and 99.66% in CATH based on a non-manual classification in TM-fold that gave rise to a TM-score of 0.9142. This is an indication that they also belonged to the same fold family.

Table 3.3 TM-fold structural classification of pB263R model 1a

Rank[1]	PDB hit	SCOPe 2.06 classif.	TM score	RMSD[2]	IDEN[3]	Coverage[4]	Posterior probab. in SCOP[5]	Posterior probab. in CATH[6]	p-values
1	4b0aA	Not manually classified in SCOPe 2.06	0.711	1.49	0.162	0.715	0.9881	0.9832	4.2×10^{-13}
2	1vokA	d.129.1.1	0.670	1.64	0.135	0.696	0.9918	0.9966	3.4×10^{-14}
3	1jfiC	d.129.1.1	0.664	1.45	0.115	0.684	0.9958	0.9954	3.0×10^{-14}
4	1mp9A	d.129.1.1	0.657	2.09	0.125	0.677	0.9958	0.9910	9.1×10^{-14}
5	1nh2A	d.129.1.1	0.656	1.31	0.167	0.684	0.9920	0.9914	6×10^{-14}
6	3eikA	d.129.1.0	0.645	1.43	0.146	0.688	0.9920	0.9369	8×10^{-15}
7	1pczA	d.129.1.1	0.643	1.88	0.158	0.665	0.9918	0.9964	3.4×10^{-14}
8	2z8uB	d.129.1.1	0.618	1.68	0.161	0.677	0.9919	0.9098	2.4×10^{-14}
9	2glsL	d.15.9.1	0.486	5.48	0.061	0.681	0.2902	0.1236	9.3×10^{-7}
10	3ng0A	d.15.9.0	0.481	5.66	0.057	0.707	0.2736	0.1150	1.1×10^{-6}

Note: CATH: class, architecture, topology, and homology; PDB: protein data bank; SCOPe 2.06: structural classification of proteins;[1]rank of PDB structures based on TM score of the structural alignment between the pB263R model 1a and known structures in the PDB library; [2]root mean square deviation between residues structurally aligned by TM-align; [3]percent sequence invariability in structurally identical regions; [4]alignment coverage by TM-align, equivalent to the number of structurally aligned residues divided by the length of the queried protein sequence; [5]probability of two proteins structures being in the same fold family in SCOPe; [6]probability of two proteins structures being in the same fold family in CATH.

3.3.5 The pB263R Model 1a is Stereochemically Fit

The Z score predicted of the pB263R model 1a by the ProSA-webserver (Wiederstein & Sippl, 2007)was −4.56. This was within the recommended range for protein structures of the same size (Figure 3.5A). From the Ramachandran plot obtained from rampage (Ramachandran *et al*., 1963) the predicted model (pB263R Model 1a) had 85.8% in the favored region and 9.6% in the allowed regions. Thus, showing that the model was fit stereochemically (Figure 3.5B).

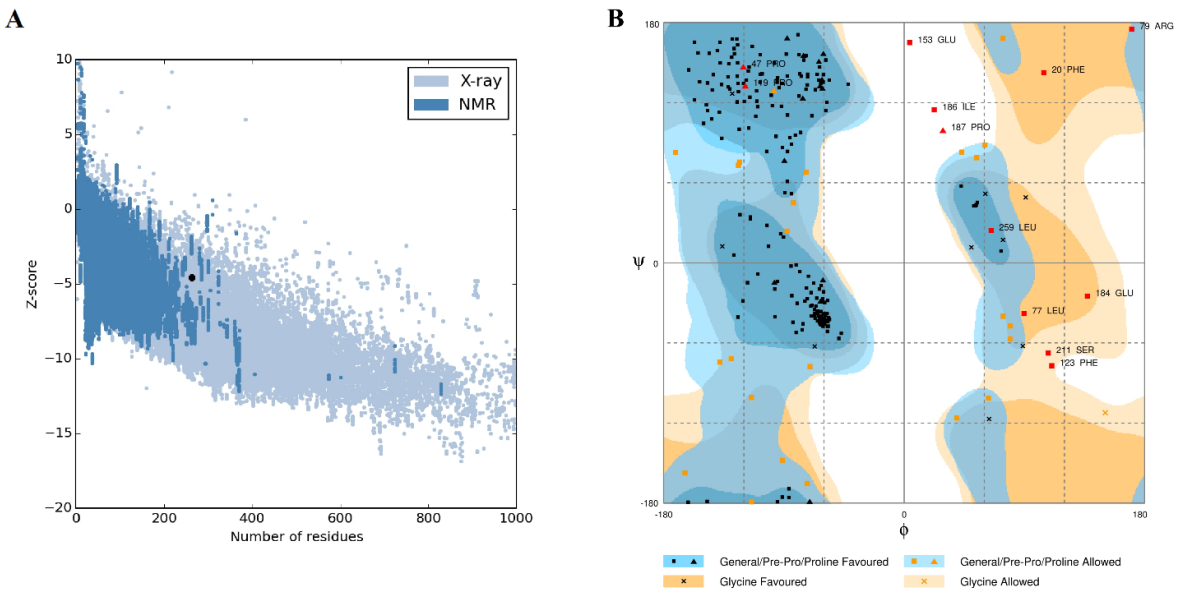


Figure 3.5 pB263R model 1a *in silico* predicted structure validation

The pB263R model 1a was validated using the ProSA-web server and rampage. (A) ProSA-web plot showing Z-score of the predicted structure is −4.65 (black dot) relative to Z-scores of similar-sized protein structures, solved using NMR and X-ray crystallography. (B) Ramachandran plot result from rampage showed 85.8% of residues lie in the most preferred region, 9.6% of residues in allowed regions, and 4.6% in the outlier region.

3.3.6 Prediction of pB263R Model 1a Biological Function and Ligand Binding Site

The default TM-site confidence score (CS_t) for a good ligand-binding site prediction is 0.35. Two TBP isoform structures of *Arabidopsis thaliana* (1qn4A and 1qnbA) were identified to have significantly higher CS_t of 0.39 and 0.37 for 1qn4A and 1qnbA respectively (Table 3.4) (Patikoglou *et al*., 1999). These results infer that the two TBPs share binding sites with pB263R model 1a. The ligand binding site residues for both the x-ray solved structures of (1qn4A and 1qnbA) were transferred to pB263R model 1a as putative binding site residues (Table 3.4).

Table 3.4 The top five binding site pockets and residues for pB263R Model 1a predicted by TM-site

Rank	CSt[1]	Cluster size[2]	Representative template[3]	TATA box sequence	Ligands[4]	Predicted binding site residues in UP-B263R model 1a
1	0.39	32	1qn4A_BS01_NUC	TGCC[CATTTATA]GC (TATAAATG)	Nucleic acid (32)	N52, N54, F90, N91, K108, F110, T113, E115, Q117, I157, Q158, N160, E200, D201, S205, F207, R217, N219, L229, and N232
2	0.37	27	1qnbA_BS02_NUC	TGCC[CATTTATA]GC (TATAAATG)	Nucleic acid (27)	K49, A50, N52, C89, F90, E95, I98, M104, K106, K108, P119, G120, N160, D201, S202, N219, F221, K223, K225, and N227
3	0.19	3	5fmfQ_BS02_NUC	C[CTTTTATA]G (TATAAAAG)	Nucleic acid (3)	K49, A50, T88, C89, F90, N91, A93, E95, S97, M104, K106, K108, F110, P119, G120, I122, D201, S202, F221, K223, and K225
4	0.16	1	5fz5O_BS01_NUC	C[CTTTTATA]G (TATAAAAG)	Nucleic acid (1)	N52, N54, L60, F90, F110, S112, E115, Q117, Q158, D201, and L229
5	0.16	1	5fz5O_BS02_NUC	T[ATTATATA]CA (TATATAAT)	Nucleic acid (1)	K49, N50, N52, C89, F90, E95, L99, K106, K108, P119, and S202

Note: [1]Confidence score of the TM-site prediction (range: 0–1); [2] total number of templates in a cluster; [3] single complex structures with the most representative ligand in the cluster; [4] names of the identified ligand.

During the prediction of the ligand binding site, phenylalanine residues (F90, F110, and F221) that are important in distorting stacking interactions at dinucleotide step sites and adding stability in kinking were detected as conserved residues on the curved underside of the C terminal domain (CTD) in the pB263R model 1a structure (Figure 3.6). The positions of the three phenylalanine residues matched those responsible for kinking the TATA element in 1qn4A (Patikoglou *et al*., 1999).

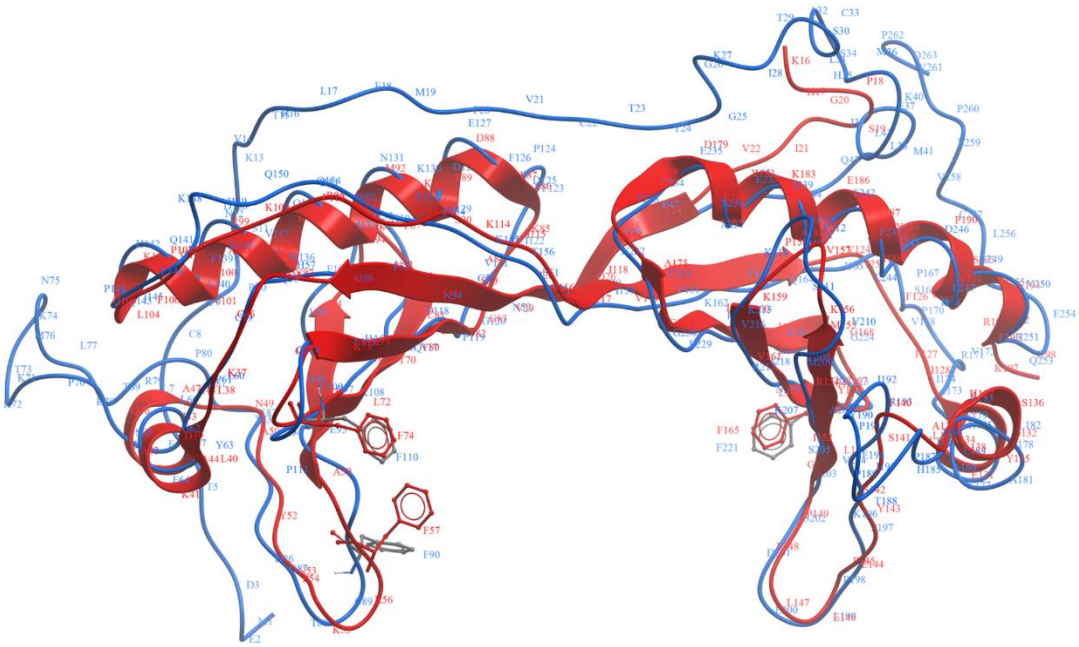


Figure 3.6 pB263R model 1a (blue) and experimentally solved 1qn4A (red) structural superimposition

The predicted conserved residues include intercalating phenylalanine (F90, F110, and F221) of pB263R model 1a on the underside of CTDs aligned to F57, F74, and F165 of 1qn4A.

Independently, eight pockets in pB263R were predicted by RaptorX binding site recognition server (http://raptorx.uchicago.edu/documentation/#goto2) and were shown to exceed a threshold multiplicity value of 40 (Table 3.5). The bases identified at the binding site include TTAATAAG, which is composed of adenine and thymine bases with guanine in the eighth pocket. The presence

of A/T bases implied the presence of a somewhat functional TATA-like element, suggesting the possibility that pB263R binds to TATA-like elements in ASFV.

Table 3.5 RaptorX binding site server predicted binding site residues and pockets.

Pocket	Pocket multiplicity[1]	Ligand[2]	Binding residue
1	146	DT	D201, K203, S205, N219, and L229
2	141	DT	C89, F90, S97, K106, K108, and Q117
3	110	DA	N52, N54, F110, E115, Q117, and I157
4	104	DA	N52, I157, M158, N160, R217, L229, and G230
5	96	DT	K49, A50, N52, L99, K106, P119, G120, and N160
6	94	DA	K49, N160, K162, F221, K225, and N227
7	94	DA	D201, S202, F221, K223, and K225
8	82	DG	F90, F110, T113, and E115

Note: [1]Frequency with which the predicted pocket was found in the query protein template structure; [2]DA: 2'-deoxyadenosine-5'-monophosphate, DG: 2'-deoxyguanosine-5'-monophosphate, DT: 2'-deoxythymidine-5'-monophosphate.

The conserved phenylalanine residues were also observed from multiple sequence alignment of TBP sequences. Additionally, the presence of conserved glycine residues in pB263R protein (Gly114, Gly120, Gly224, and Gly230) was consistent in most eukaryotic TBPs (Figure 3.7).

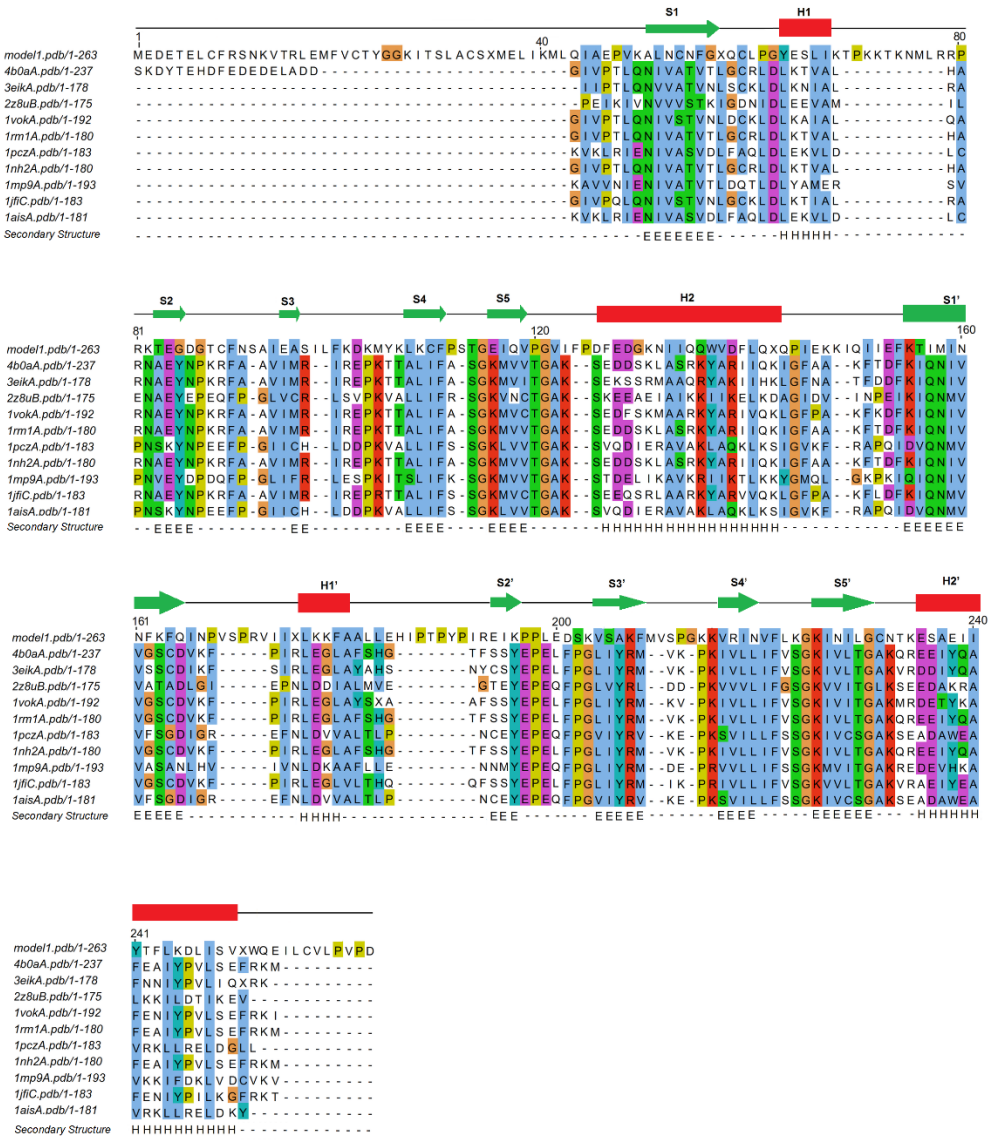


Figure 3.7 Multiple sequence alignment of TBP sequences of the top ten closest TBP structural homologs of pB263R

4b0aA, *S. cerevisiae* TBP transcription initiation factor II D subunit 1; 3eikA, E. *cuniculi* TBP; 2z8uB, *M. jannaschii* TBP; 1vokA, *A. thaliana* TBP; 1rm1A, *S. cerevisiae* TBP; 1pczA, *P. woesei* TBP; 1nh2A, *S. cerevisiae* TBP CTD; 1mp9A, *S. acidocaldarius* TBP; 1jfiC, *Homo sapiens* TBP; 1aisA, *P. woesei* TBP.

These conserved glycine residues are vital in forming a 3D structure that accommodates a short turn between crucial β strands (S4, S4' and S5, S5') (Figure 3.8 and Figure 3.2). The Gly120 and Gly230 have also been shown to make water-mediated contacts via carbonyl oxygen atoms (Tsai & Sigler, 2000). Appropriately positioned proline residues (P111, P119, P124, P144, P167, P170, P189, P191, P198, and P212) were observed in tight turns of α helices and β strands within the predicted TBP structure, and most valine residues comprising the hydrophobic core were replaced with asparagine (N) residues that were presumably involved in hydrogen bond formation through A and T bases (V → N227, L → N219, V → N160, V → N52 (S/T) → N54, and V → Q117).

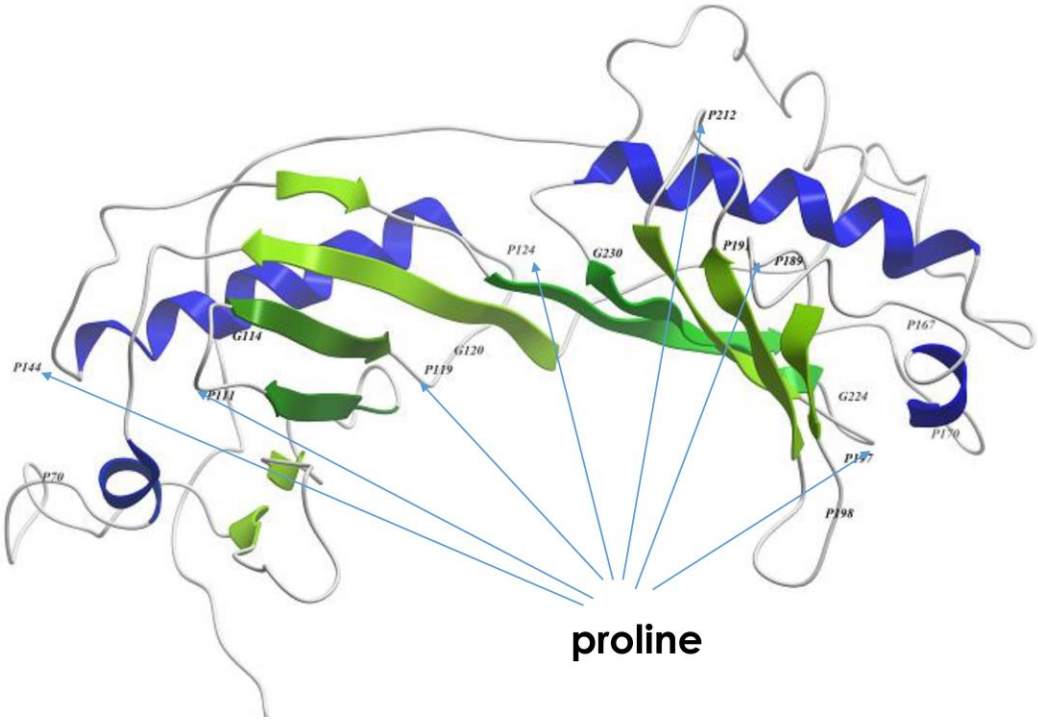


Figure 3.8 Conservation of proline and glycine residues

Conservation of pB263R glycine and proline in the turns of α helices and β strands within the predicted TBP topology.

The highest predicted molecular function for pB263R based on gene ontology terms was a DNA-binding protein (GO: 003677; GO score=0.78). The highest biological processes predicted were, transcription initiation from RNA polymerase II promoter (GO: 0003677; GO score=0.78) and any process regulating transcription that is DNA dependent (GO: 0006355; GO score=0.78). The highest cellular components predicted were a DNA-directed RNA polymerase II holoenzyme (GO: 0016591, GO score=0.52) and RNA polymerase I transcription factor complex (GO: 000120, GO score=0.52) (Table 3.6). These results suggest that pB236R binds to promoters and modulates the RNA polymerase II enzyme involved in transcription.

Table 3.6 Consensus prediction of GO terms.

Molecular function	GO-score	Biological process	GO-Score	Cellular component	GO-score
GO:0003677-DNA binding	0.78	GO:0003677-Transcription Initiation from RNA Pol II promoter	0.78	GO:001591 DNA-directed RNA Pol II holoenzyme	0.52
GO:0008135-Translation –factor activity/Nucleic acid binding	0.52	GO: 0006355-Regulation of transcription DNA dependent	0.78	GO: 000120 RNA Pol I transcription factor complex	0.52
GO: 0001071-Nucleic acid binding activity/Transcription activity	0.52	GO: 0032196 Transposition	0.52		
GO: 0003702 RNA pol II transcription activity	0.45	GO: 0006384 Transcription initiation from RNA pol III promoter	0.52		
GO:0005515 Protein binding	0.45	GO:0070897 DNA-Dependent transcriptional PIC assembly	0.52		
		GO:0044267 Cellular protein metabolic processes	0.47		

3.3.7 The pB263R is Evolutionarily Related to TBPs of NCLDVs

The phylogenetic classification was used to infer the evolutionary relationship between the BA71V pB263R sequence and other putative TBPs of NCLDV members. The NCLDVs members used in this analysis include representative members of the superfamily Poxiridae, Mimiviridae, Iridoviridae, Phycodnaviridae, Ascoviridae, and Marseillevirus retrieved from the UNIPROT database. TBPs from other prokaryotes organisms and archaea were also included in the analysis. The maximum likelihood bootstrapped phylogenetic tree showed moderate similarity between the BA71V pB263R sequence and TBPs of NCLDVs forming a single clade (Figure 3.9). The close evolutionary relationship of pB263R with other TBP homologs in NCLDVs implies that it is a potential TBP.

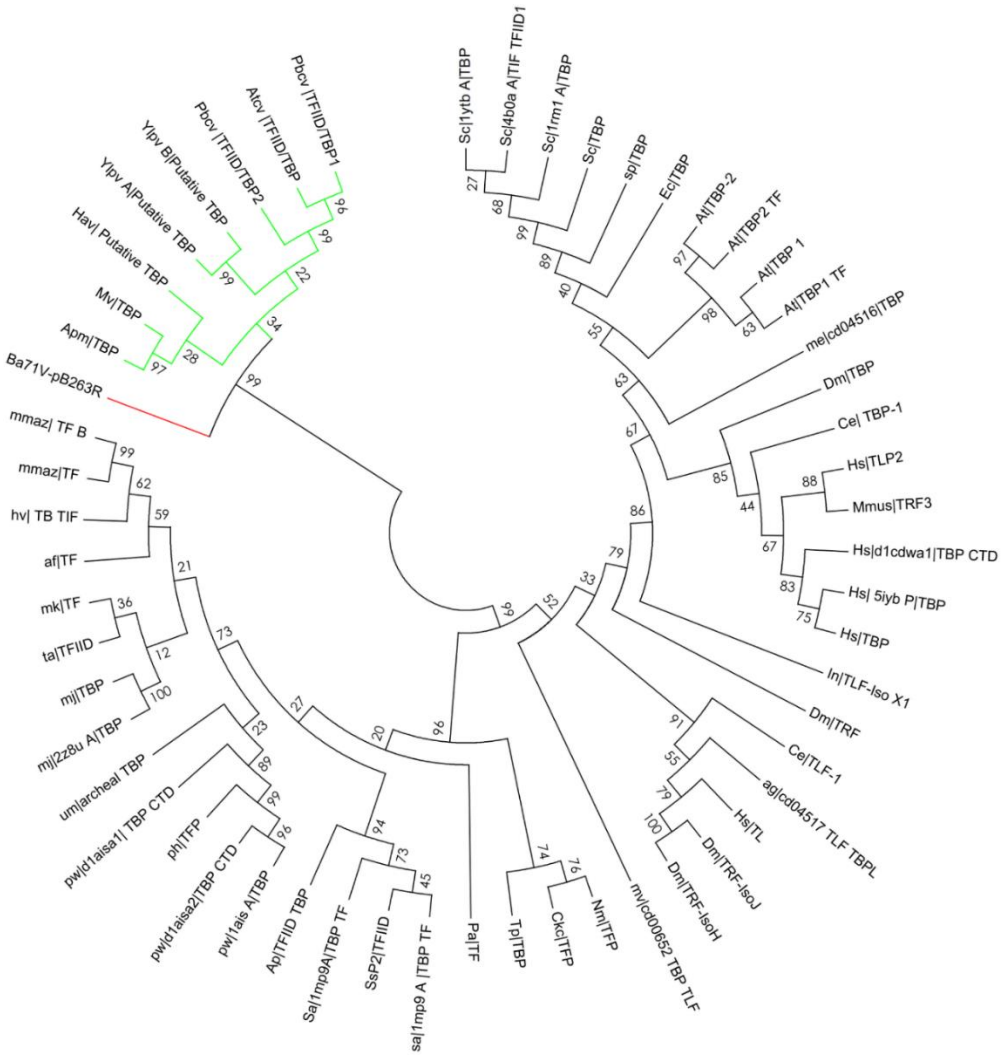


Figure 3.9 Bootstrapped phylogenetic tree classifying pB263R

Light green designates NCLDV TBPs, and red designates the BA71V pB263R sequence. Nodes designate maximum likelihood bootstrap percentages. The NCLDVs TBPs (light green) and pB263R (red) indicate close evolutionary relations. Apm, *A. polyphaga* mimivirus; af, *Archaeoglobusfulgidus*; ag, *Anoplophoraglabripennis*; Ap, *Aeropyrumpernix* K1; at, *A. thaliana*; atcv, *A. turfacea* Chlorella virus; ce, *Caenorhabditis elegans*; dm, *Drosophila melanogaster*; ec, *E. cuniculi*; hav, Heterosigmaakashiwo virus; hs, *H. sapiens*; hv, *Haloferaxvolcanii*; In, *Ipomoea nil*; me, *Mortierellaelongata*; mj, *M. jannaschii*; mk, *Methanopyruskandleri*; mmaz, *Methanosarcinamazei* Go1; mmus, *Mus musculus*; mv, Megavirus Courdo 7; mv, *Mortierellaverticillata*; nm, *Nitrosopumilusmaritimus*; pbcv, *P. bursaria* Chlorella virus; pw, *P. woesei*; sa, *S. acidocaldarius*; sc, *S. cerevisiae*; sp, *Schizosaccharomyces pombe*; SsP2, *S. solfataricus* P2; Tp, *Thermofilum pendens*; pB263R, protein B263R; Ylpv A, Yellowstone Lake phycodnavirus 1; Ylpv B, Yellowstone Lake phycodnavirus 3.

3.4 DISCUSSION

The current study was aimed at identifying a protein with features of binding to TATA-like sequences. Of all the uncharacterized proteins, only pB236R showed sequence similarity to TATA-binding proteins of various organisms in archaea groups. The identified templates used in modeling the 3D structure of pB263R had good alignment and thus the modeling accuracy was expected to be high (Roy *et al*., 2010). Moreover, the topology of pB263R model 1a had a TM-score that was higher than 0.5. Also, the pB263R model 1a structure had typical features of a TBP such as a dual-symmetric alpha/beta (α/β) structure and branching stirrups that were formed on the concave underside of the structure for binding to the minor groove of the TATA-like elements of the DNA sequence (Juo *et al*., 1996; Patikoglou *et al*., 1999). The pB263R model 1a also had a TBP-like fold topology in the core region from amino acid residues 45-263. Furthermore, based on posterior probability fold family classification, the modeled structure had a similar fold to the TBP fold family.

The TM site prediction algorithm identified two (1qn4A and 1qnbA) TBP isoforms in *Arabidopsis thaliana* that shared similar binding site residues with pB263R model 1a. The ligands identified for pB263R Model 1a were nucleic acids having the TATAAATG TATA-box motif (Table 3.4). Additionally, three conserved phenylalanine residues F90, F110, and F221 were aligned to F57, F74, and F165 of 1qn4A. These phenylalanine residues were positioned on the curved underside of the C-terminal domain of pB263R model 1a from structural superposition. Insertion of these phenylalanine residues between two T/A pairs produces two sharp kinks in the helix, unstacking and pulling the T/A bases apart. Any loss of stacking energy is thus compensated by extensive van

der Waals interaction or pi-pi interaction between the A/T and the phenylalanine rings (Mondal *et al*., 2014). Other studies have also reported the occurrence of these phenylalanine residues and their association with TATA-like elements in inducing transcriptional activation (Juo *et al*., 1996; Patikoglou *et al*., 1999). Additionally, positively charged arginine and lysine residues such as K49, K106, K108, K203, K215, K162, R217, K223, and K225 that are possibly involved in hydrogen bonding to the negatively charged phosphate backbone of DNA were observed from the putative transfer of residues from TM-site and RaptorX binding site recognition servers (Tables 3.4 and 3.5) (Mondal *et al*., 2014). Proline residues (P111, P119, P124, P144, P167, P170, P189, P191, P198, and P212) were observed in tight turns of α helices and β strands within the predicted TBP structure further adding proof to the secondary structure integrity of the TBP (Krieger *et al*., 2005).

The study, however, did not demonstrate the exact mechanisms of how to pB263R model 1a could produce a bend or a kink in DNA. Nonetheless, it has been shown that TBPs with three phenylalanine residues exist, such as those of *C. parvum, T. gondii, and P. marinus* and they typically lack TBP-associated factors like Mot1 regulators of TBP function (Koster *et al*., 2015). Additionally, *Plasmodium falciparum* TBPs (PfTBP) containing three conserved phenylalanine residues can recognize TATAA motifs in the promoter region of the knob associated histidine rich protein promoter and close variations of TATA-like motifs in promoters such as TGTAA located in the glycophorin-binding protein promoter region (Ruvalcaba-Salazar *et al*., 2005). The *Plasmodium falciparum* intergenic region is particularly rich in A+T-rich sequences (>90%) (McAndrew *et al*., 1993), and regions resembling classical TATA motifs (consensus TATAAA/TGTAA) are abundant upstream of most genes (McAndrew *et al*., 1993; Ruvalcaba-Salazar et al., 2005). Therefore, it can be inferred that TBPs with three DNA kinking phenylalanine residues may show diversity in identifying TATA-like variations. From the pB263R model 1a

function prediction using the TM site algorithm, the TATAAATG motif is similar to the TATA box sequences (Juo *et al*., 1996; Patikoglou *et al*., 1999). It is therefore highly likely that pB263R model 1a binds to late promoter TATA-like motifs. Previous studies have detected TATA-like motifs in the adjoining sequences of late ASFV genes (García-Escudero & Viñuela, 2000; Rodríguez & Salas, 2013). Again, conserved motifs essential for binding to TATA-like elements have been verified in the *P. falciparum* (TATAA) region located 81 base pairs upstream of the transcription start site of the promoter region (Ruvalcaba-Salazar *et al*., 2005), while for *A. thaliana* the sequence (TATAAATG) has been identified (Juo *et al*., 1996). ASFV late promoters have the sequence TATATA, all these organisms share similar TATA motifs of the first four bases that are known to be critical for TATA binding (Juo *et al*., 1996; Patikoglou *et al*., 1999). The RaptorX binding site server also predicted a somewhat similar TATA-like motif. This strongly suggests that the predicted pB263R model 1a binds to these conserved TATA-like sequences in ASFV. Besides, 'TATA' sequence replacement in late genes with 'GCGC' is deleterious to transcription activity, implying that the motif exists and is vital for late promoter function (García-Escudero & Viñuela, 2000; Rodríguez & Salas, 2013).

Further evidence for the identity of pB236R comes from analyzing the hydropathy profile important in the three-dimensional folding and function of a protein (Gasteiger *et al*., 2003). Again, despite the relatively low level of direct sequence homology in HH-pred, the hydropathy profiles of the ASFV TBP core domain and the equivalent regions of its homologs from other organisms are remarkably similar, as exemplified by the comparison of the *Plasmodium falciparum,* yeast, human, and ASFV profiles (Appendix VII).

The results of Gene Ontology (GO) terms for pB236R were in agreement with an earlier benchmarked dataset having a GO score cut-off of 0.5, in which 74.6% of cellular processes, 76.9% of biological processes, and 85.1% of molecular functions are correctly assigned (Roy *et al*., 2010). The predicted GO scores were all above the 0.5 cut-off for cellular processes, biological processes, and molecular functions. These results also suggest that pB263R binds to conserved promoter motifs in ASFV DNA through reversible non-covalent interactions hence modulating RNA polymerase assemblage at these promoters.

Phylogenetic classification showed that pB236R has an evolutionary relationship with TBPs of other NCLDVs members. Furthermore, the phylogenetic classification of proteins and nucleotides has demonstrated functional similarities, organism types, and structural designs (Thomas & Chiang, 2006). Therefore, this phylogenetic classification study suggests that pB236R is a presumed TBP with homologs in other NCLDVs.

3.5 CONCLUSION

This chapter presents the search and characterization of an ASFV transcription factor using an assortment of different algorithms that employ sequence-based, structure-based, and classification-based approaches. The study results identified a transcription factor pB263R having the requisite features of a TBP. This in retrospect bridges the identified scientific gap of identifying possible transcriptional factors which could interact with TATA-like elements.

CHAPTER FOUR

IN SILICO VALIDATION OF ANTIVIRALS THAT TARGET TATATA MOTIFS SIMILAR TO THOSE IN THE ASFV PROMOTERS

4.1 INTRODUCTION

The DNA minor grooves are receptors of a large number of binding agents such as protein and small molecules. Hence, they are commonly targeted during drug design (Nelson *et al.*, 2007). There is, therefore a need to exploit molecules that can bind to the minor groove to inhibit the replication of DNA viruses. Reversible minor groove binders can be viewed as potential antivirals that can mitigate ASFV replication and transcription (O'Donnell *et al*., 2015). Potential compounds with antiviral activity can be identified and used to design drugs against ASFV. The viral genomes can be targeted by these chemical compounds which act as antiviral agents leading to viral replication and transcription inhibition (Beerman *et al*., 1991; Lown *et al*., 1989; Yakimovich *et al*., 2017). Studies have shown that conserved AT-rich promoter motifs in ASFV are responsible for transcription of late genes (Cackett, Matelska, *et al*., 2020; García-Escudero & Viñuela, 2000; Rodríguez & Salas, 2013).The TATATA promoter motif is an essential region involved in transcription of late genes in ASFV and has been characterized (Cackett, Matelska, *et al*., 2020; Cackett, Sýkora, *et al*., 2020; García-escudero & Viñuela, 2000; Oliveira *et al*., 2017, 2018; Rodríguez & Salas, 2013). The replacement of the TATATA motif with equivalent C and G bases is lethal and leads to a stagnation of ASFV transcription (García-Escudero & Viñuela, 2000). Therefore, this study used these uniquely conserved TATATA motifs in ASFV as potential druggable targets for minor groove binders. Studies have shown that some minor groove binding ligands such as congocidine and distamycin can disrupt transcription (Broyles *et al*., 2004; Chiang

et al., 1994; Knutson *et al*., 2006). However, their primary disadvantage is that they exhibit high cytotoxicity to the host (Bialer *et al*., 1980). Hence, the need to identify minor groove binders that are less cytotoxic or non-cytotoxic.

Initial computational and structural studies on the binding of ligands to A/T elements have elucidated the possibility of targeting motifs of the B-DNA minor groove (Harshman & Dervan, 1985; Kopka *et al*., 1985). Notable studies have been made to understand the energetical aspects involved in DNA-ligand binding (Shaikh *et al*., 2004; Zakrzewska *et al*., 1983). This study evaluated the results of Monte Carlo docking, predicted free energy, and molecular dynamic simulation of tris-benzimidazole, congocidine 3, and congocidine 2, to a synthetic DNA duplex having the TATATA motif similar to those observed in ASFV late promoter regions. This knowledge is therefore important in verifying whether or not TATATA motifs can be targeted by ligands that can separate antiviral activity from cytotoxicity.

4.2 MATERIALS AND METHODS

4.2.1 Docking Preparation

The two-dimensional (2D) ligand structures of congocidine and tris-benzimidazole were retrieved from PubChem (Kim *et al*., 2016). The selection of the two congocidine congeners (congocidine 3 and congocidine 2) was based on reported minimal cytotoxicity compared to congocidine (Bialer *et al*., 1980). The tris-benzimidazole selection was based on its bioisosteric ability to bind AT-rich motifs (Aymami *et al*., 1999; Stockert, 2014). The congocidine congeners were sketched in the Edit molecule tool and minimized in ICM - pro version 3.83 (Abagyan *et al*., 1994).

The congocidine-DNA complex PDB ID 473D containing the TATATA motif, 5'-d(CGTATATACG)2 having a resolution of 1.58 Å was retrieved from PDB (Abrescia *et al*., 1999). The PDB structure was imported in ICM-pro and prepared by removing all water molecules and optimizing hydrogen. Thereafter, the 473D was converted to an ICM object for manipulation. The congocidine ligand was removed from the receptor of 473D for redocking and potential test ligand docking studies of congocidine 2, congocidine 3, and tris-benzimidazole.

The setup receptor tool in ICM-pro molsoft was used to generate a receptor map of the binding site using a grid size of 0.5 Å (Abagyan *et al*., 1994). The 473D receptor was trimmed to remove the terminal unbound guanine, cytosine, and nickel found in the synthetic DNA terminals which were not hydrogen-bonded to conform to the normal G≡C found in DNA structural composition.

The test ligands congocidine 2, congocidine 3, and tris-benzimidazole were converted to Structural Data Format (SDF) and used to search for optimal docking conformations during the simulations. For each ligand, default grid potential maps were calculated and docking runs were performed with a thoroughness value set at 10. The thoroughness value represents the time length of the Monte carlo simulation (Neves *et al*., 2012)

4.2.2 Redocking and Docking of Ligands

Redocking of ligands can determine if an algorithm can accurately reproduce the experimental binding mode of a preexisting x-ray-solved PDB structure. This is of prime importance in estimating the docking score parameters. Congocidine was redocked to 5'd (GTATATAC)2 to estimate the binding score (S) using ICM-pro. The test ligands were used to assess if there was an improvement in the binding score (S). The test set ligands were docked and the top three binding score values from the hit list were sampled for assessment based on ICMs accuracy from a previously benchmarked set (Lam *et al*., 2018; Neves *et al*., 2012). The top-scoring docking poses were visualized using the ICM browser V.3.8-5 (Raush *et al*., 2009).

4.2.3 Minimization of DNA-Ligand Complexes

The sampled test ligand-DNA complexes from ICM-pro were minimized using Chimera software (Pettersen *et al*., 2004). The steepest descent minimization was performed followed by the conjugate gradient minimization. The steepest descent steps used a value of 100 and a steepest descent step size of 0.02 Å. The conjugate gradient steps used a value of 10 and an update interval value of 10. Dock Prep was used to prepare structures by adding missing hydrogens, adding charges, and specifying the saved charges of the ligand-receptor complexes (Pettersen *et al*., 2004).

The terminal phosphates were deleted from the nucleic acid chains and the net charges specified for the respective ligands were set to Gasteiger.

4.2.4 Free Energy Calculations

The preddicta algorithm was used to estimate the ligand-DNA binding affinity (Shaikh & Jayaram, 2007). The energy function used to predict the ligand-DNA calculated binding energy is denoted by the formula:

$$\Delta G^\circ_{cbe} = \Delta H^\circ_{el} + \Delta H^\circ_{vdw} - T\Delta S^\circ_{rt} + \Delta G^\circ_{w}$$

Where ΔG°_{cbe} represents the calculated binding energy.

ΔH°_{el} represents the electrostatic term.

ΔH°_{vdw} represents the van der Waals contribution

$T\Delta S^\circ_{tr}$ represents the translational and rotational entropy on the formation of the ligand-DNA complex, T represents the temperature, ΔS°_{tr} represents the entropy

ΔG°_{w} represents the hydration term.

The calculated ΔG°_{cbe} was used in predicting the binding affinity (ΔG°_{pred}) and the thermal melting temperature (ΔT_m) for the ligand-DNA complexes using the equations below:

$$\Delta G^\circ_{pred} \text{ (predicted)} = (\Delta G^\circ_{cbe} - 62.215) / 7.909$$

$$\Delta T_m \text{ (predicted)} = (\Delta G^\circ_{cbe} - 2.438) / (-1.468)$$

4.2.5 Molecular Dynamics (MD) Simulation

The top-scoring docked poses of ligand- DNA complexes from ICM-pro version 3.83 of the test ligands were transferred to Desmond version 5.3 (Bowers *et al.*, 2006). The conformational changes between the synthetic construct and bound ligands were assessed for the stability of ligands in a dynamic environment and an interaction hypothesis was developed.

Before the MD simulation, the prep wizard tool was used to preprocess the DNA-ligand docked complexes, where bond orders were assigned (Bowers *et al*., 2006). The addition of hydrogen atoms and hydrogen-bonding networks were also optimized. A solvation model of a Monte Carlo-equilibrated, transferable intermolecular potential three-point (TIP3P) water bath was used, with box-shaped orthorhombic boundary conditions, distances of 10 Å × 10 Å × 10 Å, minimized box volume, and buffer box size calculation method. For congocidine 2-DNA, congocidine 3-DNA, and tris-benzimidazole-DNA docked complexes, 14 Na^+, 15 $Na^{+,}$ and 15 Na^+ ions, respectively, were used along with the 0.15 Na^+Cl^- system as neutralizing components. Finally, after these initial equilibration conditions, the MD simulations were carried out by applying the Optimized Potentials for Liquid Simulations 3 (OPLS3) force field and run at a timescale of 5 nanoseconds (ns) at 300 Kelvins (K) (Srivastava *et al*., 2011). The intermediary structures were saved at a time interval of 10 picoseconds (ps) and superimposed with their native structures to calculate the Ligand-Root Mean Square Deviation RMSD (L-RMSD) and Ligand-Root Mean Square Fluctuation (L-RMSF).

4.2.6 L-RMSD and L-RMSF MD Trajectory Analyses

The interaction behavior of the test ligands and synthetic DNA was evaluated using the simulation interaction diagram tool in the Desmond molecular dynamics package v 5.3 (Bowers *et al*., 2006). The L-RMSF and L-RMSD of the ligand MD trajectories were recorded, animations were extracted and two-dimensional ligand interaction diagrams were sampled systematically at intervals of 100 frames to study the synthetic DNA-ligand interaction behavior in the dynamic environment. A summary of docking, free energy calculation, and simulation studies is presented in Figure 4.1.

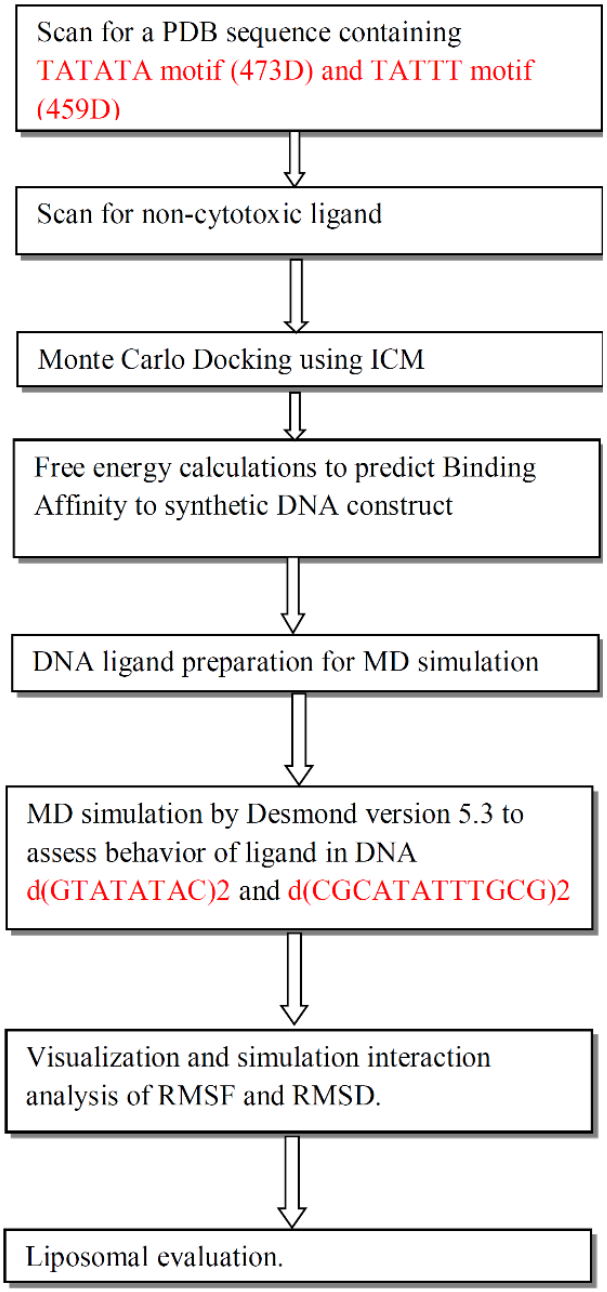


Figure 4.1 Flow diagram representing the summary of ligand binding, free energy calculation, MD simulations, and liposomal evaluation

4.3 RESULTS

4.3.1 Deleting the Terminal Atoms did not Affect the Dimensions of the Minor Groove

The complex comparisons of 5′d (CGTATATACG) 2 and 5′d (GTATATAC)2 which resulted from deleting the terminal nickel bonded guanine/cytosine bases, revealed that the dimensions of the ligand-receptor complexes, the hydrogen bonding distances, local base parameters, local step parameters, and sugar base parameters remained identical for the non-excised regions in both the native congocidine - 5′d (CGTATATACG)2 and congocidine -5′d (GTATATAC)2 duplexes (Appendices IV and V).

4.3.2 Congocidine 1 Could Significantly Redock to the Truncated 5′d (GTATATAC)2 Duplex

The binding score (S), calculated by ICM, from congocidine redocking to the minor groove truncated 5′d (GTATATAC) 2 duplex, was approximately −43.35, with an RMSD of 0.22 Å (Figure 4.2). Based on a benchmarked study of ICM-pro, a score(S) of −32 or lower is important an important statistical threshold in discriminating non-binders from binders https://www.molsoft.com/icmpro/faq-docking.html#faq-score (Neves *et al*., 2012).

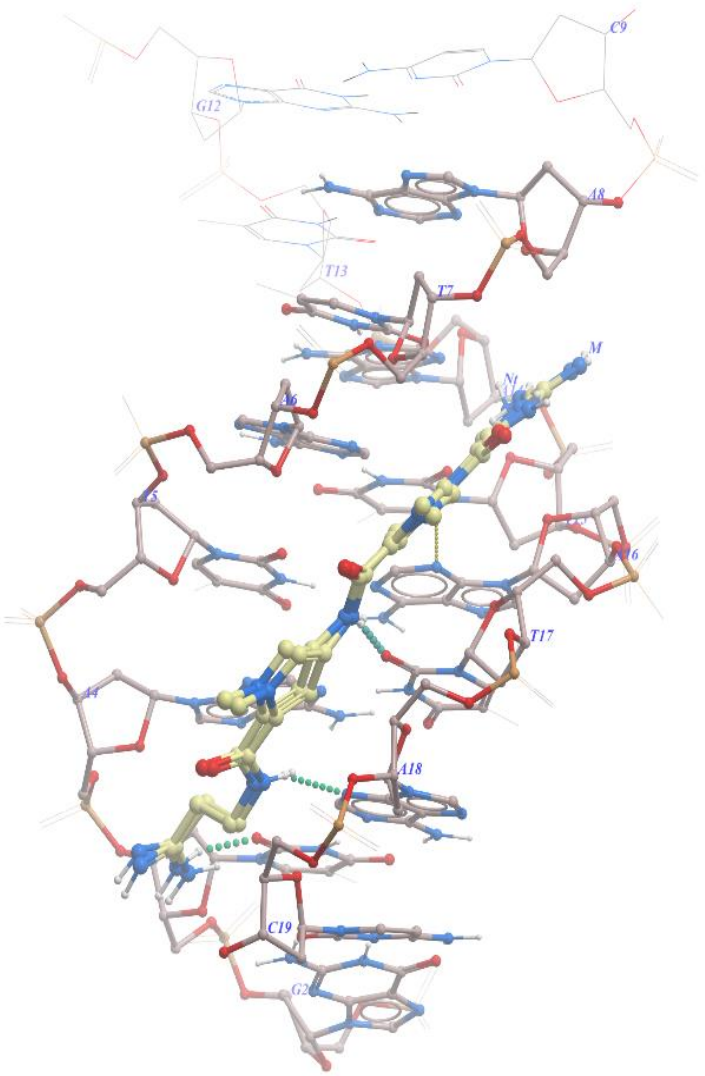


Figure 4.2 Predicted docking pose of congocidine to 5′d (GTATATAC) 2 against the truncated 473D derived structure

Hydrogen bonds in dotted spheres are formed with acceptor atoms O2 of thymine and N3 of adenine. Nt is netropsin/congocidine, and M is a redocked ligand. ICM-pro reproduced the docked pose

4.3.3 Congocidine 2 Congocidine 3 and Tris-Benzimidazole Could Significantly Dock to the 5′d (GTATATAC) 2 Duplex

The tris-benzimidazole, congocidine 3, and congocidine 2 were docked to 5′d (GTATATAC) 2 to validate if they had a significant binding ability (Table 4.1). The top-three conformer poses from the hit list had score ranges of −48.21 to −57.83, −44.83 to −47.27, and −44.08 to −50.29 for congocidine 2, congocidine 3, and tris-benzimidazole respectively (Table 4.1). The obtained score range was comparatively higher than that observed for the redocked congocidine parent molecule (−43.35), signifying that the test ligands were an improvement to the parent compound. The scores of the top three docked pose strongly suggest that tris-benzimidazole, congocidine 3, and congocidine 2 could significantly target TATATA motifs just like congocidine (Table 4.1).

Table 4.1 Top 3 ICM docking parameters for MGBs to 5′d (GTATATAC) 2

Name	Score[1]	Hbond[2]	Hphob[3]	VwInt[4]	Eintl[5]	Dsolv[6]	SolEl[7]
Congocidine	−43.35	−8.10	−6.10	−47.68	6.01	27.50	7.10
Congocidine 2	−57.83	−13.44	−7.41	−56.33	13.60	35.23	12.21
Congocidine 2	−54.94	−12.76	−7.55	−55.49	12.91	35.26	12.81
Congocidine 2	−48.21	−10.12	−7.47	−51.98	7.90	33.70	9.24
Congocidine 3	−47.27	−9.16	−8.42	−54.42	12.85	32.34	13.21
Congocidine 3	−46.70	−8.56	−8.65	−55.09	11.91	33.47	11.75
Congocidine 3	−44.83	−8.12	−8.15	−53.65	10.80	33.95	9.65
Tris-benzimidazole	−50.29	−8.45	−9.02	−53.62	10.42	29.34	13.28
Tris-benzimidazole	−47.84	−9.30	−8.67	−48.51	9.55	29.29	12.18
Tris-benzimidazole	−44.08	−5.89	−9.16	−54.48	11.35	31.60	11.72

Note: [1]Score is the ICM score (−32 and lower are generally considered good scores); [2]Hbond is hydrogen bond energy (lower values are better); [3]Hphob is the hydrophobic energy of the surface exposed to water (lower values are better);[4]VwInt is the van der Waals interaction energy (lower values are better); [5]Eintl is the internal conformation energy of the ligand (lower values are better); [6]Dsolv is the desolvation of exposed H-bond donors and acceptors (lower values are better); [7]SolEl is the solvation electrostatic energy change upon binding (lower values are better).

The docked poses of the three DNA-ligand complexes showed the ligands covered the TATATA motif of the synthetic DNA sequence 5′d (GTATATAC)2 in both reverse and forward orientation. At the center of the duplex, tris-benzimidazole, congocidine 3, and congocidine 2 fit snugly in the minor groove, within the A and T regions of the synthetic DNA construct. A typical binding pattern characterized by hydrogen bonding of the NH amide nitrogen to acceptor atoms O2 of thymine and N3 of adenine was observed in the top-three docked conformers of the congocidine congeners to the central TATATA motif (Figure 4.3A and Figure 4.3B). A similar hydrogen-bonding pattern was detected in the tris-benzimidazole top-three hit-list docked to the central TATATA of the synthetic DNA duplex. The inwards facing NH amide groups of the benzimidazole subunits of tris-benzimidazole formed hydrogen bonds between the adjacent O2 to (T) or N3 (A) bases of the

synthetic 5′d(GTATATAC)2 DNA, these hydrogen bonds were somewhat similar to those detected in congocidine and congocidine congeners (Figure 4.3C). These results show that the tris-benzimidazole and congocidine congeners could act as bioisosteres of each other. Additionally, for the tris-benzimidazole top-scoring conformer, there was displacement by one base pair so that the 3 amino-1-pyrrolidinyl ring was located in a segment in which the minor groove had widened out because of the presence of the first GC base pair. From the simulation experiments of the top three sampled stacked conformers of congocidine 2 and congocidine 3, the likelihood of hydrogen bonding between the terminal amino groups and the electronegative O4 atoms of the ribose sugar of the DNA strands was observed in the absence of water and neutralizing ions. This may be attributed to the enhanced degrees of freedom of flexible terminal amino groups in the ligands which brought the terminal amino groups and electronegative O4 atoms within hydrogen-bonding distance (Figure 4.3A). Some of the sampled poses of the terminal amino groups were observed to form hydrogen bonds with acceptor O2 of cytosine. This was feasible since the protruding NH_2 group of c2-guanine effectively caused steric hindrance.

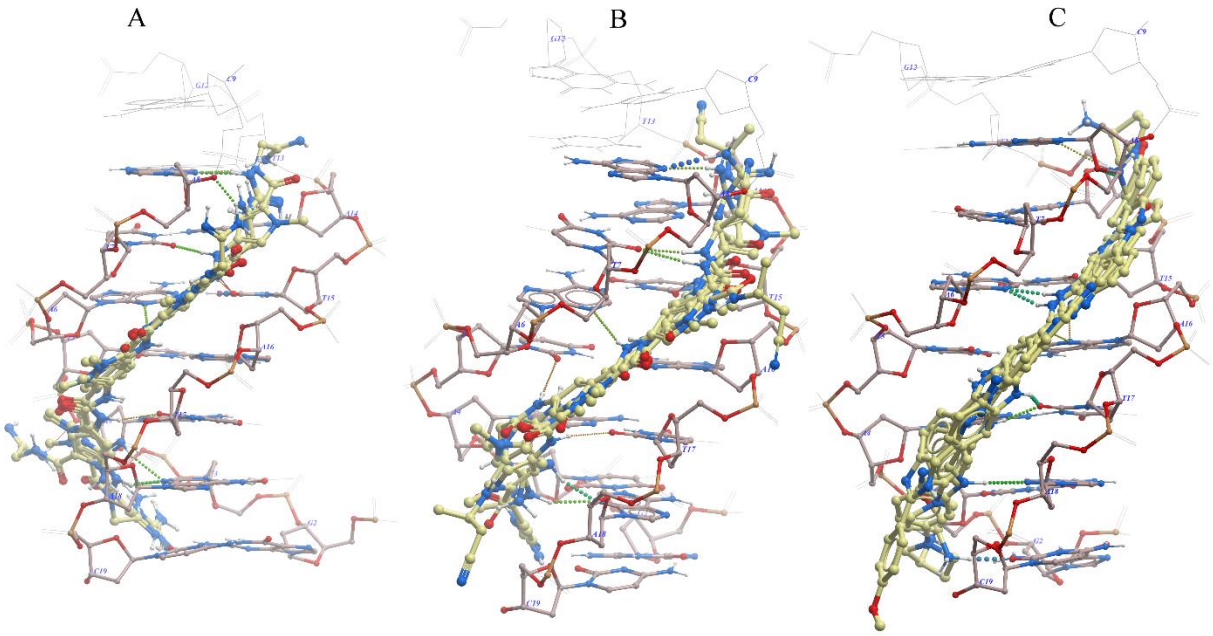


Figure 4.3 Minor groove binders docked to the synthetic DNA 5′d (GTATATAC) 2

Congocidine 2 **(**A), congocidine 3 (B), and tris-benzimidazole (C) (hydrogen bonding is represented in dotted spheres). The simulated structures of the complexes are shown in a skeletal rendering. Most of the ligands are docked to the DNA minor groove.

4.3.4 Shape Complementarity to the Minor Groove was Observed in Docking Test Ligands to the TATATA Region

All the test ligands (congocidine 3, congocidine 2, and tris-benzimidazole) in this study produced an arc-like conformation of the molecules. From the docking poses, the crescent-shaped curvature of congocidine 3, congocidine 2, and tris-benzimidazoles were visibly iso-helical to the turn of the B-form DNA double helix within the minor groove floor (Figure 4.3). A combination of shape isohelicity between the ligand and receptor and associated van der Waals influence has been described as one of the influential factors in the ligand binding process (Fuchs *et al*., 2011; Spitzer

et al., 2009). Significantly higher van der Waals interaction energy values were observed from the tris-benzimidazole, congocidine 2, and congocidine 3, compared to the parent compound congocidine (Table 4.1). This difference in van der Waals interaction energy is a result of the larger binding length due to an extra added benzimidazole and N-methylpyrrole rings along the AT tract of the minor groove in comparison to congocidine. This extra ring probably increased the van der Waals volume of interaction. Generally, close van der Waals contacts between CH groups on the pyrrole rings of the drug molecule and adenine c-2 hydrogens have been described as contributors to van der Waal interaction energy (Kopka *et al*., 1985). Additionally, close van der Waals contacts between unsaturated pi-electrons of the N-methylpyrrole ring in MGBs and the sugar ring-phosphate backbone chains have been revealed to play a vital stability role in DNA -ligand complexes and can be viewed as pseudo intercalation (Coll *et al*., 1989; Teng *et al*., 1988). An analogous stabilizing relationship of van der Waals interaction could be exhibited in tris-benzimidazole, where interactions involving the unsaturated pi-electrons of the benzimidazole subunits have contact with the O4 of the ribose rings (Carrondo *et al*., 1989; Teng *et al*., 1988).

4.3.5 Free Energy Calculations of the Test Ligands Congocidine 3 Congocidine 2 and Tris-benzimidazole were Favorable

The results showed that the calculated binding affinities of the tris-benzimidazole and congocidine congeners had favorable negative values (Table 4.2) and (Shaikh & Jayaram, 2007), which were better than the parent compound, congocidine docked to 5′d (GTATATAC) 2. The ranked order of binding parameters was dominated by the van der Waal factor which positively correlated with high binding affinities (Table 4.2). A larger number of van der Waal contacts, between the extra ligand ring and the minor groove DNA improved the calculated binding affinities. Strong

correlations in the preddicta phenomenological function used here have shown that for $\Delta H°_{vdw}$ (van der Waal contribution), $\Delta G^{o}_{experimental}$, has an R^2 value of approximately 0.80 (Shaikh & Jayaram, 2007). Therefore, the van der Waal contribution may be sufficient enough to calculate the binding affinity. Additionally, data from the literature on the free energy of binding, for analogs of the benzimidazole ligand Hoechst 33258, have shown the dominance of van der Waal interaction (Bostock-Smith & Searle, 1999). Electrostatics and hydrogen bonds have minor roles in the prediction of binding affinity, compared to van der Waal interactions (Table 4.2).

Table 4.2 Binding affinity predictions for docked complexes to TATATA and other complexes

Ligand	$\Delta H°_{el}$[1]	$\Delta H°_{vdw}$[2]	$T\Delta S°_{rt}$[3]	$\Delta G°_{w}$[4]	$\Delta G°_{cbe}$[5]	ΔTm[6]	$\Delta G°_{pred}$[7]
Congocidine (473D)	−5.6	−26.6	25.5	−11.8	−18.5	14.3	−10.2
Congocidine redock	−5.0	−27.9	25.5	−12.0	−19.5	14.9	−10.3
Congocidine 2	−6.2	−33.5	26.0	−15.2	−28.7	21.2	−11.5
Congocidine 2	−6.4	−33.6	26.0	−15.9	−29.7	21.9	−11.6
Congocidine 2	−7.8	−30.3	26.0	−15.2	−27.3	20.3	−11.3
Congocidine 3	−5.5	−35.0	26.1	−17.9	−32.3	23.7	−12.0
Congocidine 3	−5.6	−33.7	26.1	−18.3	−31.4	23.1	−11.8
Congocidine 3	−5.8	−34.2	26.1	−17.0	−30.9	22.7	−11.8
Tris-benzimidazole	−3.4	−37.9	26.1	−19.9	−35.2	25.6	−12.3
Tris-benzimidazole	−3.5	−29.5	26.1	−18.0	−24.9	18.6	−11.0
Tris-benzimidazole	−3.1	−34.3	26.1	−19.7	−31.0	22.8	−11.8
127D Hoechst 33342	−3.6	−30.9	25.3	−16.9	−26.0	19.4	−11.2
264D Hoechst 33258	−4.8	−34.3	25.4	−18.6	−32.3	23.7	−12.0
109D Hoechst 33258	−5.6	−33.7	25.1	−15.7	−29.9	22.0	−11.7
121D Congocidine	−5.3	−27.8	25.5	−14.2	−21.8	16.5	−10.6
261D Congocidine	−8.9	−33.6	25.4	−11.6	−28.7	21.2	−11.5
2DND distamycin	−4.2	−34.2	25.8	−15.5	−28.0	20.8	−11.4
1JTL distamycin	−8.4	−32.5	25.7	−16.9	−32.1	23.5	−11.9

Note: [1]$\Delta H°_{el}$: total electrostatics (kcal/mol); [2]$\Delta H°_{vdw}$: total van der Waals (kcal/mol); [3]$T\Delta S°_{rt}$: rotational and translational entropy (kcal/mol); [4]$\Delta G°_{w}$: hydration free energy (kcal/mol); [5]$\Delta G°_{cbe}$: calculated binding energy (kcal/mol);[6] ΔTm: thermal melting temperature kelvins; [7]$\Delta G°_{pred}$: predicted binding affinity (kcal/mol).

4.3.6 Congocidine 2 MD Simulation Study Showed Terminal Heavy Atom Fluctuation

The DNA-congocidine 2 complex simulation in the dynamic environment revealed the heavy atoms 9 and 6 of the terminal guanidinium fragment had the top-most RMSF (Figure 4.4). This was because the guanine c2-NH_2 amino group in the minor groove exerted steric hindrance towards the entry of the congocidine 2 amino group. Additionally, the increased degrees of freedom of movement of rotatable bonds to the solvent-exposed terminal heavy atoms of congocidine 2 further increased the RMSF. The heavy atoms buried in the minor groove had the least amount of fluctuation (28, 30, 32,4,7,11,13,14,15,29, 33, and 37).

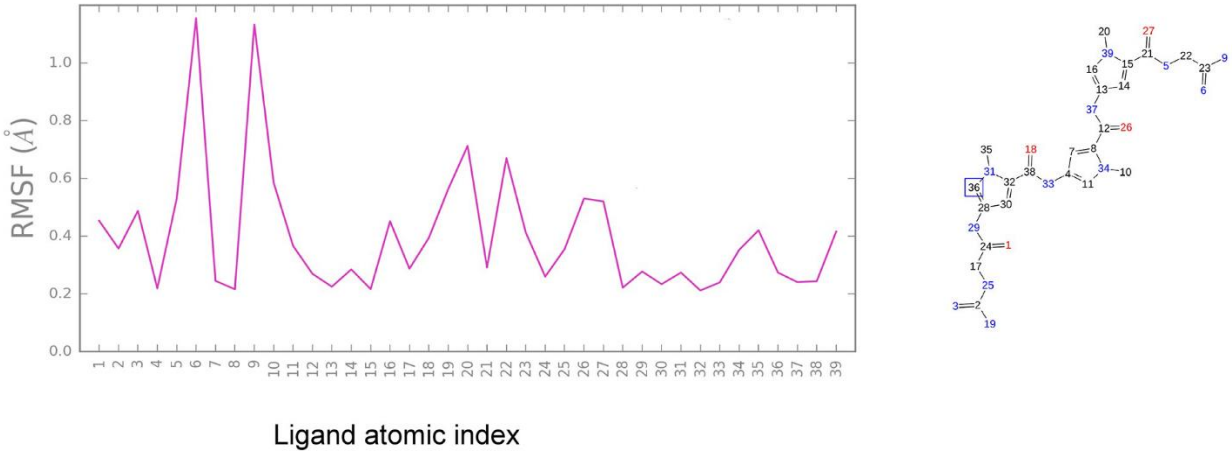


Figure 4.4 Time-dependent RMSF plot in Angstroms (Å) for congocidine 2

The fluctuations over the time course are broken down by atoms; these are compared to the reference frame as a result of superposition.

The RMSD analysis showed congocidine 2 was centered at approximately 0.8 Å, after the transition from the reference conformation at time t=0. Stabilization of the ligand was observed after 0.5 nanoseconds (Figure 4.5).

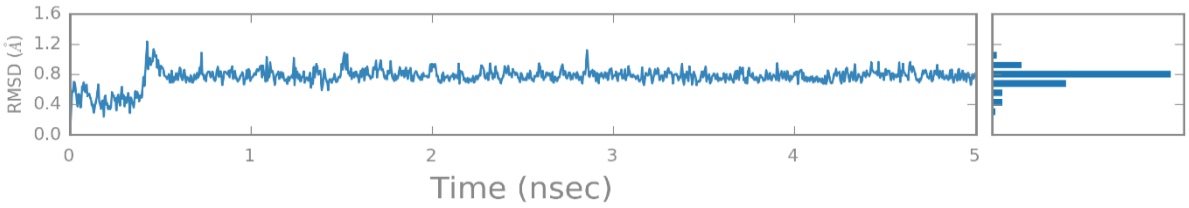


Figure 4.5 The mean RMSD of the ligand as a function of time for congocidine 2 during MD simulation

The plot on the right represents the frequency of distributions relative to the RMSD of the ligand.

From two-dimensional ligand extracts during the 5 ns simulation duration (Figures 4.6, 4.7, and 4.8), DNA- congocidine 2 sampled patterns revealed h-bonding of NH of the ligand to O2 (T) and N3 (A) acceptor atoms in both strands of 5′d(GTATATAC)2. Additionally, pi-cationic bonds during the simulation time course were observed to periodically exist between the A and T bases and the terminal ethanimidamide and guanidinium amino groups. Water molecules were involved in h-bond formation involving C, T, and A (Figures 4.6, 4.7, and 4.8).

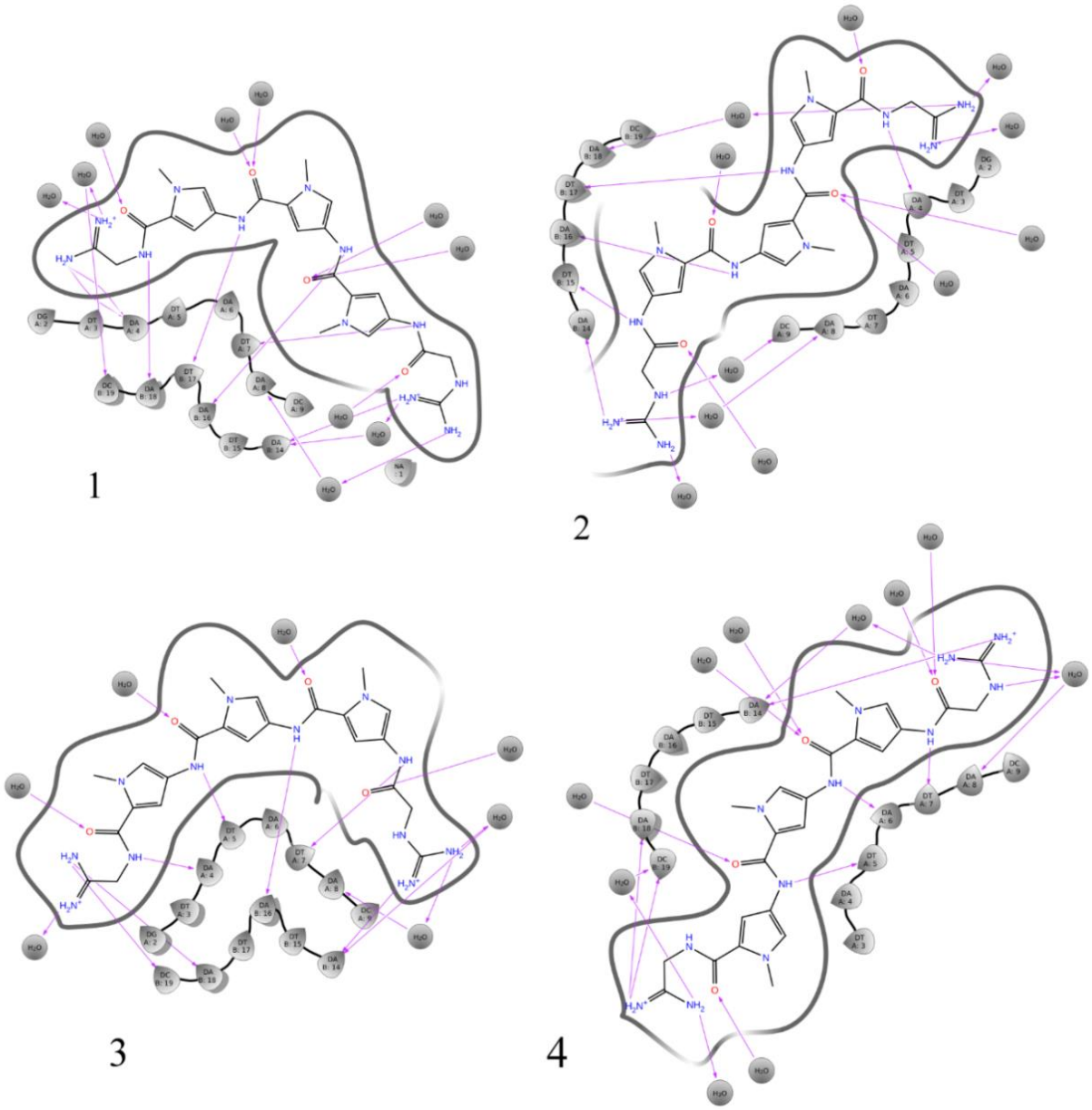


Figure 4.6 Two-dimensional systematically sampled ligand interaction diagrams of the simulation taken at intervals of 100 frames in pictures (1-4) of 5′d(GTATATAC)2 and congocidine 2

Hydrogen bonding is visible for acceptor atoms of A and T. No water molecules hydrogen bond with outward-facing oxygen atoms (pictures 1, 2, 3, and 4). Water molecules are also seen interacting with terminal amino residues of the guanidinium and ethanimidamide and hence the high RMSF in terminal heavy atoms. Water molecules are seen bridging hydrogen-bond formation in picture 1 involving (C19, A8, and A14), in picture 2 involving (A8, A18, and C9), in picture 3 involving (A8, A14, and C9), and in picture 4 involving (A8, C19, and A14). The ligand could be accommodated in one G-C base pair through a water molecule involved in bridge formation picture 2 (C9), or the terminal amino groups forming a hydrogen bond with acceptor oxygen of cytosine pictures 1, 3, and 4 (C19).

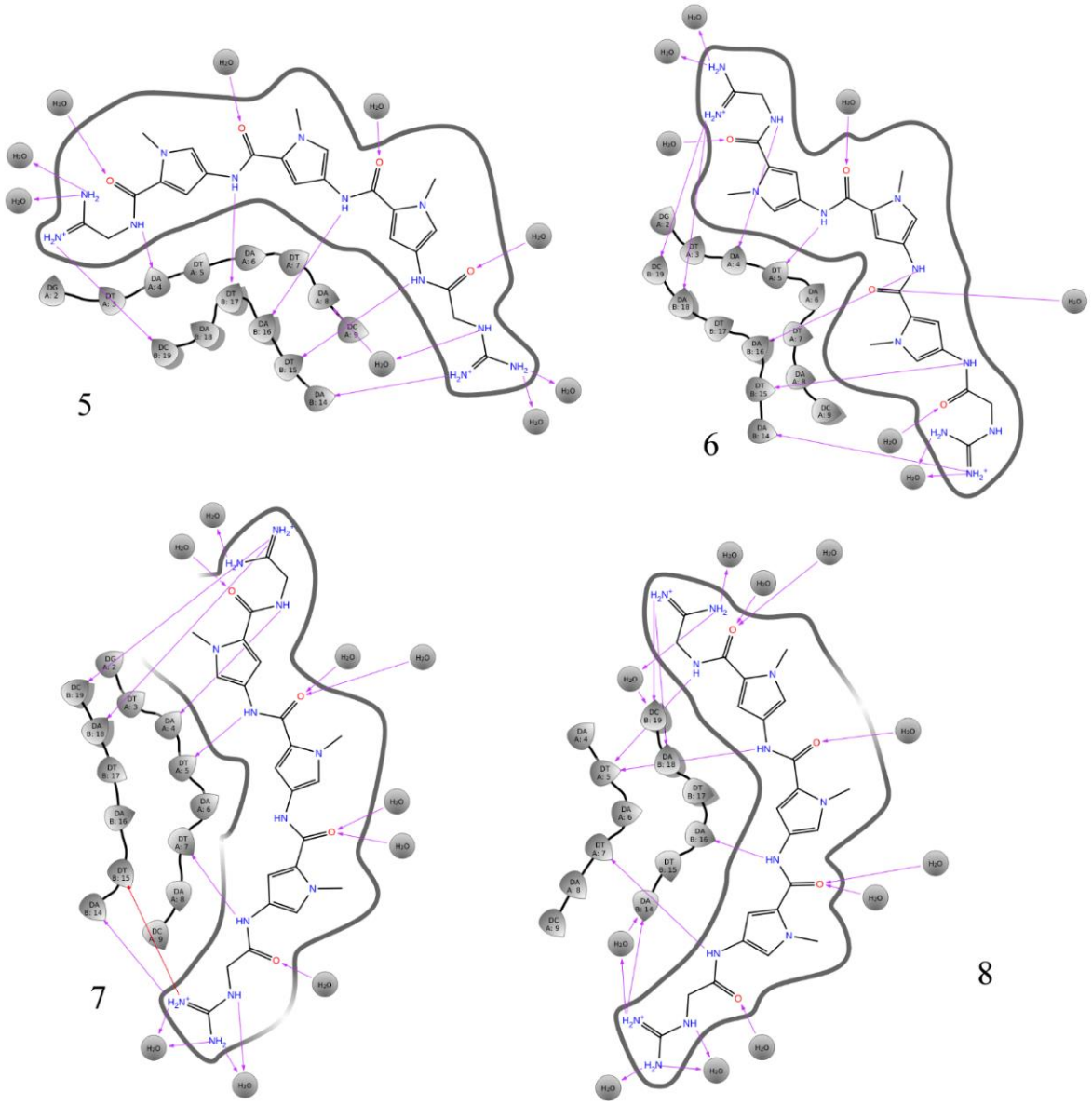


Figure 4.7 Two-dimensional ligand interaction diagrams of the simulation taken at intervals of 100 frames in pictures (5-8) of 5′d(GTATATAC)2 and congocidine 2

Water molecules are seen interacting with terminal heavy atoms in pictures 5, 6, 7, and 8. Water molecules are seen forming bridging hydrogen bonds in picture 8 involving the terminal ethanimidamide amino end (C19), and guanidinium (A14). The possibility of a pi-cation bond from the terminal amino of the guanidinium end to (T15) is also observed in picture 7. The curvature of the ligand is observed to be predominantly iso-helical to the geometry of the minor groove. The ligand could be accommodated in one G-C base pair through a water molecule involved in hydrogen bond bridge formation in picture 8 (C19) or the terminal amino groups forming a hydrogen bond with acceptor oxygen of cytosine in pictures 5,6,7, and 8 (C19).

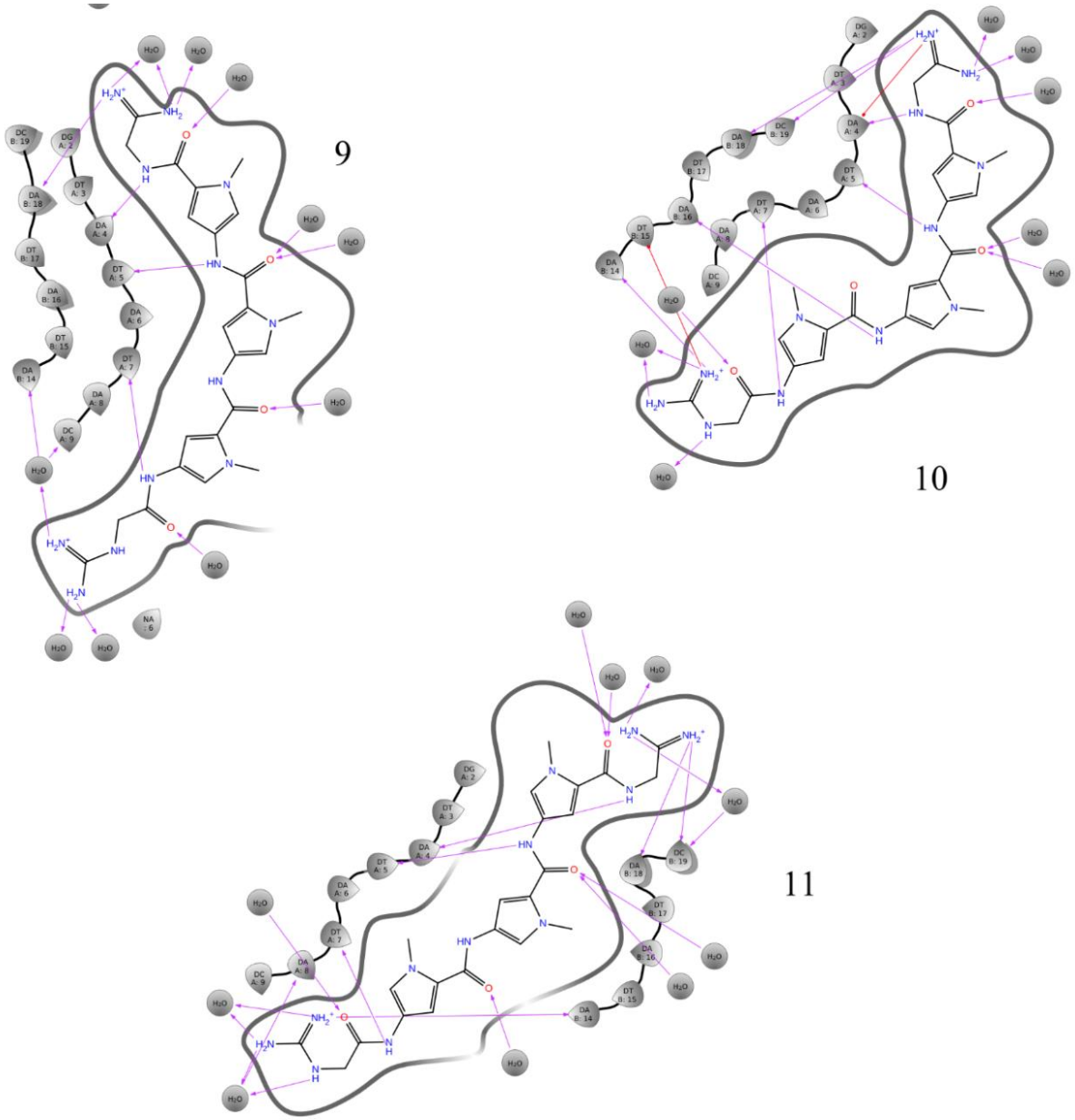


Figure 4.8 Two-dimensional ligand interaction diagram of the simulation taken at intervals of 100 frames in pictures (9-11) of the duplex DNA 5′d(GTATATAC)2 and congocidine 2

Water molecules are seen forming bridging hydrogen bond interactions in pictures 9 (A14 and C9) and picture 11 (A8 and C19). The amino group of the ethanimidamide can also form a hydrogen bond with acceptor oxygen in picture 11(C19). Water molecules are seen interacting with oxygen in pictures (9, 10, and 11). The possibility of the pi-cation bond existing between the amino groups is observed in picture 10 ethanimidamide 10 (T15) and guanidinium end (A4). The ligand could be accommodated in one G-C base pair through a water molecule involved in bridge formation picture 11 (C19) or the terminal ethanimidamide amino groups forming a hydrogen bond with acceptor oxygen of cytosine pictures 10 (C19) and 11(C19).

4.3.7 Congocidine 3 MD Simulation Study Showed Terminal Heavy Atom Fluctuation

In the congocidine 3 RMSF study, the least RMSF occurred in the deeply buried less solvent-exposed part of the minor groove involving the heavy atoms (39, 38, 36,34,33,29, 11, 10, 9, 8, 20, 19, 15, and 14) (Figure 4.9). These heavy atoms had constrained minimal space to rotate and move around. Therefore, an RMSF of around 0.2 Å - 0.3 Å deviation was observed (Figure 4.9). However, the terminal heavy atoms 5 and 6 attached to the guanidinium portion of congocidine 3 showed high RMSFs – recording a maximal atomic RMSF of 1.4 Å (5) and 1.5 Å (6). The butanenitrile end showed a maximal atomic RMSF of 1.3 Å for the heavy atom (3) (Figure 4.9). These high RMSF values are probably due to steric hindrance from the amino group of the guanine c2-NH_2 base towards the guanidinium amino group and the increased degrees of freedom of movement of rotatable bonds towards the solvent-exposed terminal heavy atoms of the ligand.

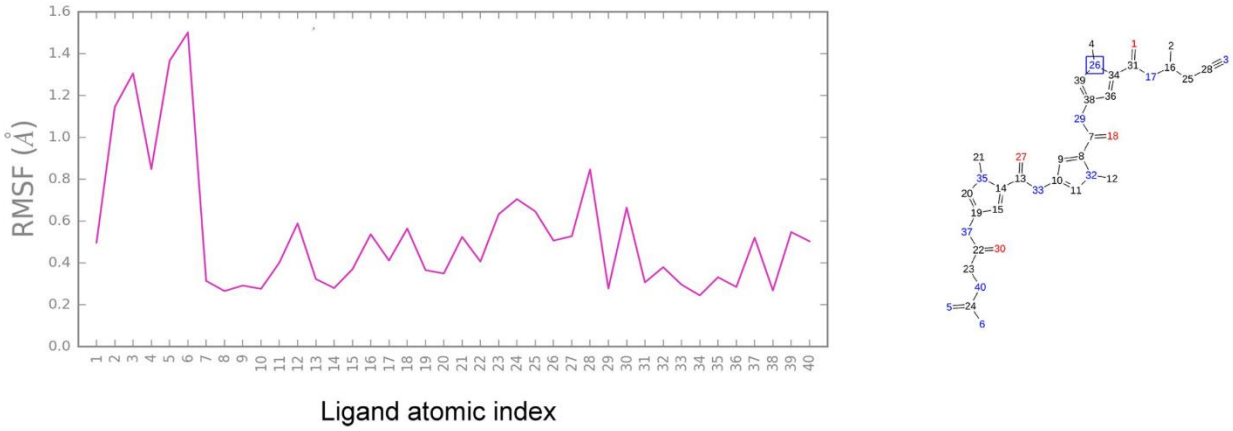


Figure 4.9 Time-dependent RMSF plot in Angstroms (Å) for congocidine 3 showing the individual atom fluctuation.

The fluctuations are broken down by atoms. These are compared to the reference frame as a result of superposition within the minor groove.

The congocidine 3 RMSD was centered on 1.2 Å, relative to the reference conformation at time t=0, after 1.6 nanoseconds from the reference conformation (Figure 4.10).

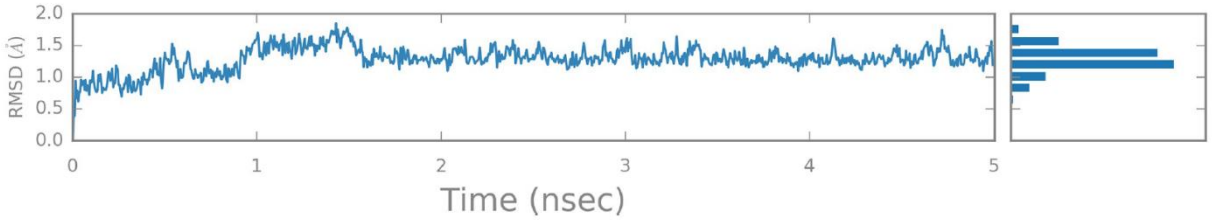


Figure 4.10 Average ligand structure RMSD of congocidine 3 as a function of time and stability profiles

The majority of the congocidine 3 ligand portion remained buried within the minor groove during the 5 ns simulation duration (Figures 4.11, 4.12, and 4.13). A systematically sampled two-dimensional ligand extraction pattern of congocidine 3-DNA interactions revealed characteristic hydrogen bonding of amide nitrogen (NH) to acceptor atoms of O2 thymine (T) and N3 adenine (A). In addition, the terminal guanidinium end was also observed to be able to hydrogen bond. Water molecules were also observed forming bridging hydrogen bonds (Figures 4.11, 4.12, and 4.13).

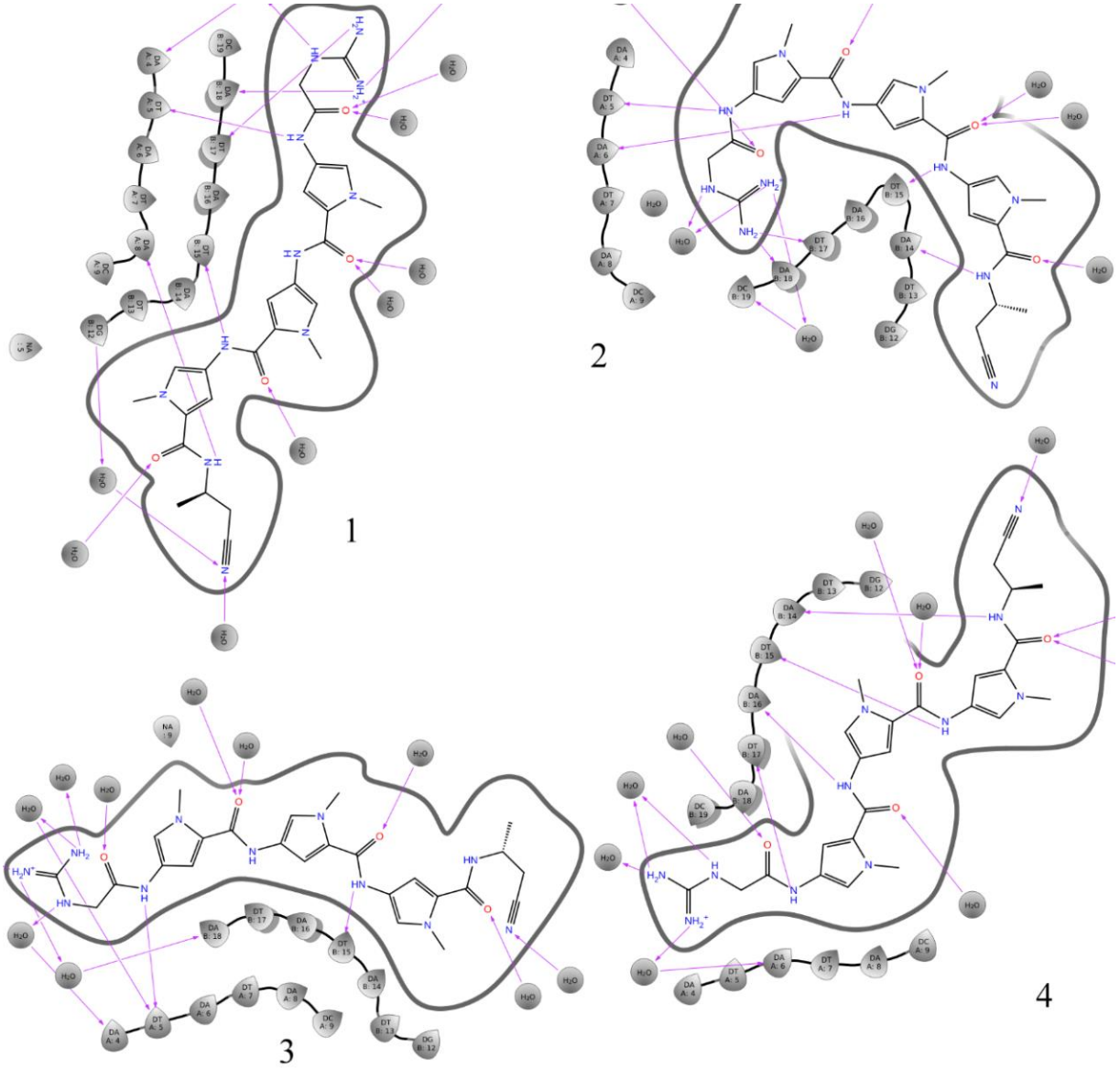


Figure 4.11 Two-dimensional ligand interaction diagrams of the simulation taken at intervals of 100 frames in pictures (1-4) of the duplex DNA 5′d (GTATATAC) 2 and congocidine 3

Characteristic hydrogen bonds involving the inward-facing NH groups are observed where they predominantly interact with acceptor atoms of N3 (A) and O2 (T) bases in pictures (1, 2, 3, and 4). Water molecules are observed to bridge hydrogen bond formation with terminal residues in picture 1 (A4 and G12 (N3)), picture 2 (C19), picture 3 (A4 and A5), and picture 4 (A6). Water molecules are observed interacting with oxygen atoms in pictures (1, 2, 3, and 4) through hydrogen bond donation. The terminal amino of ethanimidamide was also observed to be capable of hydrogen bonding to DNA A and T bases in picture 1 (A4 and A18) and picture 2 (C19 and A18). The ligand could be accommodated by one G-C base-pair through a water molecule involved in hydrogen-bond bridge formation picture 2 (C19).

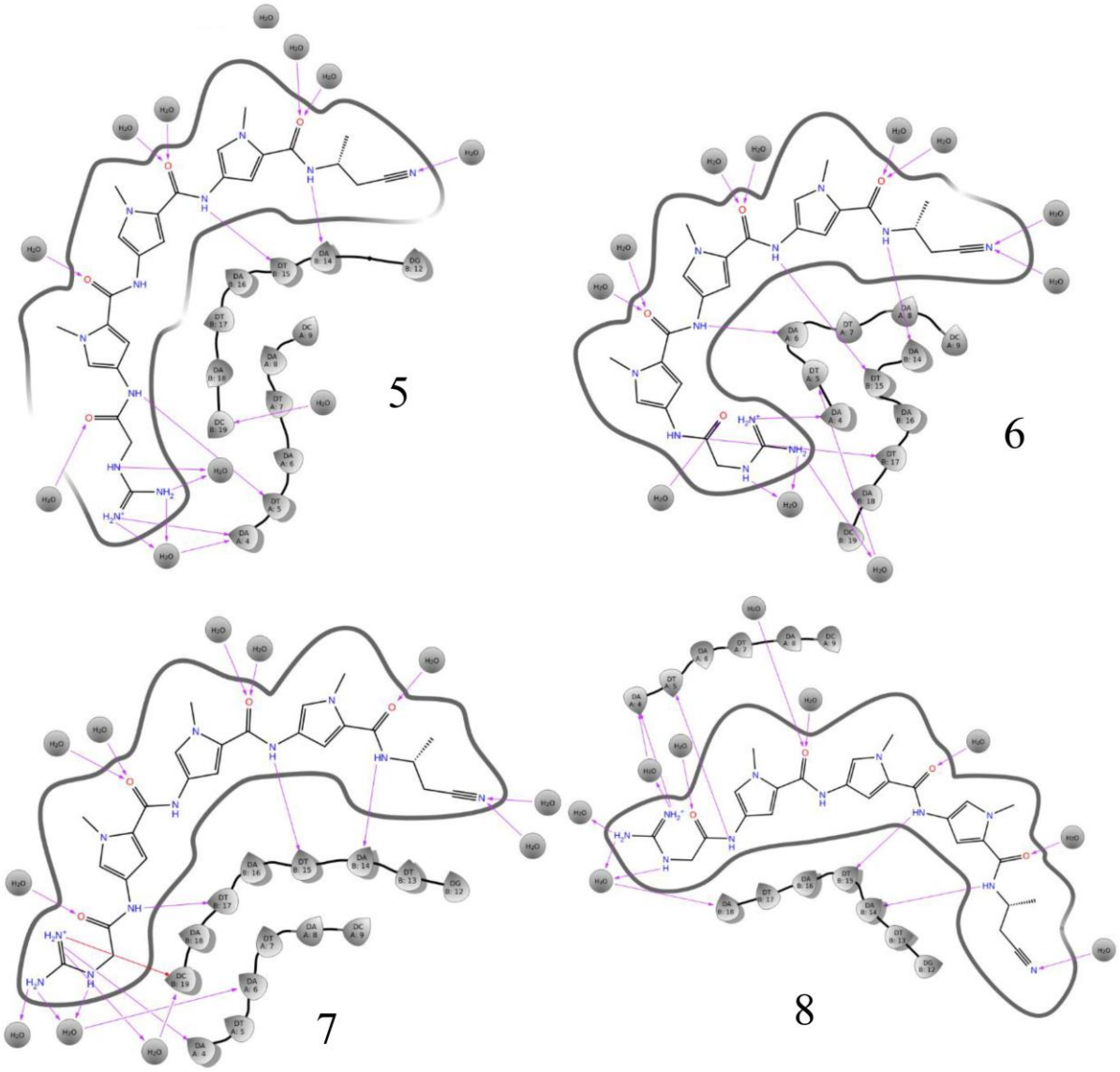


Figure 4.12 Two-dimensional ligand interaction diagrams of the simulation taken at intervals of 100 frames in pictures (5-8) of the duplex DNA 5′d (GTATATAC)2 and congocidine 3

The inward-facing NH groups are seen to form hydrogen bonds predominantly with acceptor atoms of N3 adenine (A) and O2 thymine (T) bases of the duplex. Additionally, the terminal amino groups of guanidinium could hydrogen bond with adenine (A) and O2 thymine (T) bases. Water molecules are observed involved in hydrogen bonding with terminal residues picture 5 (A4), picture 6 (T5), picture 7 (C19 and A6), and picture 8 (A4 and A18). The terminal butanenitrile is observed to interact with water molecules and pi-cationic bonds are observed between the terminal amino group in picture 7 (C 19). The outward-facing oxygen atoms interact with water molecules through hydrogen bond donation. The ligand could be accommodated in one G-C base pair through a water molecule involved in hydrogen bond bridge formation picture 7 (C19).

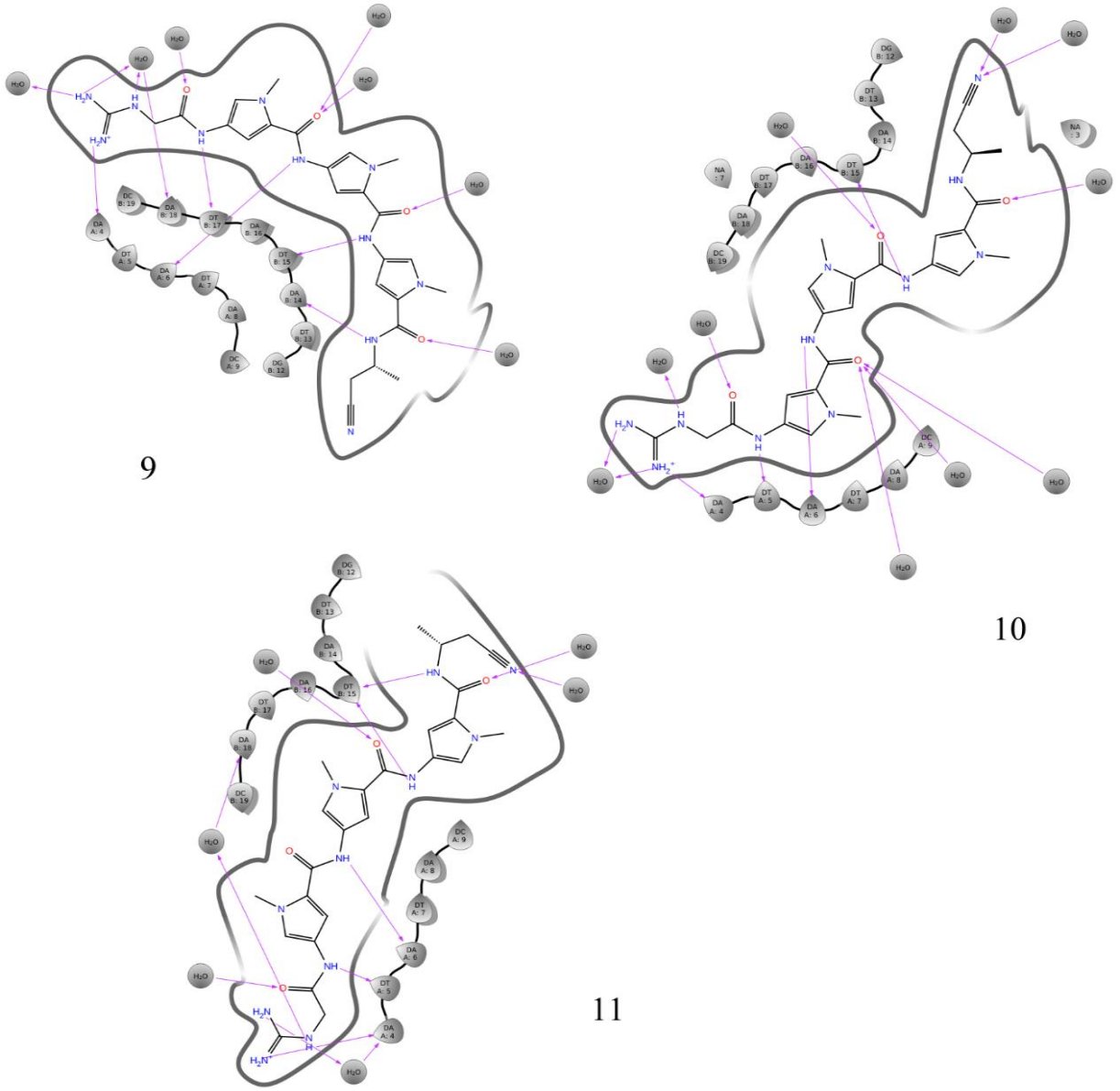


Figure 4.13 Two-dimensional ligand interaction diagrams of the simulation taken at intervals of 100 frames in pictures (9 -11) of the duplex DNA 5′d (GTATATAC)2 and congocidine 3

The inward-facing NH groups are seen to form hydrogen bonds predominantly with acceptor atoms of O2 thymine (T) and N3 adenine (A) bases of the duplex while the outer-facing oxygen atoms are also observed to interact with water molecules. Water molecules are involved in hydrogen bond mediated bridging of terminal residues in picture 9 (A18), picture 11 (A4 and A18). The terminal butanenitrile is observed to interact with water molecules in pictures 10 and 11. The terminal guanidinium is observed to interact with water molecules in pictures 9 and 10. Terminal guanidinium amino groups are also observed to interact with A and T bases in pictures 9, 10, and 11 (A4 and A18).

4.3.8 Tris-Benzimidazole MD Simulation Study Showed Terminal Heavy Atom Fluctuation

The terminal portions of tris-benzimidazole, consisting of heavy atoms (15, 2,27,14,32, 30, and 1) which are solvent-exposed and have an increased degree of freedom of movement, showed high RMSF (Figure 4.14). A high atomic RMSF of approximately 1.1 Å, of heavy atom 30 was observed at the methoxyphenyl terminal. The 3 amino-1-pyrrolidinyl group had a maximum RMSF of 0.8 Å (15) for the heavy atoms. The minimal amount of L-RMSF was observed in heavy atoms (34,36,39,26,5,6,9,12,10,35,11,13,17,16,38,21, and 22) deeply buried in the narrow and less solvent-exposed parts of the minor groove (Figure 4.14). The bonds linking these heavy atoms had constrained rotational movement and were buried within the minor groove, hence, the lower L-RMSF of approximately 0.2 Å-0.3 Å.

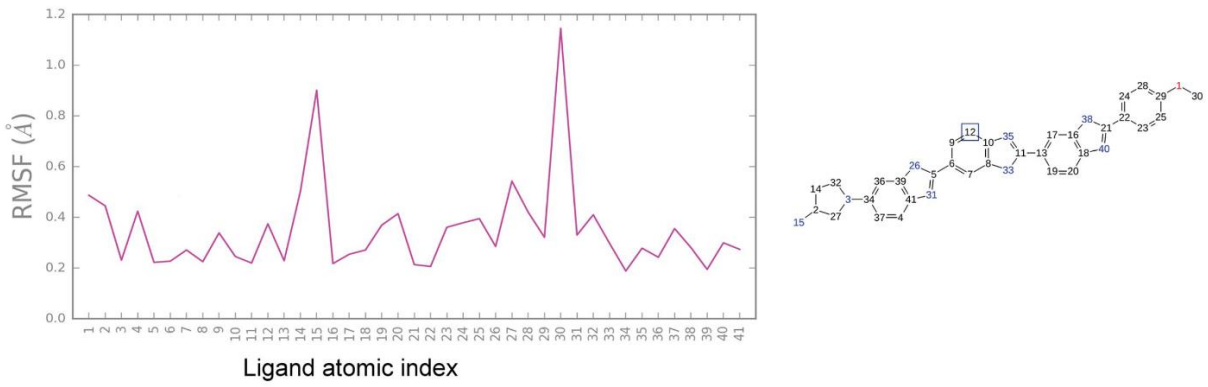


Figure 4.14 Time-dependent RMSF plot for tris-benzimidazole showing heavy atom fluctuations

The RMSD of the ligand was centered at approximately 0.5 Å and 0.6 Å from the frequency distribution on the right, relative to the start frame (Figure 4.15).

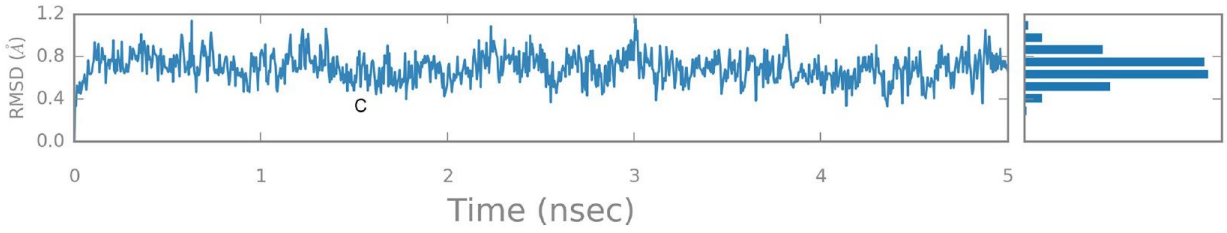


Figure 4.15 RMSD of tris-benzimidazole as a function of time and stability profile

The sampled two-dimensional ligand extraction patterns for the 5ns duration (Figures 4.16, 4.17, and 4.18) showed the tris-benzimidazole remained embedded in the minor groove (Figures 4.16, 4.17, and 4.18). For the tris-benzimidazole-DNA interactions, the inward-facing (NH) of the benzimidazole subunits of tris-benzimidazole participated in hydrogen bonding with acceptor atoms of O2 thymine (T) and N3 adenine (A) of the 5′d(GTATATAC)2 DNA duplex. In addition, the 3 amino-1-pyrrolidinyl group could participate in hydrogen bonding with A and T bases and also form pi-cation bonds with cytosine (Figures 4.16). Water molecules could hydrogen bond with A and T bases at the floor of the minor groove.

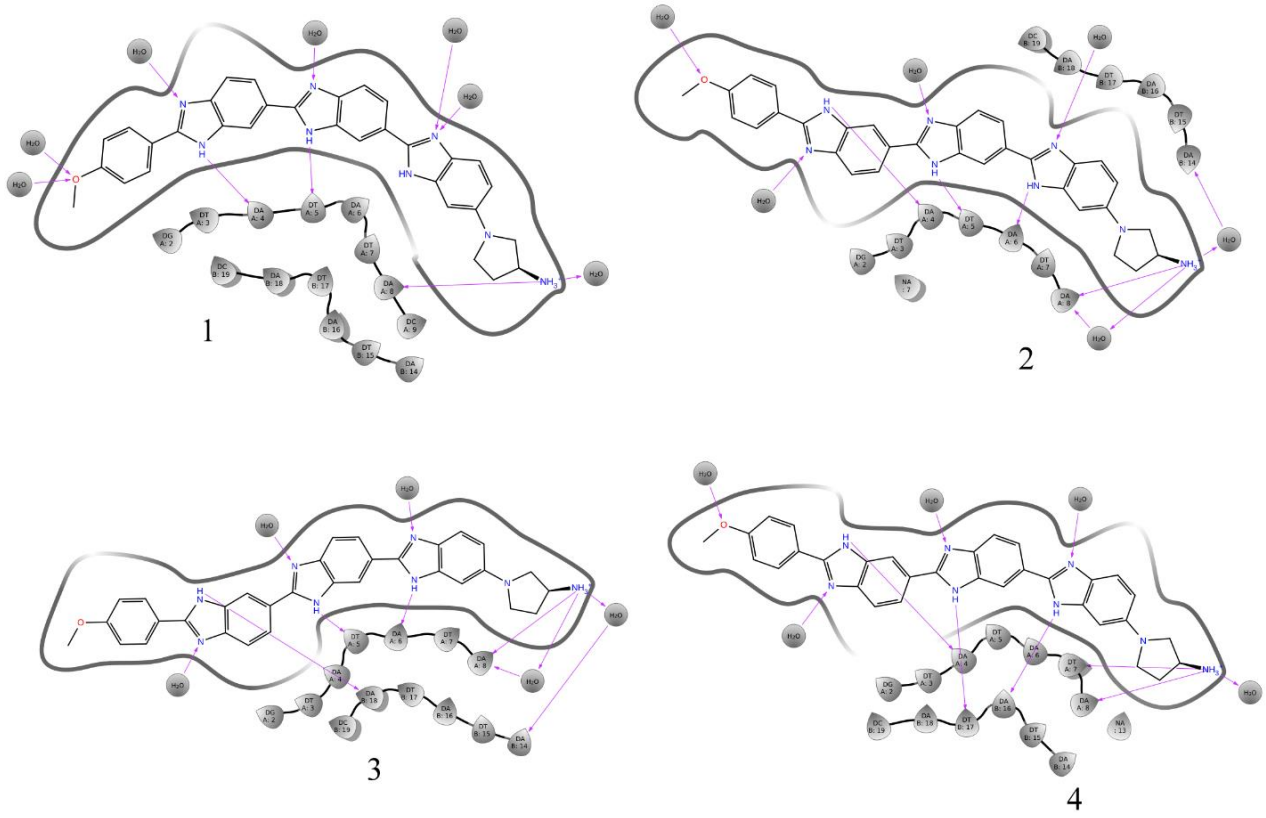


Figure 4.16 Two-dimensional ligand interaction diagram of the simulation taken at intervals of 100 frames in pictures (1-4) of the duplex DNA 5′d (GTATATAC)2 and tris-benzimidazole

The inward-facing NH groups form hydrogen bonds predominantly with acceptor N3 adenine (A) and O2 thymine (T) bases of the duplex. Water molecules are seen bridging the 3-amino-1-pyrrolidinyl group in picture 2 (A8), (A14), and picture 3(A14) through hydrogen bonding. Water molecules are also seen characteristically interacting with the outward-facing nitrogen of the benzimidazole (1, 2, 3, and 4). Water molecules are also seen interacting with the oxygen atom of the methoxyphenyl group.

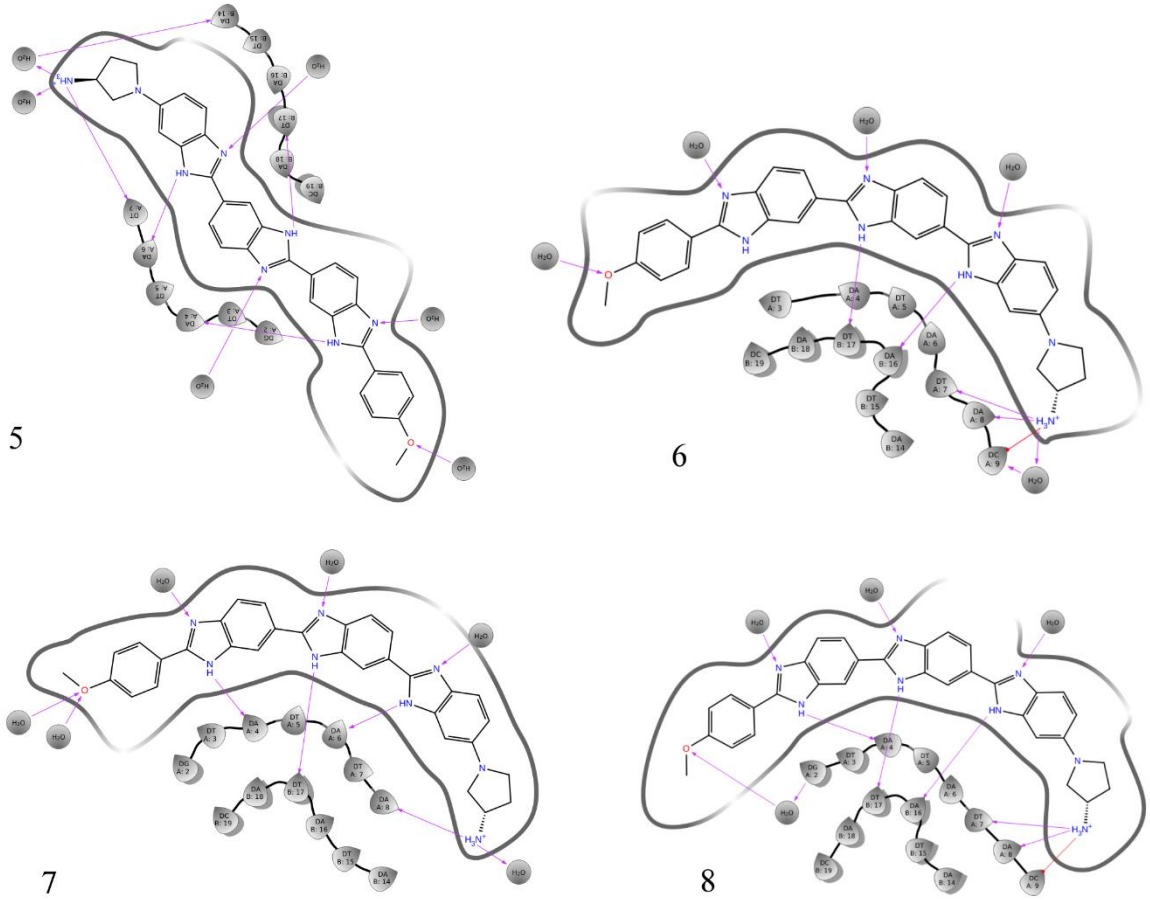


Figure 4.17 Two-dimensional ligand interaction diagrams of the simulation taken at intervals of 100 frames in pictures (5-8) of the duplex DNA 5′d (GTATATAC)2 and tris-benzimidazole

The inward-facing NH groups form typical hydrogen bonds predominantly with acceptor N3 adenine (A) and O2 thymine (T) bases of the duplex. Water molecules bridge hydrogen-bonding interactions between the amino group in the 3-amino-1-pyrrolidinyl and C9 in picture 6. Water molecules are also seen interacting with the outer-facing nitrogen atoms in pictures (5, 6, 7, and 8). Water molecules are also seen interacting with the 3-amino-1-pyrrolidinyl end. The oxygen atom of the methoxyphenyl group is also observed to interact with water molecules. Pi-cationic bonds are observed in pictures (6 and 8) between the 3- amino-1-pyrrolidinyl group and C9. From the ligand interaction in pictures (6, 7, and 8) the curvature extracted seems to be iso-helical to the concave floor of the minor groove. The tris-benzimidazole could be accommodated in one G-C base pair through a water molecule in picture 6 (C9) and picture 8 (C9).

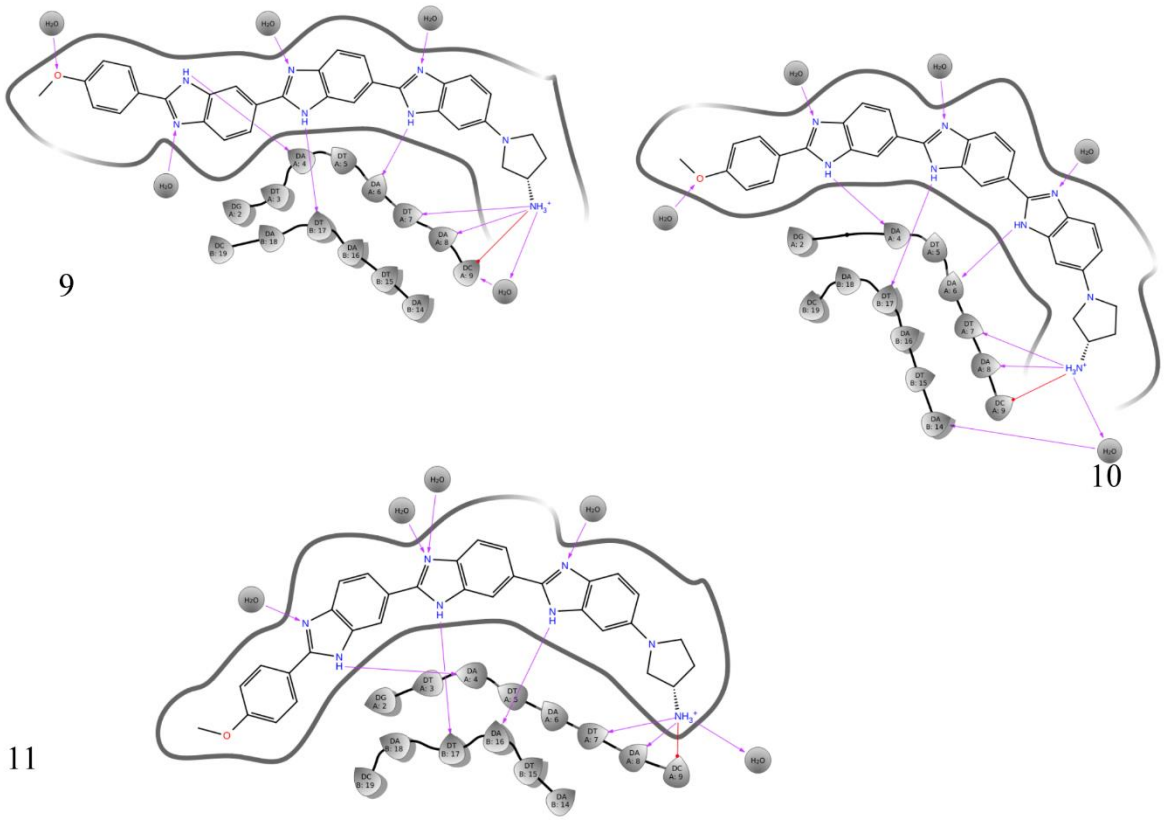


Figure 4.18 Two-dimensional ligand interaction diagrams of the simulation taken at intervals of 100 frames in pictures (9-11) of the duplex DNA 5′d (GTATATAC)2 and tris-benzimidazole

The inward-facing NH groups are observed forming hydrogen bonds predominantly with acceptor atoms of N3 adenine (A) and O2 thymine (T) bases of the duplex. Water molecules are seen bridging interactions through hydrogen bonding between the 3-amino-1-pyrrolidinyl group and acceptor atoms of cytosine (C9) in picture 9, and picture 10 (A14). Water molecules are also seen interacting with the outer-facing nitrogen atoms in pictures (9, 10, and 11). The 3-amino-1-pyrrolidinyl end is also observed to interact with water molecules. The pi-cationic bonds are observed involving 3-amino-1-pyrrolidinyl end with C9 pictures (9, 10, and 11). The ligand could be accommodated in one G-C base pair through a water molecule involved in hydrogen bond bridge formation in picture 9 (C9) and the pi-cationic bond (C9) in pictures (9,10 and 11).

4.4 DISCUSSION

Congocidine has been shown to act as an antiviral by inhibiting the replication of viruses, such as the vaccinia virus and the Shope fibroma virus (Becker *et al*., 1972), and herpes simplex virus (Bialer *et al*., 1980). However, unlike the parent drug congocidine associated with high cytotoxicity, congocidine tri pyrrole derivatives are more potent with minimal cytotoxicity (Bialer *et al*., 1980) making them potentially viable ligands for targeting the ASFV viral promoter. Moreover, an analysis of the ASFV genome base composition of both pre-replicative and post-replicative regions revealed a high prevalence of A and T nucleotides (Yáñez *et al*., 1995). Additionally, apparent conservation of NCLDVs promoter motifs has been demonstrated recently (Cackett, Matelska, *et al*., 2020; Oliveira *et al*., 2018).

The selected test minor groove binders (congocidine 3, congocidine 2, and tris-benzimidazole) docked significantly to the A/T elements of the 5′d(GTATATAC)2 covering the entire TATATA motif of the synthetic DNA construct. Thus, inhibition of ASFV late gene transcription might be postulated involving the interaction of congocidine 3, congocidine 2, and tris-benzimidazole with ASFV TATATA promoters. Late gene transcription is suggested since it occurs in the cytoplasm after viral genome replication and the genome is thus accessible to solutes. For instance, the B646L gene that codes p72 (A major capsid protein) has a crucial TATATA motif (García-Escudero & Viñuela, 2000; Oliveira *et al*., 2017, 2018; Rodríguez & Salas, 2013). The docking and simulation results showed improved binding scores (S) to the central TATATA motif of 5′d (GTATATAC) 2 using the test ligands compared to congocidine. Additionally, binding affinity predictions were improved for the test ligands compared to the parent compound congocidine both predicted and preexisting experimentally observed (Shaikh & Jayaram, 2007).

Hence, its predictable that tris-benzimidazole, congocidine 3, and congocidine 2, can target conserved TATATA promoters in ASFV such as B646L (García-Escudero & Viñuela, 2000), B438L (Inmaculada Galindo *et al*., 2000), A224L (Chacón *et al.,* 1995), E165R (Oliveros *et al.*, 1999) and I329L (Rodríguez & Salas, 2013), and may thus influence the transcription of several ASFV late genes through steric hindrance of transcription factors, and melting/unwinding of DNA. A transcription factor having TATA-binding protein-like features pB263R has been identified in ASFV (Kinyanyi *et al.,* 2018). The TBP binds to TATA motifs and is a prime anchor to other transcription factors involved in binding to DNA. The postulated ASFV TBP/DNA inhibition by tris-benzimidazole and congocidine congeners in ASFV is consistent with studies in the vaccinia virus, in which intermediate and late promoter elements are targeted by TBP (Broyles *et al*., 2004; Knutson *et al*., 2006). Moreover, late and intermediate gene transcription in vaccinia virus was repressed by the MGBs distamycin A and bisbenzimidazole attributable to the presence of essential conserved late TAAA/TAAAT(G/A) and intermediate TAAAT promoters (Broyles *et al*., 2004; Knutson *et al*., 2006; Yakimovich *et al*., 2017). Similarly, ASFV has vital conserved late TATATA motifs as promoters (Oliveira *et al*., 2017, 2018), thus providing binding sites for MGBs and TBPs. MGBs actively compete with TBPs for the minor groove binding site (Chiang *et al*., 1994). Assessments by gel mobility shift assays have further revealed that reversible MGBs are inhibitors of TBP/DNA interactions (Chiang *et al*., 1994), leading to transcription stagnation (Bellorini *et al*., 1995; Broyles *et al*., 2004; Chiang *et al*., 1994; Knutson *et al*., 2006). Therefore, the inhibition of ASFV transcription involving AT-rich late promoters using the tris-benzimidazole and congocidine congeners is predictable.

4.5 CONCLUSION

This study examined *in silico,* the potential of tris-benzimidazole, congocidine 3, and congocidine 2 in targeting TATATA motifs of a synthetic DNA 473D. The simulation study results revealed for the first time, that tris-benzimidazole, congocidine 3, and congocidine 2 could be used in targeting TATATA motifs in synthetic DNA constructs. Similar TATATA motifs are conserved in late genes in ASFV that are essential for transcription.

CHAPTER FIVE

IN SILICO VALIDATION OF ANTIVIRALS THAT TARGET TATTT MOTIFS SIMILAR TO THOSE IN THE ASFV PROMOTERS

5.1 INTRODUCTION

Putative promoter motifs have been characterized in nucleocytoplasmic large DNA viruses (NCLDVs). These include kaumoebavirus, asfarviruses, and faustoviruses (Oliveira *et al*., 2018). The ASFV is a member of NCLDVs with the TATTT promoter motif in its intergenic genomic regions involved in transcription (Cackett, Matelska, *et al*., 2020; García-Escudero & Viñuela, 2000; Oliveira *et al*., 2017, 2018; Rodríguez & Salas, 2013).

Investigational characterization of TATTT involving substitutions and genetic deletions of the promoter sequence of the BA71V p72 gene (major capsid protein), has shown that the promoter activity of the p72 gene occurs at −15 to −11 bases upstream of the transcription start site (García-Escudero & Viñuela, 2000). Replacement of the TATTT motif by a GCGCG sequence in p72 has been shown to abolish the promoter activity and subsequent transcription (García-Escudero & Viñuela, 2000). Thus the TATTT promoter motifs observed in the ASFV genome may act as druggable targets through inhibition of transcription. Various studies have shown that the TATTT motifs can occur as a monomer or as repetitions within intergenic regions close to the transcription start site (García-Escudero & Viñuela, 2000; Oliveira *et al*., 2018; Rodríguez & Salas, 2013). Additionally, copy numbers of TATTT motifs are significantly higher in number than TATATA motifs not only in ASFV but in other NCLDVs such as faustovirus and kaumoebavirus (Oliveira *et al*., 2018). And thus this TATTT motif could be more effective as a drug target in stagnating

promoter-directed transcription. Moreover, the TATTT motif can pair up as TATTT and TATATA as seen in intergenic regions of faustovirus, kaumoebavirus, and asfarvirus (Oliveira *et al*., 2018). Therefore, finding ligand/s that can target additional motifs may be advantageous in targeting transcription efficiency effectively unlike targeting a single TATATA motif. To date, no drugs have been reported to target the characterized TATTT motifs in ASFV. Thus, the objective of this study as described in this chapter was to validate through *in silico* simulation if congocidine congeners and Hoechst 33258 could be used in targeting TATTT motifs and other motifs similar to those observed in ASFV promoters using synthetic DNA constructs. The selection of congocidine 2, congocidine 3, and Hoechst 33258 was based on reported minimal cytotoxicity and bioisosteric ability (Bialer *et al*., 1980; Patel *et al*., 1991; Stockert, 2014; Yakimovich *et al*., 2017).

5.2 MATERIALS AND METHODS

5.2.1 Docking Preparation

The two-dimensional (2D) ligand structure of Hoechst 33258 was retrieved from PubChem (Kim *et al*., 2016). The Edit molecule tool in ICM -pro version 3.83 was used to sketch congocidine 3 and congocidine 2 (Abagyan *et al*., 1994).

The non-self-complementary dodecamer DNA duplex PDB ID 459D, 5′d (CG[5BrC]ATATTTGCG)2 with a resolution of 2.3 Å, the self-complementary 12-mer DNA duplex 5′d (CGCGATATCGCG)2 in complex with congocidine PDB ID 1VTJ, with a resolution of 2.4 Å and the AT-rich duplex of PDB ID 5F9I, a 20-mer DNA duplex, 5′d (CCAATAATCGCGATTATTGG)2 with a resolution of 3.0 Å were all retrieved from the nucleic acid database for simulation studies, based on the fact they have AT motifs similar to those observed in ASFV promoter regions.

The retrieved PDB structures were imported in ICM-pro and prepared by deleting all water molecules and optimizing hydrogen (Abagyan *et al*., 1994). Thereafter, the 1VTJ, 459D, and the 5F9I derivative were converted to ICM objects for manipulation. The congocidine and tris-benzimidazole ligands were removed from the receptors of 1VTJ and 459D respectively. The bromine atom was also removed from the 459D structure. The duplex, 5′d (CCAATAATCGCGATTATTGG)2 of PDB ID 5F9I was truncated to 5′d (CCAATAATCG)2.

The “setup receptor” tool created a binding site receptor map using a default grid size of 0.5 Å for 1VTJ and 459D (Abagyan *et al*., 1994). ICM pocket finder was used to map out the binding site

because the synthetic crystallized 5′d (CCAATAATCG)2 construct had no ligand bound to it. The Hoechst 33258, tris-benzimidazole and congocidine congeners were converted to Structural Data Format (SDF).

5.2.2 Docking of Ligands

The binding modes of the ligands were predicted in ICM-pro using semi-flexible ligand docking in absence of water (Abagyan *et al*., 1994). The top three binding scores from docking compiled from the hit list were retained for assessment of the binding behavior according to a previously benchmarked set (Neves *et al*., 2012). For each test ligand, grid potential maps were calculated and docking runs were performed with a thoroughness value set at 10. The thoroughness value represents the time length of the Monte carlo simulation is scaled between 1-10 (Neves *et al*., 2012)

5.2.3 Minimization of DNA-Ligand Complexes

The receptor-ligand complexes from ICM-pro were minimized using Chimera (Pettersen *et al*., 2004). The steepest descent minimization was performed followed by the conjugate gradient minimization. The steepest descent steps used a value of 100 and a steepest descent step size of 0.02 Å. The conjugate gradient steps used a value of 10 and an update interval value of 10. Dock prep was used to prepare structures by adding missing hydrogens, adding charges, and specifying the saved charges of the ligand-receptor complexes. The net charges specified for the respective ligands were set to Gasteiger.

5.2.4 Free Energy Calculations

The preddicta algorithm was used to estimate the ligand-DNA binding affinity (Shaikh & Jayaram, 2007). The energy function used to predict the ligand-DNA calculated binding energy is denoted by the formula:

$$\Delta G^{\circ}_{cbe} = \Delta H^{\circ}_{el} + \Delta H^{\circ}_{vdw} - T\Delta S^{\circ}_{rt} + \Delta G^{\circ}_{w}$$

Here G°_{cbe} represents the calculated binding energy, H°_{el} represents the electrostatic term, ΔH°_{vdw} represents the van der Waals contribution, $T\Delta S^{\circ}_{rt}$ represents the translational and rotational entropy on the formation of the ligand-DNA complex, and ΔG°_{w} is the hydration term. The calculated ΔG°_{cbe} was used in determining the binding affinity (ΔG°_{pred}) and the thermal melting temperature (ΔT_m) for the ligand-DNA complexes using the equations below:

$$\Delta G^{\circ}_{cbe} = 7.909 \times \Delta G^{\circ}_{pred} + 62.215$$

$$\Delta G^{\circ}_{cbe} = -1.468 \times \Delta T_m + 2.438$$

5.2.5 Molecular Dynamics (MD) Simulation

The top-scoring docked poses of congocidine congeners - DNA and Hoechst 33258- DNA ligand-receptor complexes were transferred to Desmond version 5.3 from ICM-pro version 3.83 (Bowers *et al*., 2006). The conformational changes between the 5′d (CGCA [TATTT] GCG) 2 and bound ligands - congocidine 2, congocidine 3, and Hoechst 33258 were assessed for their stability and ligand interactions in a dynamic environment.

Ligand-DNA docked complexes were preprocessed before the MD simulation, using the prep wizard tool in maestro (Bowers *et al*., 2006), where bond orders were assigned, addition of hydrogen atoms, and hydrogen-bonding networks were also optimized. A solvation model of a

Monte Carlo-equilibrated TIP3P water bath was used, with box-shaped orthorhombic boundary settings, buffer box size calculation method, distances of 10 Å × 10 Å × 10 Å, and minimized box volume. For DNA-congocidine 2 , DNA-congocidine 3 and DNA-Hoechst 33258 docked complexes, 22 Na^+, 23 Na^+ and 23 Na^+ ions, respectively, were used along with the 0.15 Na^+Cl^- system as neutralizing components. Finally, after these initial equilibration conditions, the MD simulations were run by application of the OPLS3 force field at a timescale of 5 nanoseconds and a default 300 K temperature.

5.2.6 Ligand-RMSD and Ligand-RMSF MD Trajectory Analyses

The interaction behavior between the DNA and ligands was investigated by the simulation interaction diagram tool executed in the molecular dynamics package Desmond v 5.3 (Bowers *et al*., 2006). The L-RMSF and L-RMSD of the ligand MD trajectories were recorded, animations were extracted and two-dimensional ligand interaction diagrams were sampled systematically at intervals of 100 frames to study the synthetic DNA-ligand interaction behavior in the dynamic environment

5.2.7 Swiss Absorption Distribution Metabolism and Excretion (SwissADME) Prediction

The *in silico* ligand properties, such as solubility and LogP were determined by SwissADME to evaluate the liposomal packaging ability of test minor groove binders (Daina *et al.*, 2017).

5.3 RESULTS

5.3.1 Congocidine Congeners and Hoechst 33258 Significantly Docked to the TATTT Motif

The synthetic construct 5′d(CGCATATTTGCG)2 of the non-self-complementary dodecamer duplex has the TATTT motif embedded within it. From the Monte Carlo docking runs, the top three stacked poses of the test ligand-DNA complexes had a score range of −65.54 to −70.34 for congocidine 2, −59.72 to −68.71 for congocidine 3, and −42.48 to −45.04 for Hoechst 33258 (Table 5.1).

Table 5.1 ICM top 3 docking parameters to 5′d(CGCATATTTGCG)2

Name	Score[1]	Hbond[2]	Hphob[3]	VwInt[4]	Eintl[5]	Dsolv[6]	SolEl[7]
Congocidine 2	−70.34	−14.58	−7.86	−60.55	20.25	36.90	2.53
Congocidine 2	−69.92	−12.40	−8.17	−64.53	12.80	32.10	6.62
Congocidine 2	−65.54	−13.56	−7.88	−57.65	15.44	31.06	7.92
Congocidine 3	−68.71	−12.03	−9.38	−64.41	22.42	34.09	5.59
Congocidine 3	−62.99	−10.13	−8.67	−60.17	17.74	30.58	4.04
Congocidine 3	−59.72	−9.64	−8.27	−52.18	13.53	27.39	−1.71
Hoechst 33258	−45.04	−3.64	−8.51	−46.35	3.77	21.04	0.67
Hoechst 33258	−43.74	−4.47	−8.17	−45.12	3.76	22.70	1.72
Hoechst 33258	−42.48	−4.17	−8.61	−42.53	3.15	19.92	2.14

Note: [1]Score is the ICM score (−32 and lower are generally considered good scores); [2] Hbond is hydrogen bond energy (lower values are better);[3]Hphob is the hydrophobic energy of the surface exposed to water (lower values are better);[4]VwInt is the van der Waals interaction energy (lower values are better);[5]Eintl is the internal conformation energy of the ligand (lower values are better);[6]Dsolv is the desolvation of exposed h-bond donors and acceptors (lower values are better); [7]SolEl is the solvation electrostatic energy change upon binding (lower values are better).

For the docked congocidine 2, the flexible amino-guanidinium region was observed of being capable of avoiding the protruding amino group of G22 and forming a hydrogen bond with N3 of

G22 (2.28 Å), hydrogen bonds could also be formed between the ligand (NH) amide groups and acceptor O2 of thymine (2.04 Å) of T5 and acceptor O2 of T7 (2.41 Å) and acceptor O2 of T8 (2.17 Å). For congocidine 2 coverage of the TATTT of the motif was feasible within the 5′d(CGCATATTTGCG)2 duplex (Figure 5.1).

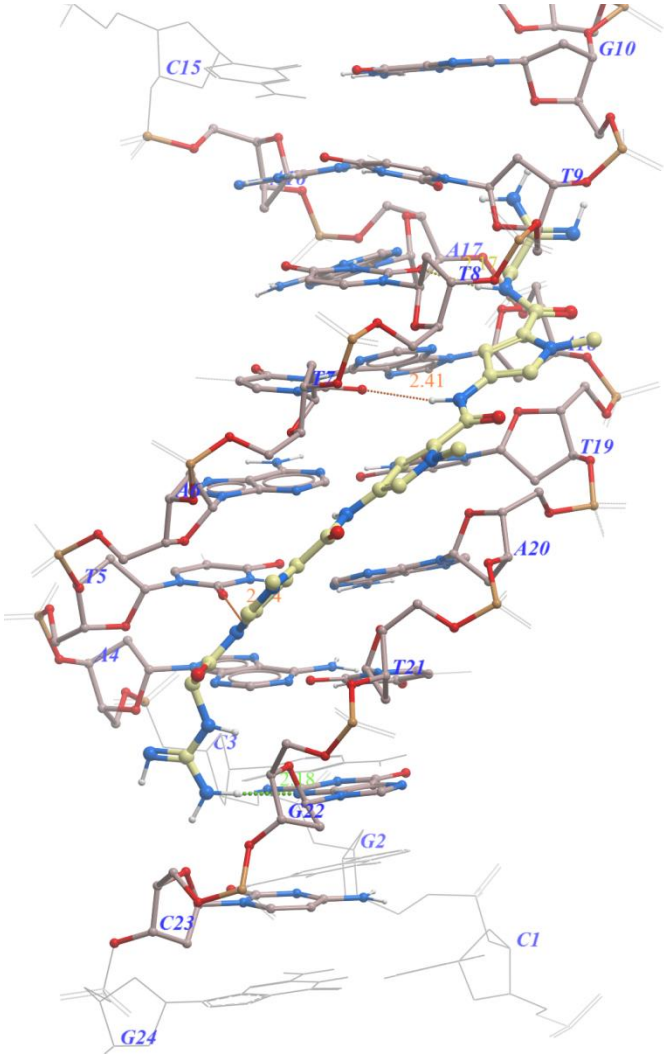


Figure 5.1 Simulated stereo interaction diagram showing congocidine 2 bonded to the minor groove of 5′d(CGCATATTTGCG)2 duplex

Congocidine 2 forms hydrogen bonds O2T8, O2T7, O2T5, and N3G22 represented in dotted spheres.

For congocidine 3 the flexible guanidinium amino region was capable of forming a hydrogen bond with G10 (1.62 Å) (Figure 5.2). A hydrogen bond could also be formed between amide groups and acceptor N3 of A18 (2.18 Å). A bifurcating hydrogen bond could also be formed between the O4 ribose and A17 (2.04 Å and 2.28 Å) in water absence. For congocidine 3 coverage of the TATTT motif (T5-T9) was observed within the 5′d(CGCATATTTGCG)2 minor groove of the duplex.

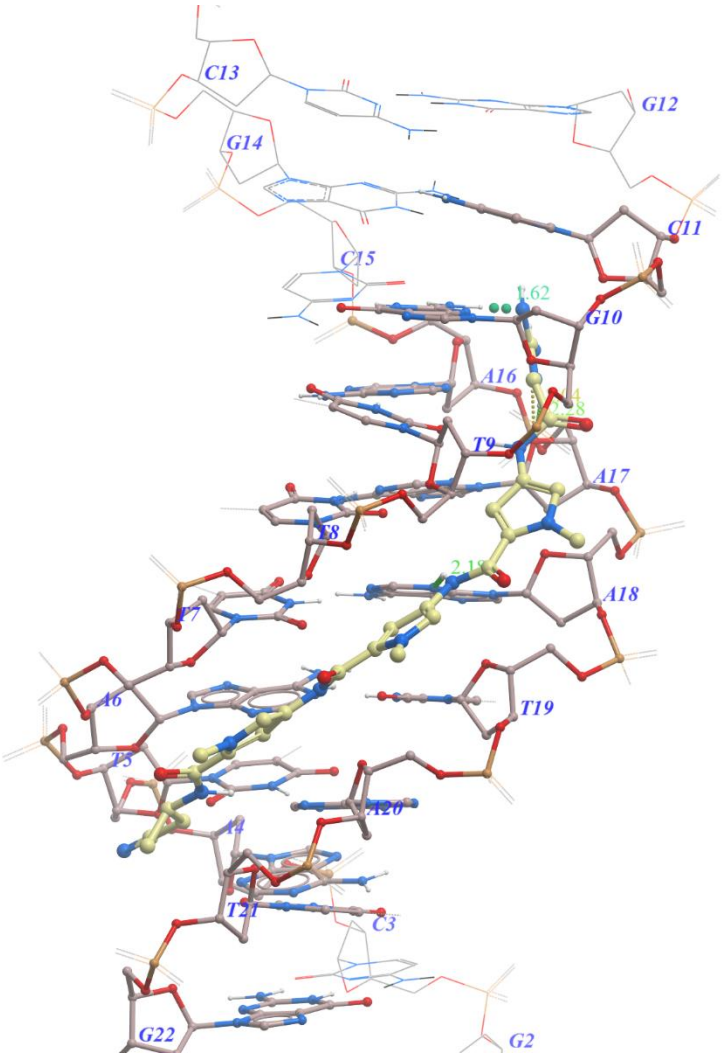


Figure 5.2 Simulated stereo interaction diagram showing congocidine 3 molecule bonded to the minor groove of 5′d(CGCATATTTGCG)2 duplex

Congocidine 3 forms hydrogen bonds N3G10, O4NH (A17), O4NH_2 (A17), and N3A18 represented in dotted spheres.

Hoechst 33258 was observed to be capable of forming a hydrogen bond with A18 (1.99 Å) and acceptor N3 of T19 (2.46 Å). For Hoechst 33258, coverage of the TATTT motif (T5-T9) was observed within the 5′d(CGCATATTTGCG)2 minor groove of the duplex (Figure 5.3).

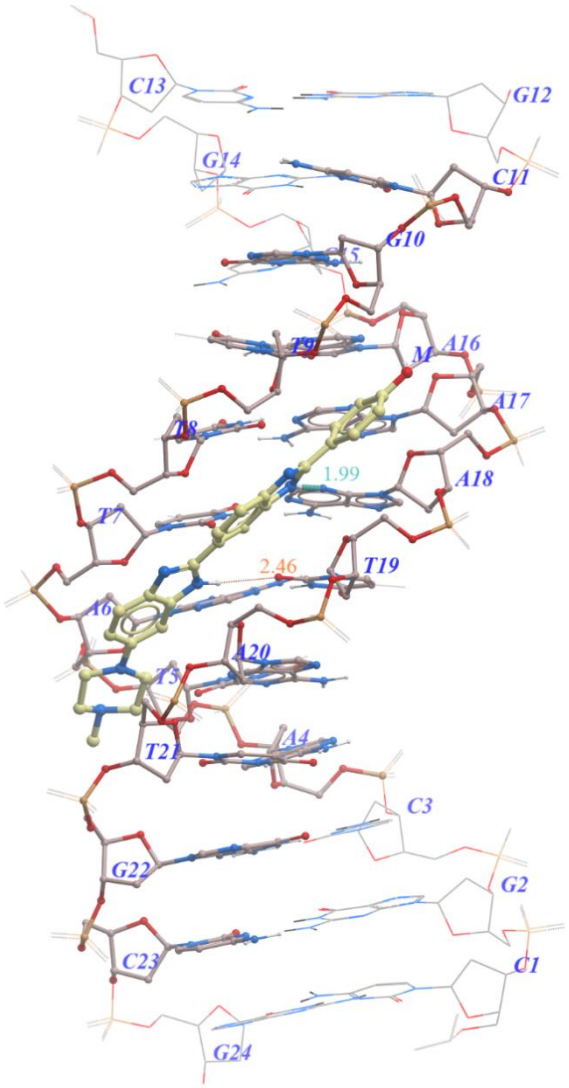


Figure 5.3 Simulated stereo interaction diagram showing Hoechst 33258 molecule bonded to the minor groove of 5′d(CGCATATTTGCG)2 duplex

Hoechst 33258 forms hydrogen bonds N3A18 and O2T19 represented in dotted spheres.

5.3.2 Free Energy Calculations of Congocidine Congeners Tris-benzimidazole and Hoechst 33258 to the Duplex 5′d(CGCATATTTGCG)2

The docked single structure complexes of congocidine 3, congeners 2, tris-benzimidazole, and Hoechst 33258 to 5′d (CGCATATTTGCG)2 had the predicted binding affinities ($\Delta G\ ^{\circ}_{pred}$) remaining comparable to the known experimental values (Table 4.2 and Table 5.2). Both the van der Waals factors (ΔH°_{vdw}) and thermal melting temperatures (ΔTm) seemed to positively correlate with predicted binding affinities for the congocidine congeners, Hoechst 33258 and tris-benzimidazole.

Table 5.2 Free energy predictions for docked complexes to 5′d(CGCATATTTGCG)2

Ligand	ΔH°_{el}[1]	ΔH°_{vdw}[2]	$T\Delta S^{\circ}_{rt}$[3]	ΔG°_{w}[4]	ΔG°_{cbe}[5]	ΔTm[6]	$\Delta G\ ^{\circ}_{pred}$[7]
Congocidine 2	−8.8	−42.4	26.1	−16.7	−41.7	30.1	−13.1
Congocidine 2	−8.2	−41.3	26.1	−17.1	−40.6	29.3	−13.0
Congocidine 2	−7.4	−42.0	26.1	−17.4	−40.7	29.3	−13.0
Congocidine 3	−8.2	−45.7	26.1	−19.8	−45.8	32.8	−13.7
Congocidine 3	−7.6	−42.4	26.1	−16.7	−41.7	30.7	−13.1
Congocidine 3	−7.0	−45.6	26.2	−19.9	−46.3	33.2	−13.7
Tris-benzimidazole	−5.4	−39.1	26.1	−22.3	−40.6	29.3	−13.0
Hoechst 33258	−4.8	−38.5	25.3	−19.1	−37.0	26.8	−12.5
Hoechst 33258	−4.2	−33.3	25.3	−18.3	−30.5	22.4	−11.7
Hoechst 33258	−4.3	−36.9	25.3	−19.2	−35.1	25.6	−12.3

Note:[1]ΔH°_{el}: total electrostatics (kcal/mol);[2]ΔH°_{vdw}: total van der Waals (kcal/mol);[3]$T\Delta S^{\circ}_{rt}$: rotational and translational entropy (kcal/mol);[4]ΔG°_{w}: hydration free energy (kcal/mol);[5]ΔG°_{cbe}: calculated binding energy (kcal/mol);[6]ΔTm: thermal melting temperature in kelvins; [7]$\Delta G\ ^{\circ}_{pred}$: predicted binding affinity (kcal/mol).

5.3.3 Congocidine 2 TATTT Molecular Dynamics Simulation Study Showed Terminal Heavy Atom Fluctuation

The RMSF analysis of congocidine 2 showed heavy atoms, 3(1.4 Å), 6(1.2 Å), 9(1.3 Å), and 19(0.9 Å) had high L-RMSFs (Figure 5.4). These heavy atoms showed the high RMSFs, which were due to the guanine amino group (c2-NH_2) that exerted steric hindrance towards the entry of the guanidinium amino group in the minor groove. Also, the terminal ligand ends had improved degrees of freedom of movement of rotatable bonds towards the solvent-exposed terminal heavy atoms, thus the high L-RMSF (Figure 5.4). The heavy atoms in the buried portion of the minor groove had less solvent exposure and minimal space to rotate around, thus the lower RMSF of around 0.2 Å - 0.3 Å was observed in heavy atoms (13,14,7,8,30,32, 33, and 37) (Figure 5.4). Congocidine 2 remained embedded in the minor groove during the simulation.

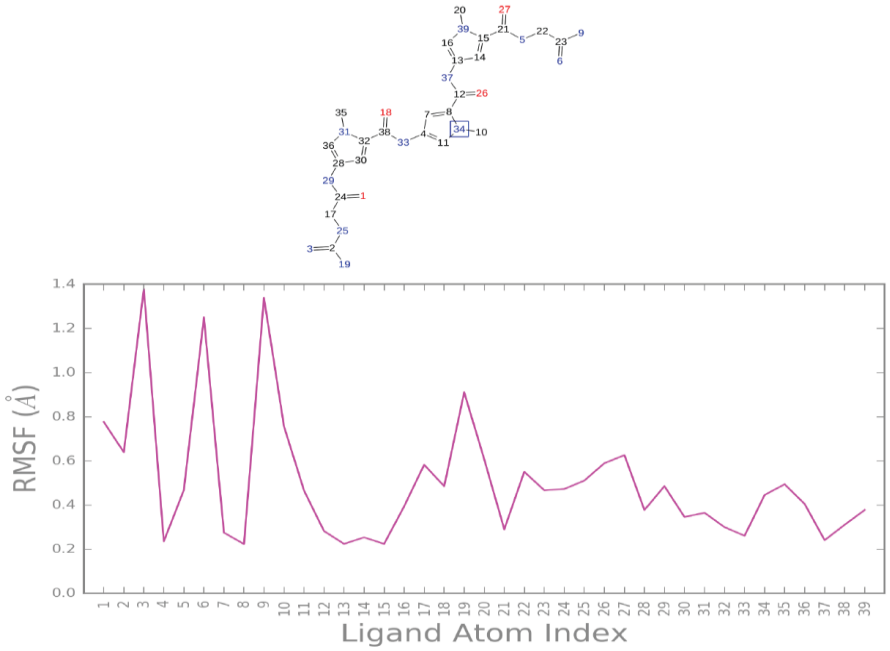


Figure 5.4 Time-dependent L-RMSF plot in Angstroms (Å) for congocidine 2 within 5′d (CGCATATTTGCG) 2 showing the heavy atom fluctuation

The fluctuations are broken down by atoms; these are compared to the reference frame as a result of superposition over the time course of the simulation within the minor groove.

The ligand RMSD seemed to have centered at approximately 0.8 Å, relative to the reference conformation at time t=0 (Figure 5.5).

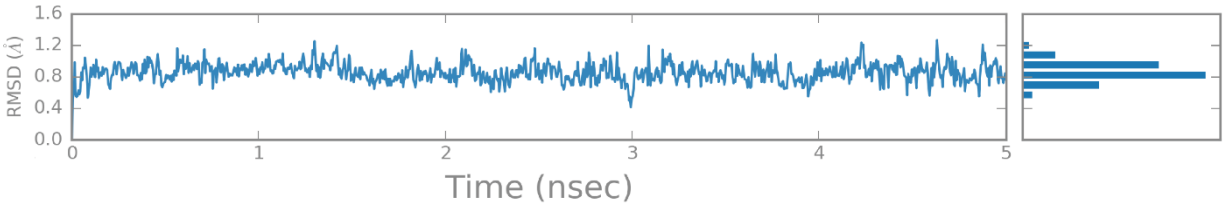


Figure 5.5 The L-RMSD of congocidine 2 as a function of time and stability profiles

The sampled 2D ligand image extracts during the 5 ns simulation duration showed characteristic h-bonding of amide nitrogen (NH) of congocidine 2 to acceptor atoms of O2 thymine (T) and N3 adenine (A). Besides, the terminal guanidinium end was capable of h-bonding directly to bases or through water-mediated hydrogen bonding with bases at the minor groove floor (Figures 5.6, 5.7, and 5.8).

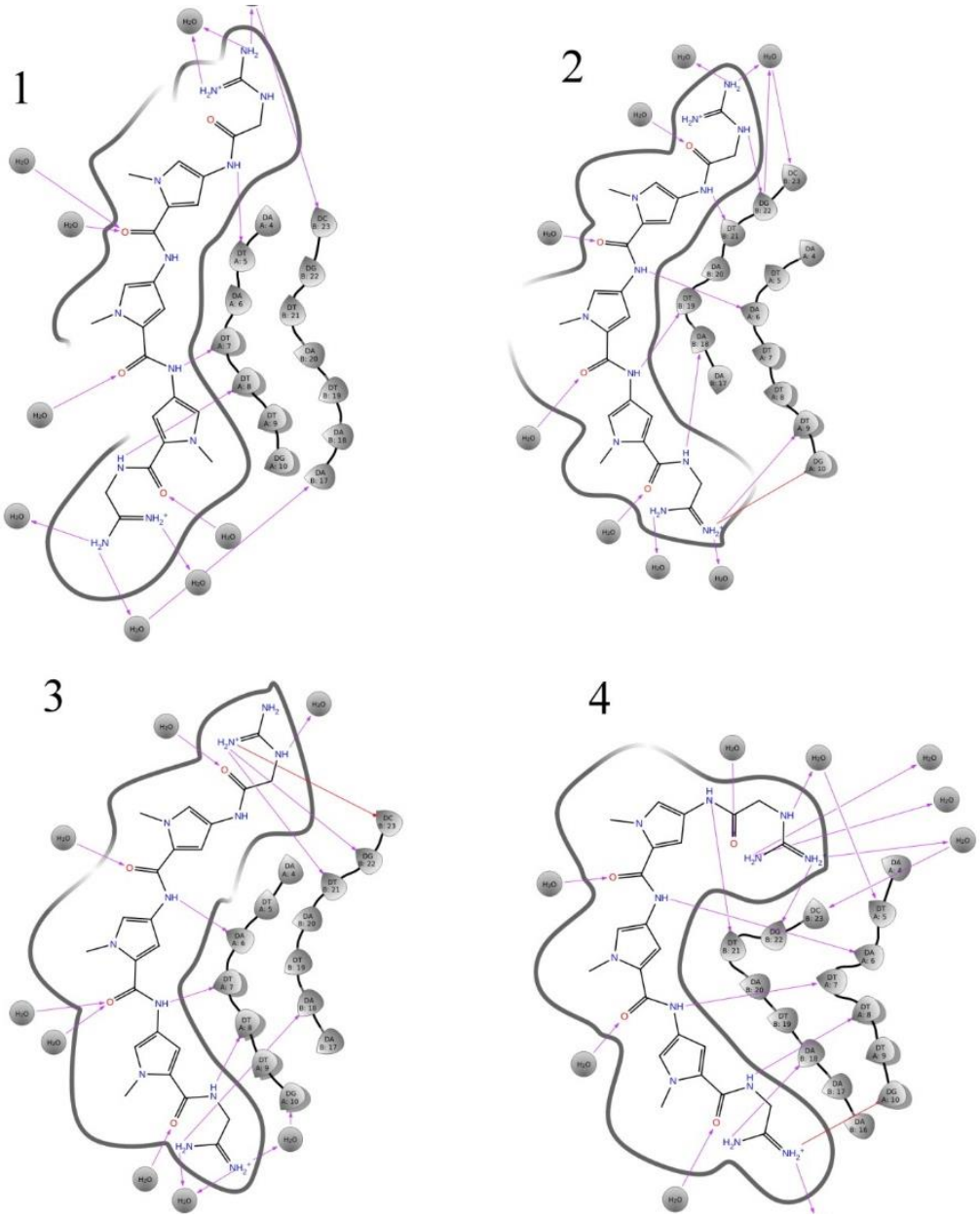


Figure 5.6 Two-dimensional ligand interaction diagrams of the simulation taken at intervals of 100 frames in pictures (1-4) of the duplex 5′d (CGCATATTTGCG)2 and congocidine 2

Water molecules mediate hydrogen bonding to terminal residues in picture 1 (C23, A17), picture 2 (C23), picture 3 (G10), and picture 4 (T5 and C23). Water molecules are also seen interacting with terminal amino residues in pictures (1, 2, 3, and 4) hence the high RMSF in terminal residues. The guanidinium end is involved in hydrogen bonding in pictures 2 (G22), picture 3(G22, T21), and picture 4(G22). The possibility of a pi-cation bond involving the terminal ethanimidamide end and guanidinium end is observed in pictures 2 (G10), picture 3 (C23), and picture 4 (G10), the flexible ethanimidamide end and guanidinium end heavy atoms could be accommodated by a maximum of two GC base pairs.

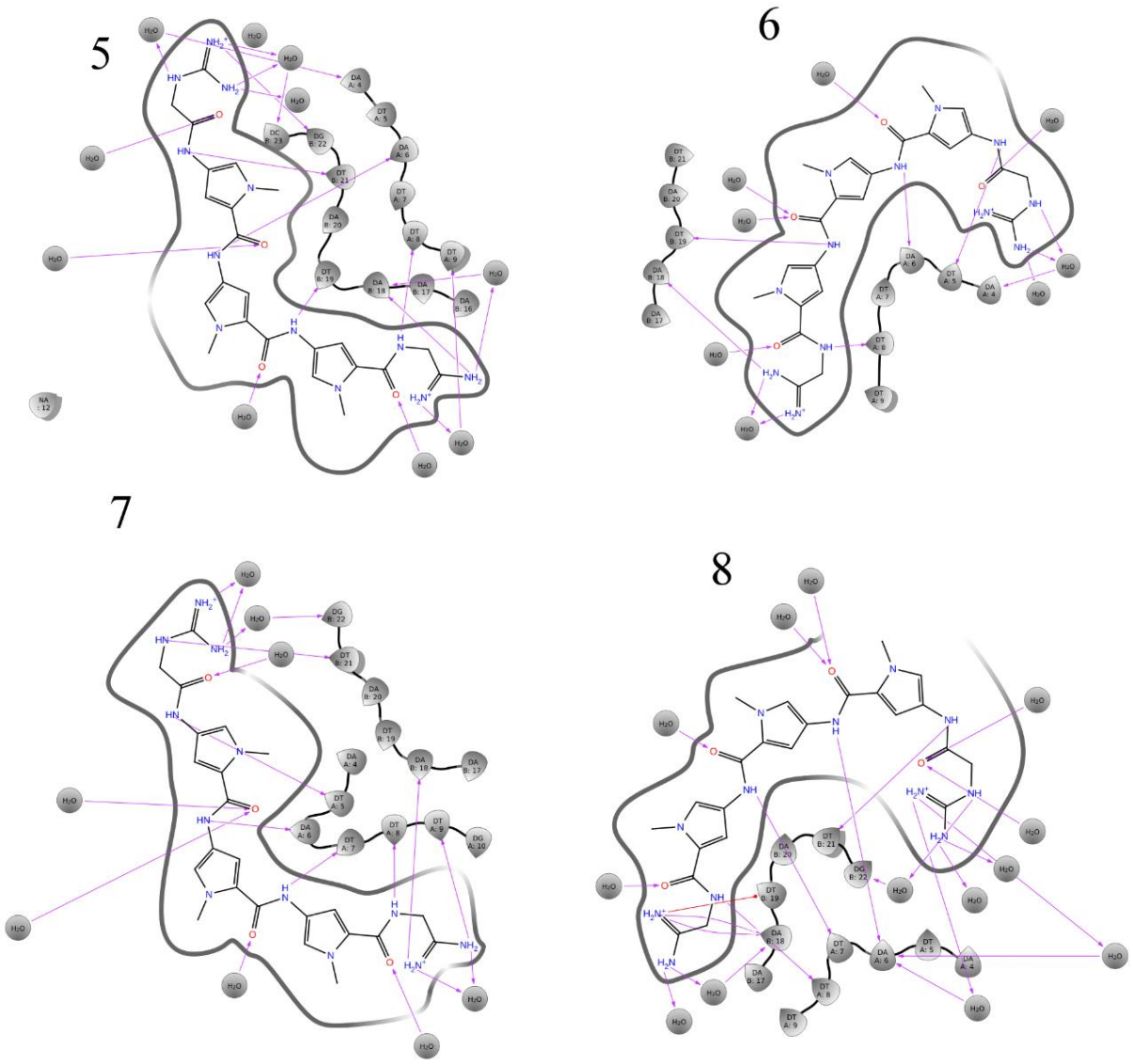


Figure 5.7 Two-dimensional ligand interaction diagrams of the simulation taken at intervals of 100 frames in pictures (5-8) of the duplex 5′d(CGCATATTTGCG)2 and congocidine 2

Water molecules mediate hydrogen bonding to terminal residues in picture 5 (C23, G22, T9, A4, and A18) picture 6 (A4), picture 7 (G22), and picture 8 (A18, A6, and G22). Water molecules are also seen interacting with terminal amino residues in pictures (5, 6, 7, and 8), hence the high RMSF in terminal residues. Hydrogen bonding is visible for acceptor atoms of A and T. No water molecules are observed deep within the minor groove, rather, they appear on the outer surface interacting with oxygen atoms pictures (5, 6, 7, and 8). The possibility of a pi-cation bond from the terminal amino ends is also observed in picture 8 (T19). The flexible terminal heavy atoms of the ligand could be accommodated in the first and second GC base pairs, as observed in picture 5 (C23, G22), picture 7 (G22), and picture 8 (G22 (N3)).

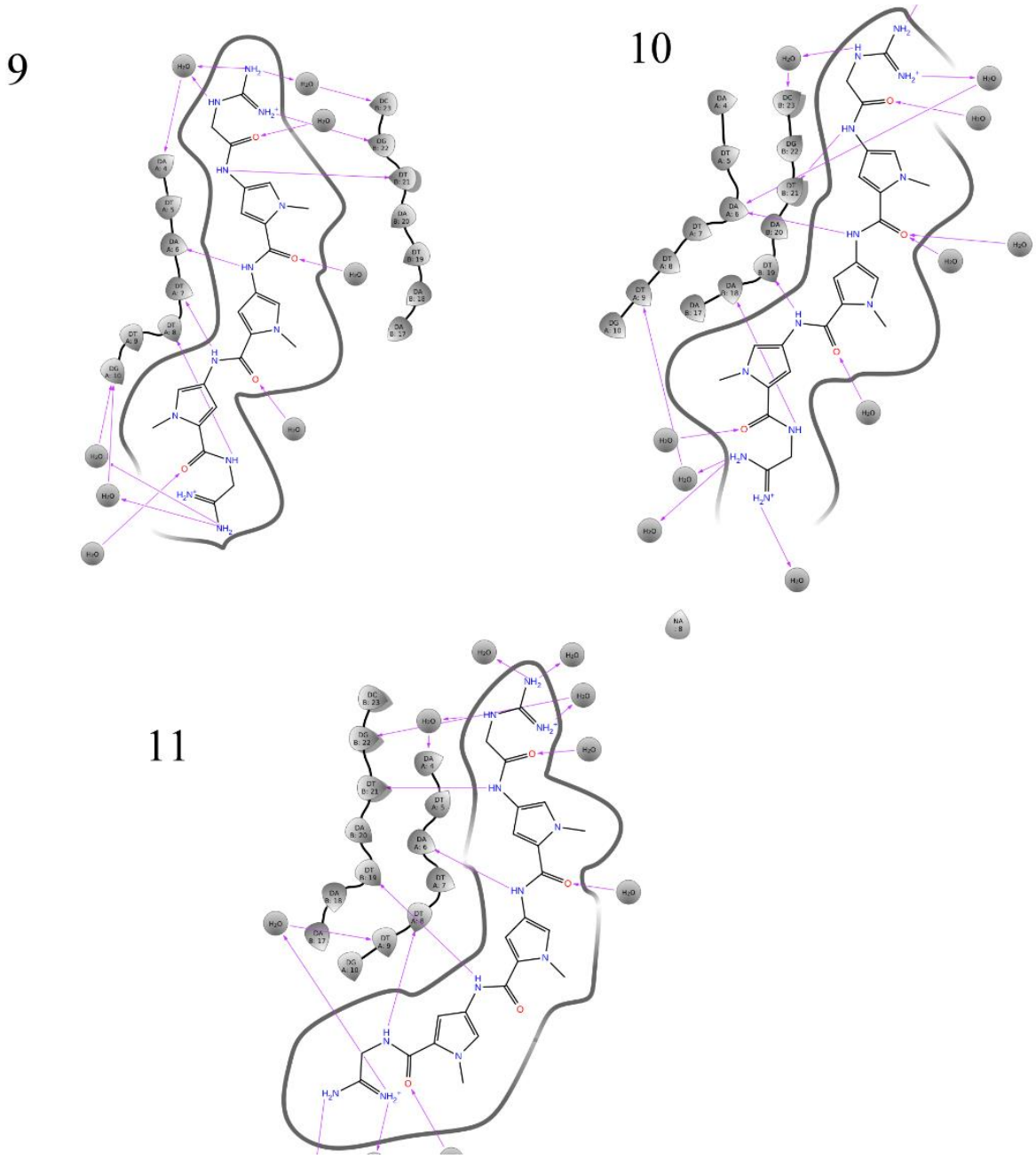


Figure 5.8 Two-dimensional ligand interaction diagrams of the simulation taken at intervals of 100 in pictures (9-11) of the duplex DNA 5′d(CGCATATTTGCG)2 and congocidine 2

Water molecules are involved in hydrogen bond bridging to terminal residues in picture 9 (C23, A4, G10), picture 10 (C23, T9, and G10), and picture 11 (G22, T9, and A6). Water molecules are also seen interacting with terminal heavy atoms in pictures (9, 10, and 11), hence the high RMSF in heavy atoms. The interaction patterns appear to be non-static and seem to adjust to fit the changing conformation, to maximize van der Waal contacts. The flexible guanidinium end and ethanimidamide end heavy atoms could be accommodated by G and C bases in the first and second base pairs in picture 9 (C23, G22, and G10).

5.3.4 Congocidine 3 TATTT Molecular Dynamics Simulation Study Showed Terminal Heavy Atom Fluctuation

In the RMSF study of congocidine 3 with the synthetic DNA construct containing the TATTT motif, the highest RMSF was approximately 1.2 Å in the heavy atom 3 (Figure 5.9). The heavy atoms of the ligand terminals showed high RMSFs, due to an increase in rotational movement space of the solvent-exposed ends of the ligand (Figure 5.9). Rotational freedom and solvent exposure also influenced the heavy atoms (40,21,12,4,30,27,18,1,2,25, and 28). Steric hindrance from the c2-NH_2 of guanine to the guanidinium amino groups also influenced the RMSF. The heavy atoms (19,15,20,14,9,8,11,10,39,38,36, and 34) of congocidine 3 buried in the less solvent-exposed fragment of the minor had lower RMSF.

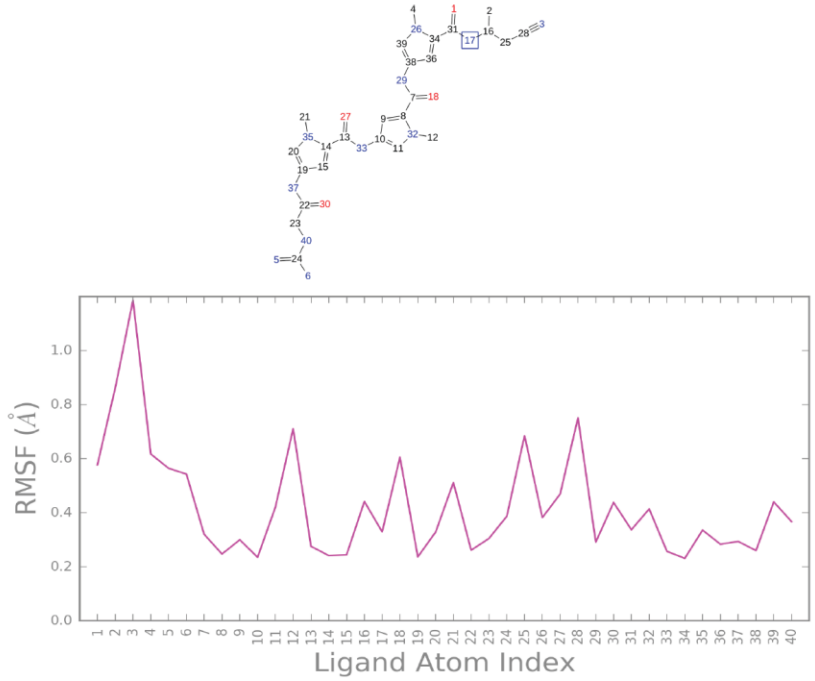


Figure 5.9 Time-dependent RMSF plot in Angstroms (Å) for congocidine 3 showing the heavy atom fluctuations

The fluctuations are broken down by atoms; these are compared to the reference frame as a result of superposition over the time course of the simulation within the minor groove.

Congocidine 3 RMSD stabilized after 0.5 nanoseconds and remained centered around 1.4 Å, in reference to the original conformation at time t=0 (Figure 5.10).

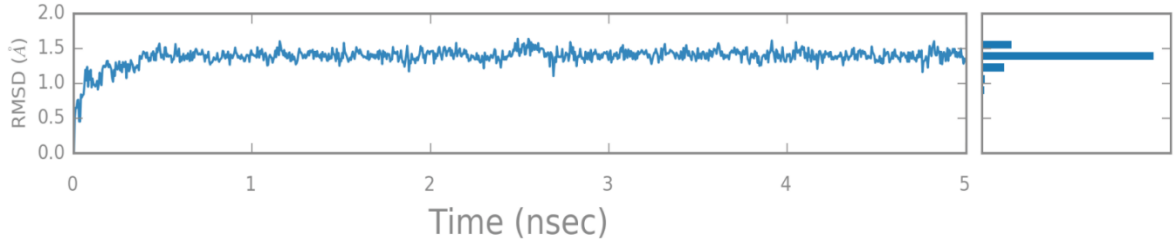


Figure 5.10 RMSD of congocidine 3 structure as a function of time and stability profiles

The plot on the right represents the frequency of distributions relative to the RMSD of the ligand

Most of the congocidine 3 ligand fragments remained buried in the less solvent-exposed part of the minor groove during the 5 ns simulation duration (Figures 5.11, 5.12, and 5.13). The sampled 2D ligand extraction patterns revealed congocidine 3 - DNA interactions were characterized by hydrogen bonding of congocidine 3 amide nitrogen (NH) group to O2 thymine (T) and N3 adenine (A) acceptor atoms.

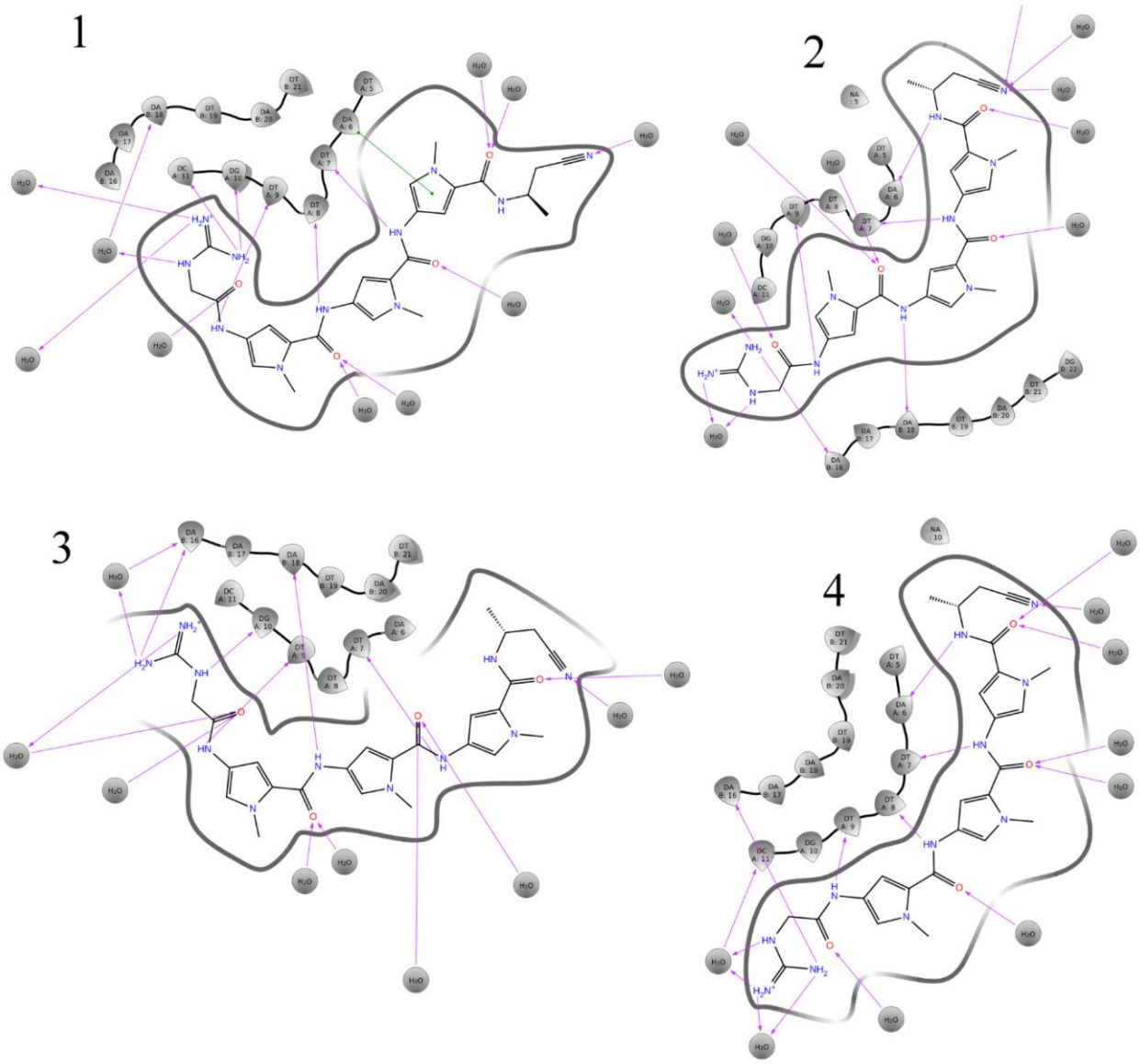


Figure 5.11 Two-dimensional ligand interaction diagrams of the simulation taken at intervals of 100 frames in pictures (1-4) of the duplex DNA 5′d(CGCATATTTGCG)2 and congocidine 3

Water-mediated hydrogen bonding is observed in terminal residues in pictures 1(A18), 3(A16), and 4(C11). Water molecules are also seen interacting with terminal heavy atoms of guanidinium and butanenitrile in pictures (1, 2, 3, and 4), hence the high RMSF in terminal residues. Hydrogen bonding is visible for acceptor atoms of A and T. The interaction patterns appear to be non-static and seem to be adjusting to fit the changing DNA conformation, to maximize contacts. No water molecules are observed deep within the minor groove, rather they appear on the outer surface interacting with oxygen atoms pictures (1, 2, 3, and 4). The flexible terminal heavy atoms could be accommodated by one or two GC base pairs as seen in picture1 (G10, C11), picture 3(G10), and picture 4(C11). Coverage of the TATTT motif is seen in pictures (1, 2, 3, and 4). A pi-pi bond is observed in picture 1.

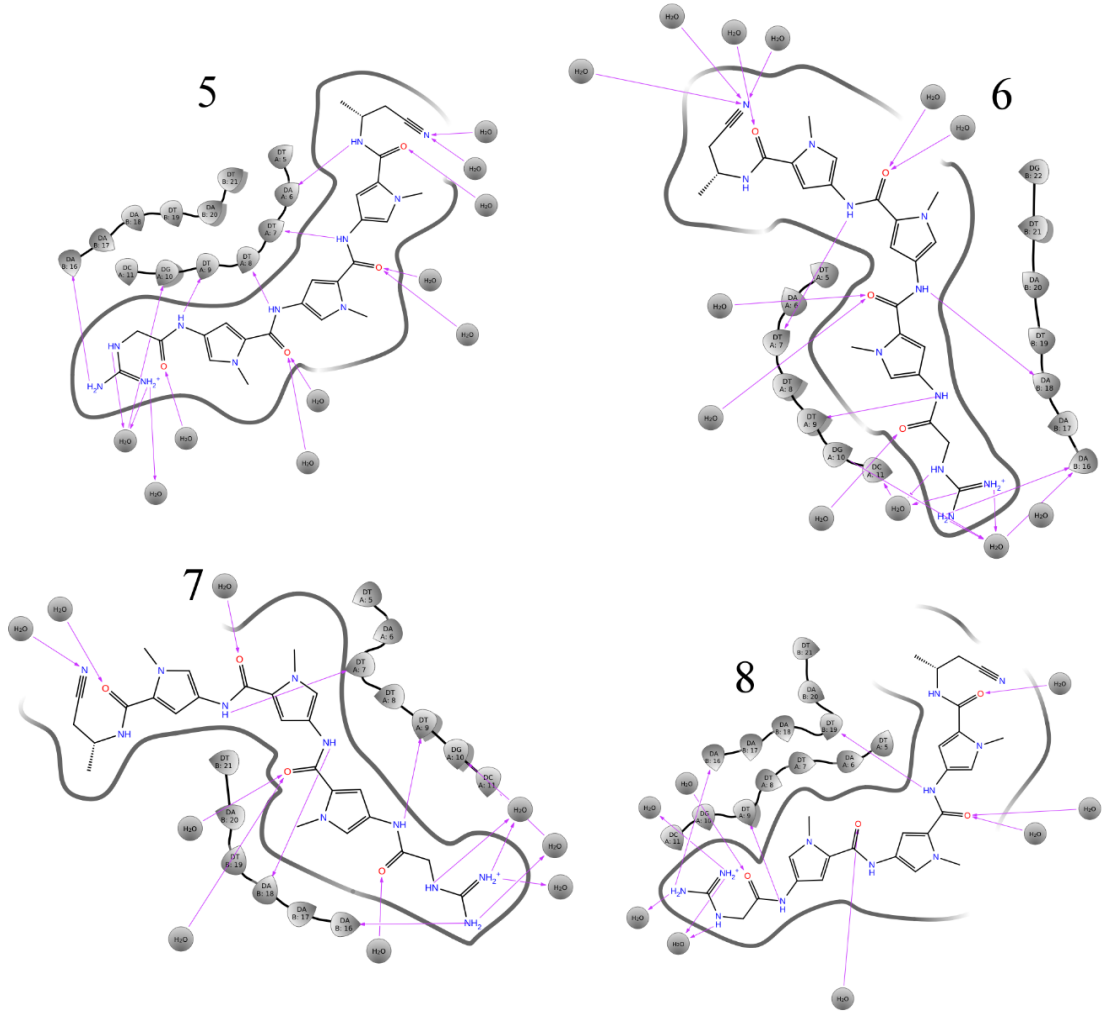


Figure 5.12 Two-dimensional ligand interaction diagrams of the simulation taken at intervals of 100 frames in pictures (5-8) of the duplex 5′d (CGCATATTTGCG)2 and congocidine 3

Water-mediated hydrogen bonding involving ligand terminal heavy atoms to bases is observed in pictures 5 (G10), picture 6 (C11), and picture 7 (G10, C11). Water molecules are also seen interacting directly with terminal heavy atoms of guanidinium and butanenitrile (5, 6, 7, and 8), hence the high RMSF in terminal residues. Hydrogen bonding is visible for acceptor atoms of A and T. No water molecules are observed deep within the minor groove, rather they appear on the outer surface interacting with oxygen atoms pictures (5, 6, 7, and 8). From observations here, the flexible terminal residues of the ligands could be accommodated by G and C bases in the first and second base pairs visible in picture 5 (G10), picture 6 (G10, C11), and picture 7 (G10). Coverage of the TATTT motif is visible in pictures (5, 6, 7, and 8).

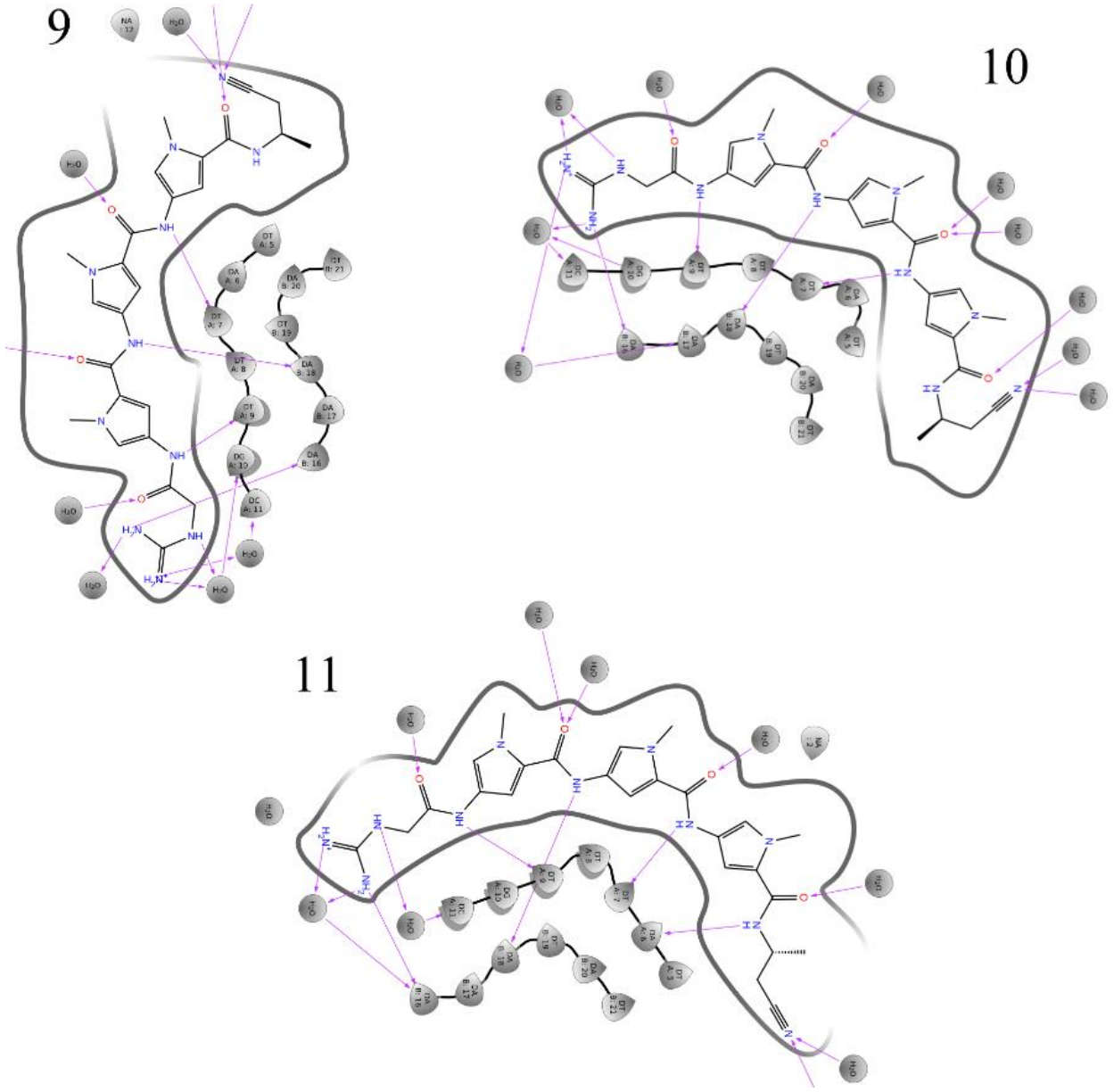


Figure 5.13 Two-dimensional ligand interaction diagrams of the simulation taken at intervals of 100 frames in pictures (9 - 11) of the duplex DNA 5′d(CGCATATTTGCG)2 and congocidine 3

Water-mediated hydrogen bonding involving ligand terminal heavy atoms to bases is observed in pictures 9(G10, C11), 10(A17), and 11(A16). Water molecules are also seen interacting with terminal heavy atoms of butanenitrile and guanidinium in pictures (9, 10, and 11), hence the high RMSF.The flexible terminal residues could be accommodated by one or two GC base pairs visible in picture 9 (G10, C11), picture 10 (G10, C11), and picture 11(C11). Complete coverage of the TATTT is seen in pictures (9, 10, and 11).

5.3.5 Hoechst 33258 TATTT Molecular Dynamics Simulation Study Showed Terminal Heavy Atom Fluctuation

In the RMSF analysis of Hoechst 33258 with the synthetic DNA construct containing the TATTT motif, the highest amount of RMSF of roughly 1.3 Å was observed in the heavy atoms 12, 13, 19, and 20. The free single bonds around heavy atoms 27 and 28 of the piperazine rings conferred some rotational freedom of movement of the heavy atoms 12, 13, 19, and 20 (Figure 5.14). The heavy atoms on the buried part of the minor groove had less space to move around and the double bonds further minimized the rotational freedom of movement. Thus, the lower RMSF in these regions of the ligand fragment.

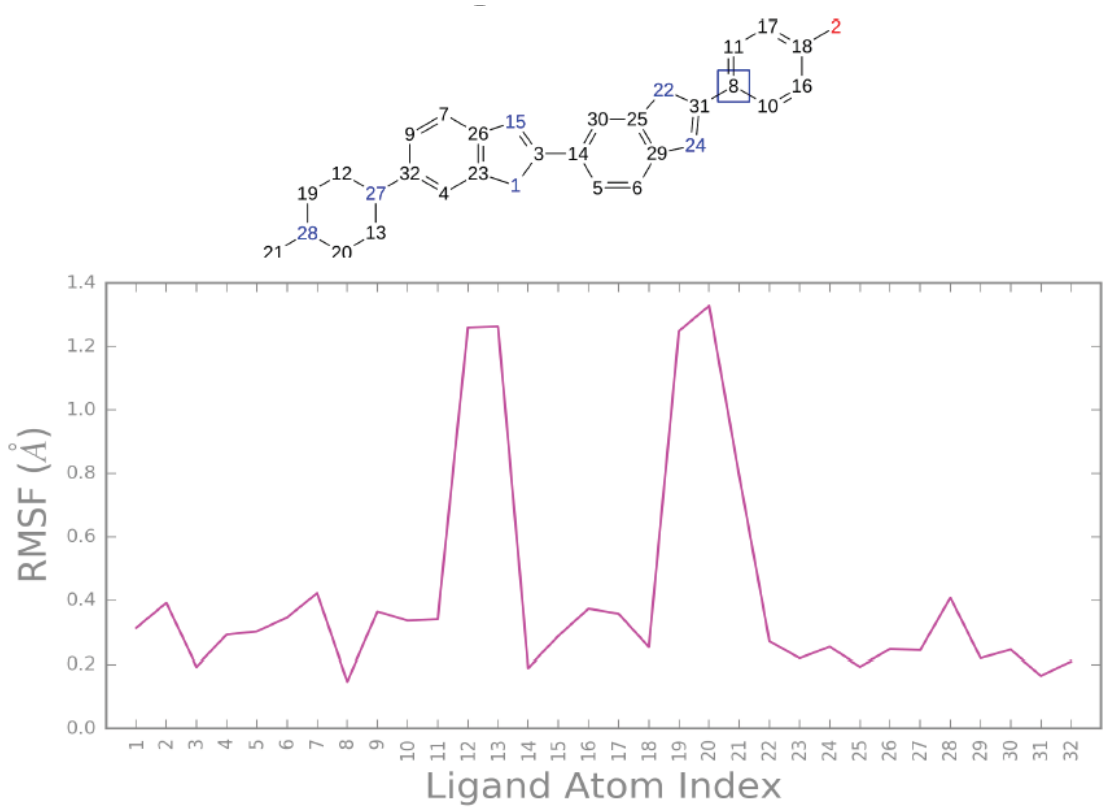


Figure 5.14 Time-dependent L-RMSF plot of Hoechst 33258 showing the heavy atom fluctuations

The fluctuations are broken down by atoms; these are compared to the reference frame as a result of superposition over the time course of the simulation within the minor groove.

Hoechst 33258 L-RMSD was centered at 0.4 Å, in comparison to the starting conformation at time t=0 (Figure 5.15).

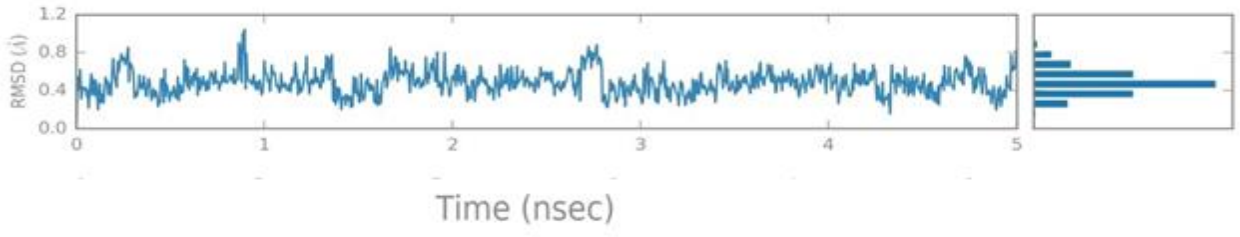


Figure 5.15 RMSD of Hoechst 33258 as a function of time and stability profiles during the MD simulation

The plot on the right represents the frequency of distribution relative to the RMSD of the ligand

The sampled two-dimensional bonding patterns during the 5 ns simulation duration showed that Hoechst 33258-DNA interactions were characterized by continuous hydrogen bonding of the ligand (NH) to acceptor atoms of O2 thymine (T) and N3 adenine (A) in the dynamic environment. Water molecules mediated hydrogen bonds were formed between the ligand terminal heavy atoms and bases at the bottom of the minor groove (Figures 5.16, 5.17, and 5.18).

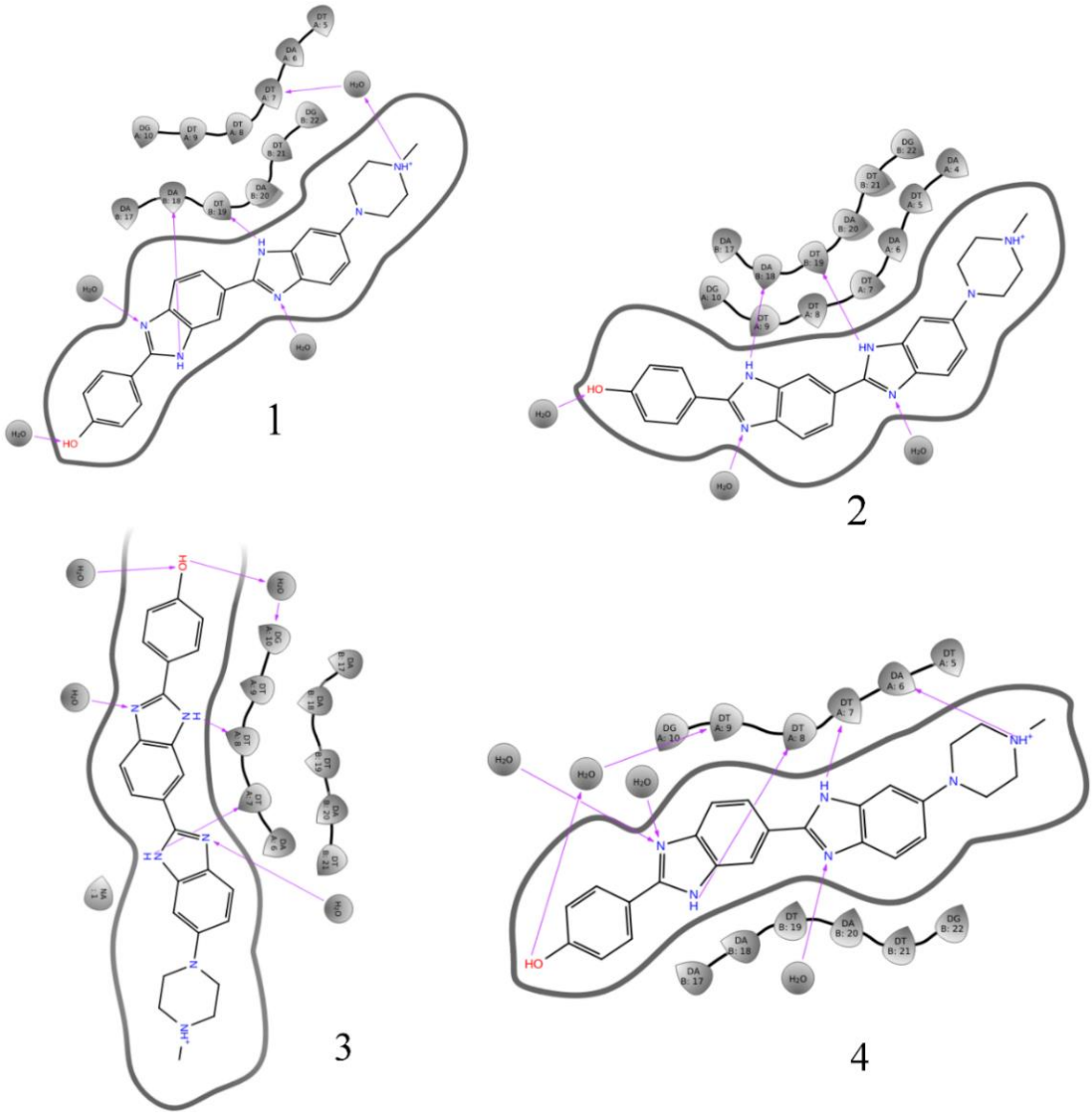


Figure 5.16 Two-dimensional ligand interaction diagram of the simulation taken at intervals of 100 frames in pictures (1-4) of the duplex DNA 5′d(CGCATATTTGCG)2 and Hoechst 33258

Water molecules interact with the terminal hydroxyl of Hoechst 33258 in pictures (1, 2, 3, and 4). Water molecules are also involved in hydrogen bonding with NH^+ of the piperazine ring and (T7) in pictures 1 and 3. The hydroxyl (OH) group is involved in hydrogen bonding in picture 3 (G10) and picture 4 (T9). Hydrogen bonding is visible to acceptor atoms of A and T in the dynamic environment. No water molecules are observed deep within the minor groove, rather they appear on the outer surface interacting with nitrogen atoms of the benzimidazole rings (pictures 1, 2, 3, and 4).

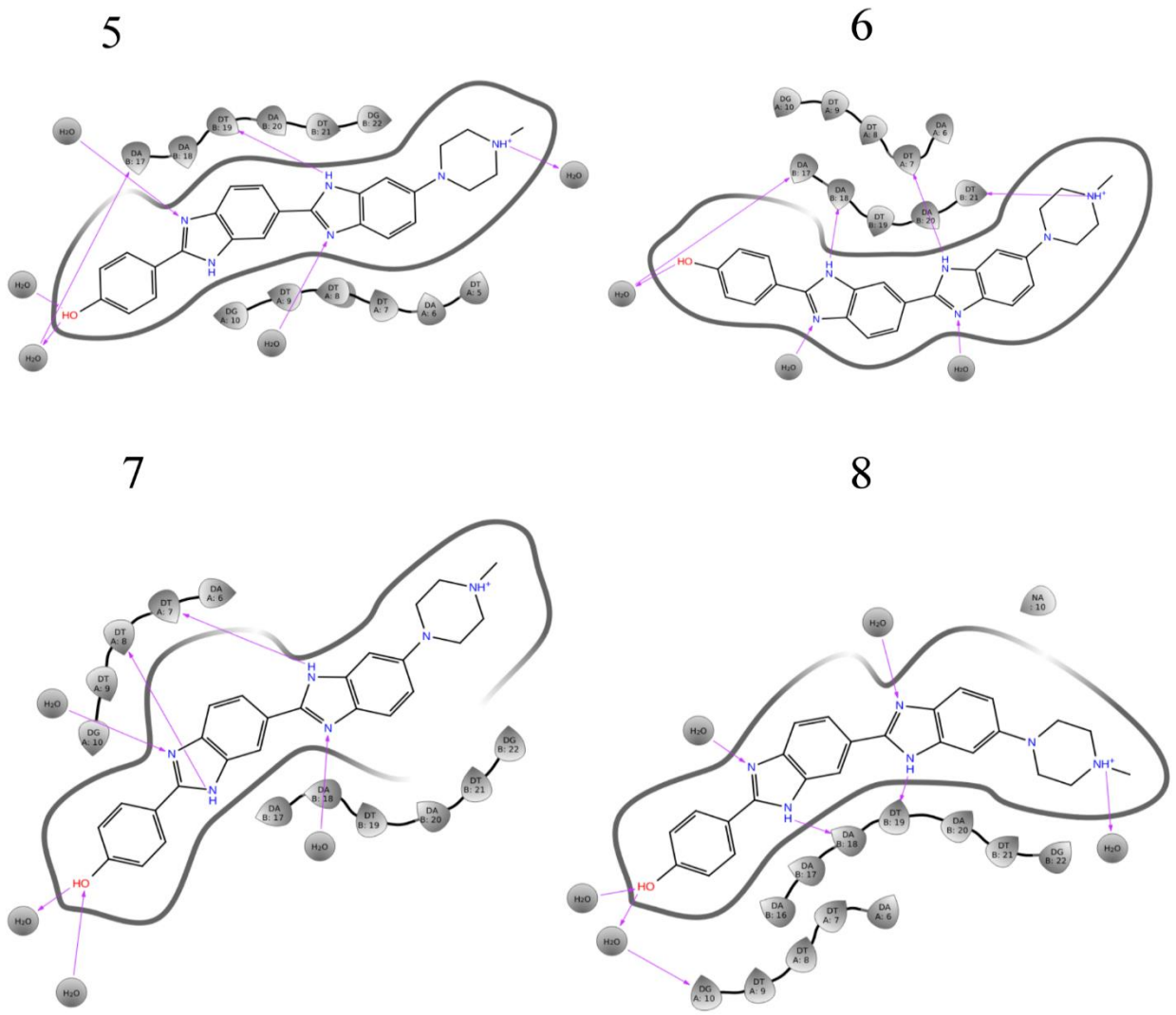


Figure 5.17 Two-dimensional ligand interaction diagrams of the simulation taken at intervals of 100 frames in pictures (5-8) of the duplex DNA 5′d(CGCATATTTGCG)2 and Hoechst 33258

Water molecules are involved in mediating hydrogen bonds between ligand terminal heavy atoms and DNA bases in pictures 5(A17), 6(A17), and 8(G10). Hydrogen bonding is visible to A and T bases. The interaction patterns are non-static and seem to be adjusting to fit the changing DNA conformation to maximize contacts. No water molecules are observed deep within the minor groove, rather they appear on the outer surface interacting with nitrogen atoms attached to the benzimidazole rings (pictures 5, 6, 7, and 8). The flexible terminal residues could be accommodated by a G-C base pair (G10 (N3)).

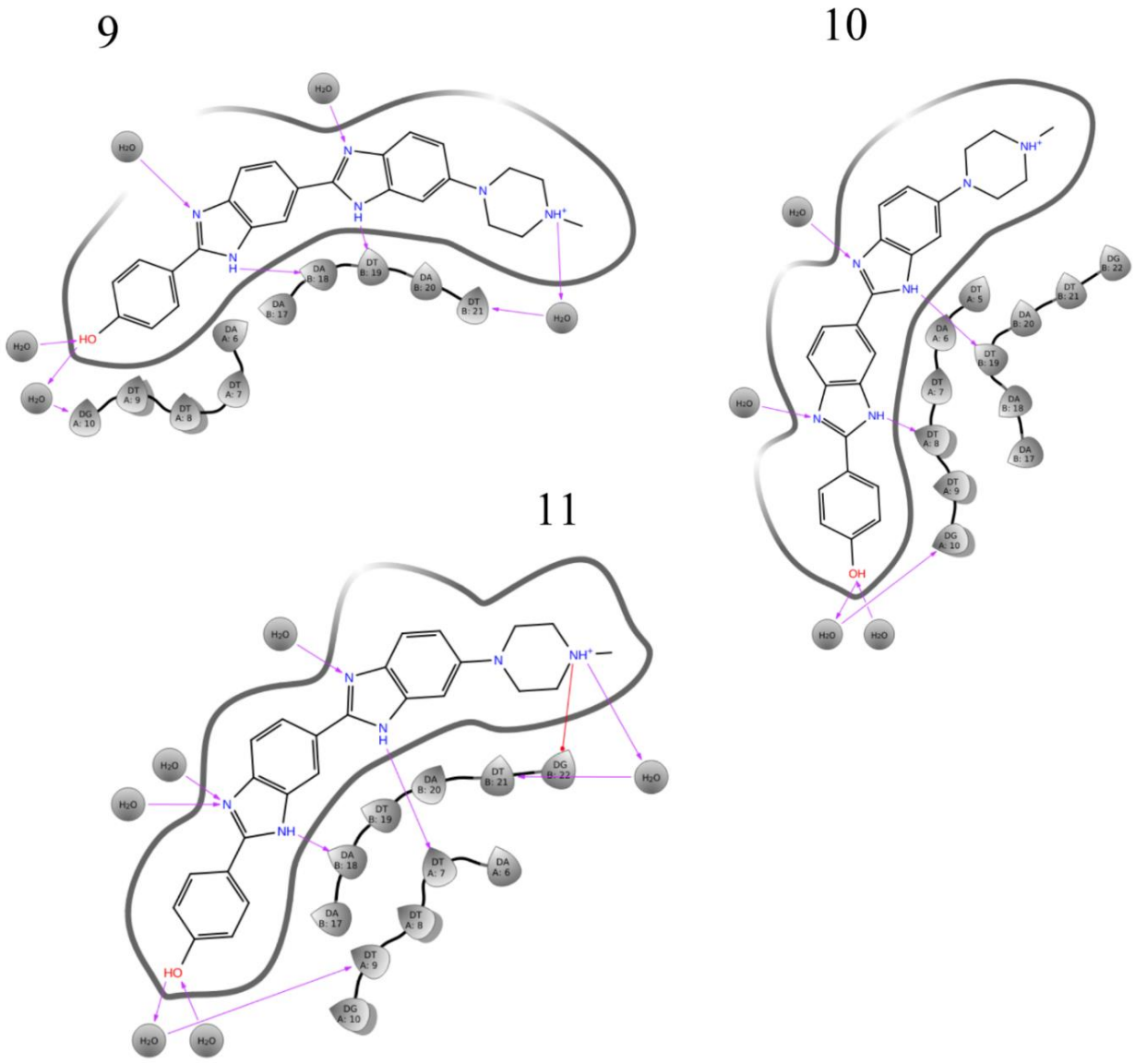


Figure 5.18 Two-dimensional ligand interaction diagrams of the simulation taken at intervals of 100 frames in pictures (9 -11) of the duplex DNA 5′d(CGCATATTTGCG)2 and Hoechst 33258

Water molecules are involved in bridging terminal residues through hydrogen bond formation in picture 9(T21, G10), picture 10(G10), and picture 11(T9, T21). Water molecules are also seen interacting with nitrogen atoms of benzimidazole in pictures (9, 10, and 11). Hydrogen bonding is visible for acceptor atoms of A and T. The interaction patterns are non-static and seem to be adjusting to fit the changing DNA conformation. No water molecules are observed deep within the minor groove. A pi-cationic bond is observed involving G22 and NH^+ of the piperazine ring in picture 11. The flexible terminal residues could be accommodated by a G-C base pair (G10 (N3)).

5.3.6 Analysis of the Potential of the ATAT Motif

The 5′d(CGCGATATCGCG)2 was chosen to verify if the flanking CGCG may significantly affect the docking of ligands and binding affinity in the minor groove and if there was a general preference for the AT motifs. The DNA duplex extracted from the 1VTJ has an ATAT motif that is flanked by two CGCG 4-mer motifs in a 2:1 ratio (Coll *et al*., 1989).

5.3.7 Congocidine 1 Congocidine 2 and Tris-benzimidazole can Dock to the 5′d (CGCGATATCGCG)2 Duplex DNA

For the docked simulation test set, congocidine 2 had a score range of −58.14 to −62.01 for the top 3 sampled conformer poses (Table 5.3). For the highest-scoring docking simulation (Fig 5.19), the flexible terminal guanidinium regions were observed to be capable of forming hydrogen bonds with the O4 atoms of ribose, (T18, 1.98 Å, and A19, 2.11 Å) and acceptor O2 of (T8 1.96 Å). Hydrogen bonds could also be formed between amide groups and acceptor N3 of adenine (A7, 2.37 Å).

Table 5.3 ICM top 3 docking parameters to 5′d(CGCGATATCGCG)2

Name	Score[1]	Hbond[2]	Hphob[3]	VwInt[4]	Eintl[5]	Dsolv[6]	SolEl[7]
Congocidine redock	−33.65	−8.10	−6.10	−47.68	6.01	27.50	7.10
Congocidine 2	−62.01	−14.75	−7.24	−58.88	16.37	37.36	12.18
Congocidine 2	−58.19	−12.58	−7.56	−57.99	15.43	36.89	12.54
Congocidine 2	−58.14	−12.79	−7.51	−59.12	15.86	38.59	9.72
Congocidine 3	−53.51	−10.59	−8.31	−57.34	13.73	34.31	11.29
Congocidine 3	−52.85	−11.27	−8.01	−55.06	15.67	34.15	11.33
Congocidine 3	−51.71	−10.59	−8.51	−57.14	17.23	34.93	13.23
Tris-benzimidazole	−57.47	−9.75	−8.84	−59.73	4.05	33.30	11.94
Tris-benzimidazole	−51.34	−9.21	−8.75	−56.11	4.58	32.50	14.47
Tris-benzimidazole	−50.50	−6.74	−8.11	−51.42	5.61	26.80	5.03

Note: [1]Score is the ICM score (−32 and lower are generally considered good scores); [2]Hbond is hydrogen bond energy (lower values are better); [3]Hphob is the hydrophobic energy of the surface exposed to water (lower values are better);[4]VwInt is the van der Waals interaction energy (lower values are better);[5]Eintl is the internal conformation energy of the ligand (lower values are better);[6]Dsolv is the desolvation of exposed H-bond donors and acceptors (lower values are better);[7]SolEl is the solvation electrostatic energy change upon binding (lower values are better).

By visual examination, the O4 atoms of the deoxyribose ring pointed to the middle of the minor groove (Figure 5.19). Additionally, the placement of the sugar O4 atoms was observed to be regularly spaced along the two sides of the walls of the minor groove, enabling them to possibly interact with the pi-electron clouds of the aromatic N-methylpyrrole rings — a factor which may also add to the stability. Avoidance of the amino group of G4 was observed and the flexible ethanimidamide was seen to present itself to the acceptor oxygen of the cytosine side (C3, 2.06 Å) and hydrogen bond to the O4 ribose of G4 (2.17 Å) in water absence (Figure 5.19). This observation suggests that the flanking of CGCG sequences influenced ligand binding due to the protruding c2-NH_2. Additionally, the simulation runs revealed that for the test ligands of congocidine congeners their flexibility was advantageous in avoiding the protruding guanine

amino group in C3-G22 and G4-C23 alternating base pairs. From the simulation run, the placement of an N-methylpyrrole ring above the protruding amino group prevented a severe intermolecular collision with the guanine amino group G4 of the first G4-C23 base pair as compared to having a guanidinium or ethanimidamide group placed near the guanine amino group. The ethanimidamide amino group of congocidine 2 can thus present itself favorably on the second base pair (C3-G22) of the acceptor O2 of cytosine 3 (Figure 5.19), possibly explaining the favorable docking score observed for this simulation. An additional third CG base pair would probably confer more clashes and a worst-case scenario would be seen in an alternating case of more than two GC base pairs that may become increasingly severe in permitting the drug molecule to buckle away from the floor of the minor groove.

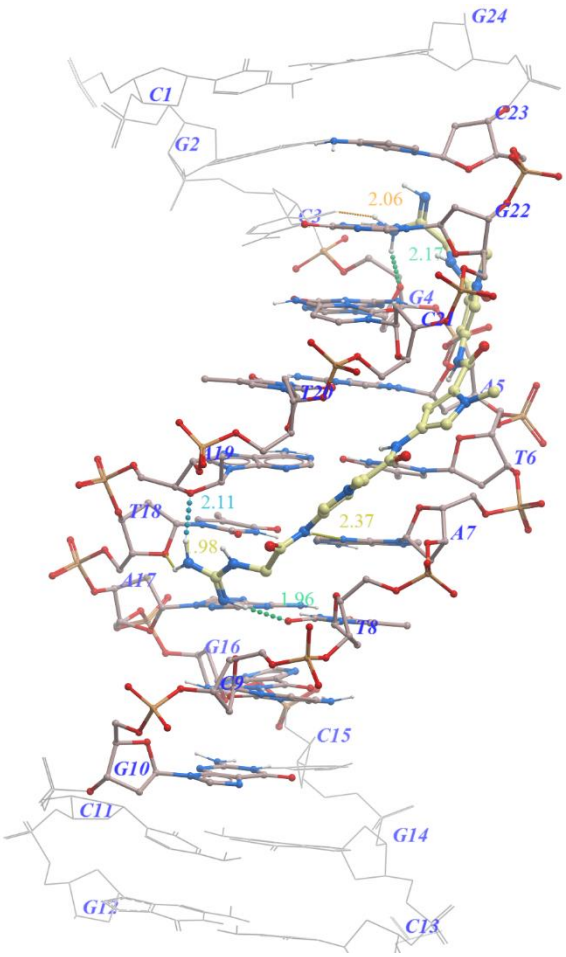


Figure 5.19 A stereoscopic simulated skeletal diagram of the DNA-congocidine 2 complex docked to the minor groove 5′d(CGCGATATCGCG)2

Congocidine 2 forms hydrogen bonds O4 T18 (NH_2), O4 A19 (NH_2), O2T8, N3T8, and O2C3 represented in dotted spheres.

The score range observed for congocidine 3 was between −51.71 to −53.51 (Table 5.3); however, the flexible terminal butanenitrile end was observed to avoid the guanine amino group in the first and second base pairs of the C3-G22 and G4-C21 (Figure 5.20). Besides, the terminal guanidinium could flexibly bend to form hydrogen bonds with the electronegative oxygen atoms of the O4 ribose (T18 2.35 Å and C9 2.2 Å) and acceptor oxygen (2.26 Å T18) (Figure 5.20) in water

absence. The three N-methyl pyrrole rings and donor NH groups were observed to cover the ATAT region during the simulation. Steric collision avoidance of amino groups of G4 and G22 was also observed and significant S-scores were observed during docking.

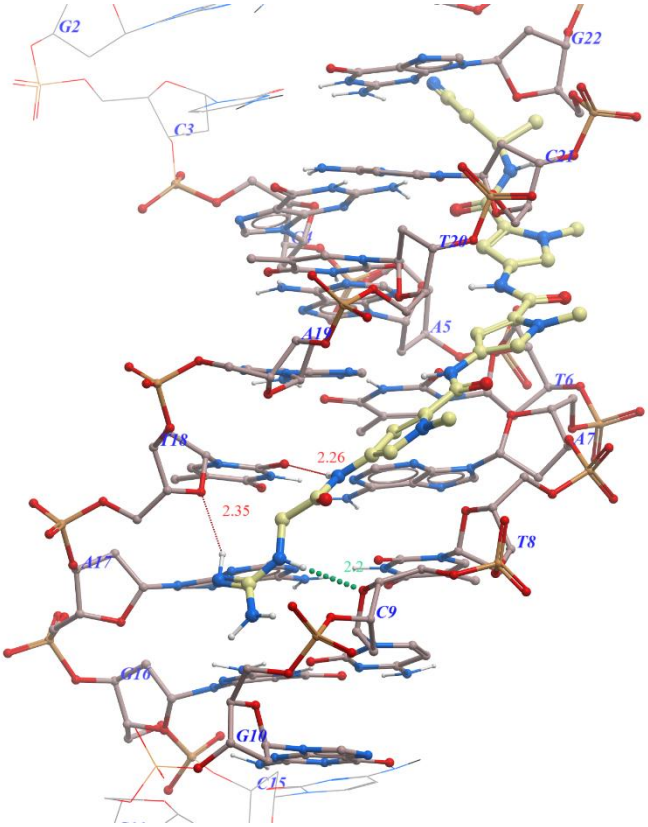


Figure 5.20 Stereo diagram rendered with the ball and stick conformation showing the DNA dodecamer 5′d(CGCGATATCGCG)2 docked to the congocidine 3

Congocidine 3 forms hydrogen bonds O4T18 (NH), O4C9 (NH), and O2T18 represented in dotted spheres.

Tris-benzimidazole top three sampled conformers had a score range of −50.50 to −57.47 (Table 5.3). Complete coverage of the ATAT region was feasible in such a way that the three benzimidazole units could cover the ATAT region. From the first pose, the terminal 3-amino-1-pyrrolidinyl group was observed to extend by two GC base pairs (Figure 5.21). The 3 amino-1-

pyrrolidinyl group presented itself first on the guanine side C9 to avoid the G16 amino group as it flexibly hydrogen-bonded with the G10 ribose O4 oxygen (1.85 Å) and then to the N3 acceptor atom of G10 (2.22 Å), thus avoiding the G10 protruding amino group. Additionally, for tris-benzimidazole docked to the central ATAT, the inward-facing NH groups of the benzimidazole subunits of tris-benzimidazole were involved in h-bonding with acceptor atoms of O2 thymine (T) and N3 adenine (A) bases of the duplex 5′d(CGCGATATCGCG)2.

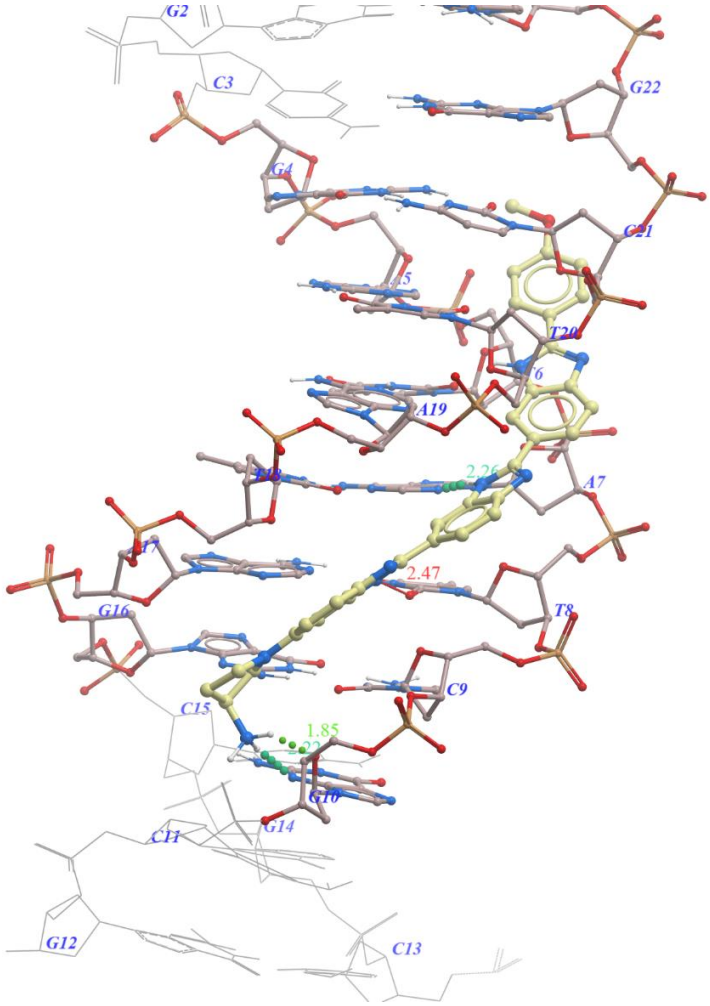


Figure 5.21 Simulated stereo diagram rendered in the ball and stick conformation showing tris-benzimidazole complex with the hydrogen bonds in spheres

Tris-benzimidazole forms hydrogen bonds N3G10, O4G10 (NH), O2T8, and N3A7 represented in dotted spheres.

5.3.8 Tris-benzimidazole and Congocidine Congeners Could Significantly Dock to 5′d (CCAATAATCG)2 Duplex DNA

The 4-mer motif ATAA is present in late promoters of the African swine fever virus genome (Appendix VI). The study, therefore, validated if congocidine 3, congocidine 2, and tris-benzimidazole could significantly dock to a decamer 5′d(CCA[ATAA]TCG)2 having the ATAA motif embedded within it. The PDB structure used in this study was 5F9I, from which the DNA duplex 5′d(CCAATAATCG)2 was extracted. From the simulations, the top three docked complexes were found to have significant S-score ranges of −59.27 to −68.67 for congocidine 2, −59.49 to −66.68 for congocidine 3, and −53.15 to −61.15 for tris-benzimidazole (Table 5.4).

Table 5.4 Top three docking parameters to 5′d(CCAATAATCG)2 for ICM.

Name	Score[1]	Hbond[2]	Hphob[3]	VwInt[4]	Eintl[5]	Dsolv[6]	SolEl[7]
Congocidine 2	−68.67	−16.41	−7.57	−53.71	14.35	34.89	3.42
Congocidine 2	−59.58	−11.67	−7.54	−52.30	10.70	29.90	2.83
Congocidine 2	−59.27	−12.83	−7.76	−55.08	21.12	34.63	6.81
Congocidine 3	−66.68	−14.51	−8.75	−53.93	17.06	32.41	3.81
Congocidine 3	−63.41	−13.16	−8.00	−51.01	10.88	29.17	2.10
Congocidine 3	−59.49	−12.27	−8.75	−53.86	16.96	33.22	5.08
Tris-benzimidazole	−61.51	−12.07	−9.68	−50.36	7.76	31.34	3.63
Tris-benzimidazole	−59.10	−9.72	−9.67	−52.92	5.28	27.96	5.96
Tris-benzimidazole	−53.15	−12.64	−9.76	−43.71	9.67	30.47	9.47

Note: [1]Score is the ICM score (−32 and lower are generally considered good scores); [2] Hbond is hydrogen bond energy (lower values are better);[3]Hphob is the hydrophobic energy of the surface exposed to water (lower values are better);[4]VwInt is the van der Waals interaction energy (lower values are better);[5]Eintl is the internal conformation energy of the ligand (lower values are better);[6]Dsolv is the desolvation of exposed h-bond donors and acceptors (lower values are better); [7]SolEl is the solvation electrostatic energy change upon binding (lower values are better).

5.3.9 Congocidine 2 Interaction Pattern with the AATAAT Motif

For congocidine 2, (Figure 5.22) typical hydrogen bonds were formed with acceptor atoms of O2 T (5, 2.23 Å) and T (38, 1.84 Å), of the duplex 5′d(GCAATAATCC)2, from the sampled top conformers. The docked poses of the ligand were oriented in such a way that there was a possibility of all three N-methylpyrrole rings and associated amides (NH) fitting in the AATAAT region of the duplex 5′d(GCAATAATCC)2 minor groove. Besides spanning the ATAA motif, the flexible amino-terminal guanidinium region was capable of forming hydrogen bonds with the ribose O4 atoms (1.77 Å) of T34. From the semi-flexible docking simulations, the guanidinium amino end was observed to avoid the protruding amino group of G 32 (Figure 5.22). A fit in the AATAAT region was good because there was no steric interference from the protruding amino group of guanine hence a deeper penetration and probably good van der Waal contacts within the minor groove.

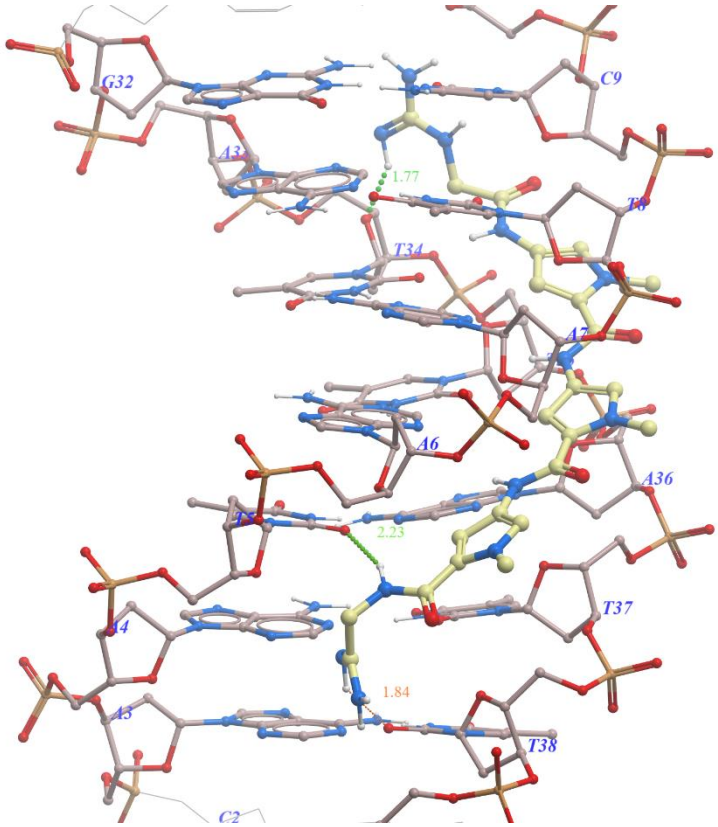


Figure 5.22 A simulated interaction of congocidine 2 docked to the minor groove duplex 5′d(CCAATAATCG)2

Congocidine 2 forms hydrogen bond interactions with O2T38, O2T5, and O4T34 (NH).

5.3.10 Congocidine 3 Interaction Pattern with the AATAAT Motif

For congocidine 3, all three N-methylpyrrole rings and associated amides fitted in the AATAAT region of the duplex 5′d(CCAATAATCG)2 minor groove (Figure 5.23). The flexible terminal guanidinium region was seen forming hydrogen bonds (T34, 1.77 Å) with the O4 atoms of the ribose lining the minor groove . The ligand guanidinium amino group end avoided the protruding amino group of the guanine (G32). Characteristic hydrogen bonding was also observed between the amide nitrogen NH to O2 thymine (T) T37 (2.06 Å) and N3 adenine (A) A36 (2.49 Å), from the docked poses of congocidine 3 to the central 5′ AATAAT motif in 5′d(CCAATAATCG)2.

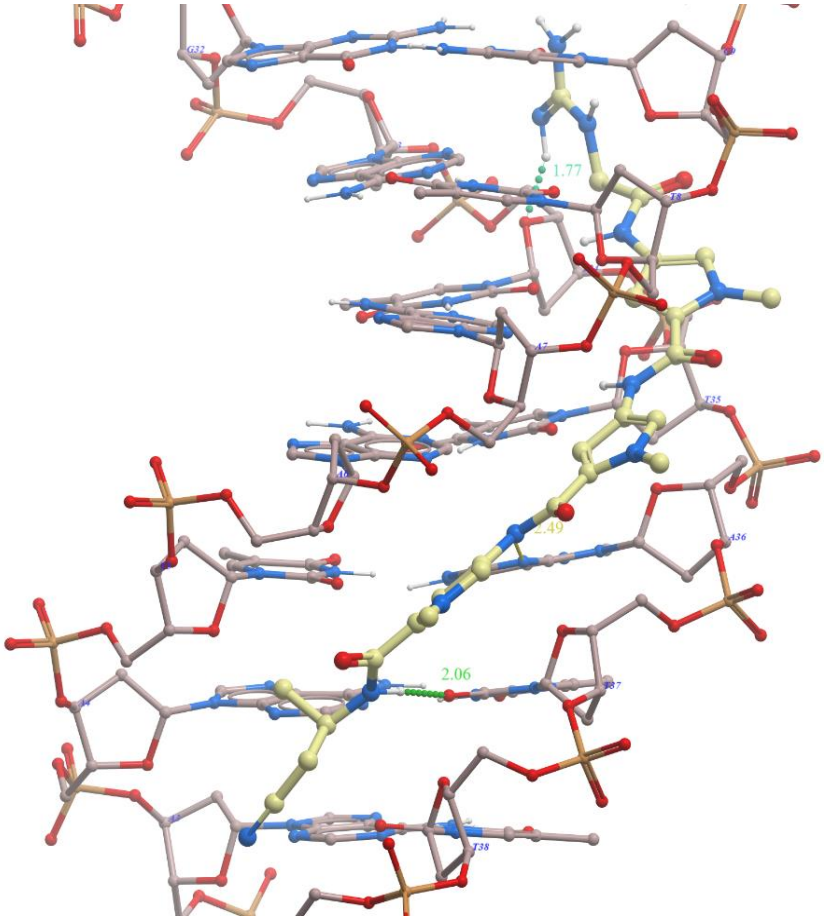


Figure 5.23 Simulated stereo diagram of a hypothetical high-scoring docking pose of congocidine 3, docked to the minor groove of 5′d(CCAATAATCG)2

Congocidine 3 forms hydrogen bonds with O2T37, N3A36, and O4T34(NH) represented in dotted spheres.

5.3.11 Tris-Benzimidazole Interaction Pattern with the AATAAT Motif

For tris-benzimidazole, complete coverage of all the benzimidazole rings fitting in the AATAAT region of the duplex 5′d(CCAATAATCG)2 was observed from multiple docking simulations. Additionally, the 3-amino-1-pyrrolidinyl group was located in the space between T38 and G39 (Figure 5.24). Direct avoidance of the c2 -NH2 amino group of (G39) was observed by the 3-amino-1-pyrrolidinyl group, resulting in hydrogen bonding between O4 (G40 2.16 Å) and guanine

N3 (G39 1.88 Å). Hydrogen bonding was observed between the amide nitrogen NH to N3 adenine (A), A36 (2.05 Å), and O2 thymine (T37, 2.16Å) and (T35 1.99Å). Moreover, for tris-benzimidazole, displacement by one base pair was observed, such that the 3 amino-1-pyrrolidinyl ring was positioned in the segment in which the minor groove widened out because of the presence of the first G-C base pair. The methoxyphenyl end was observed to cover the A7 and T8 regions of the duplex.

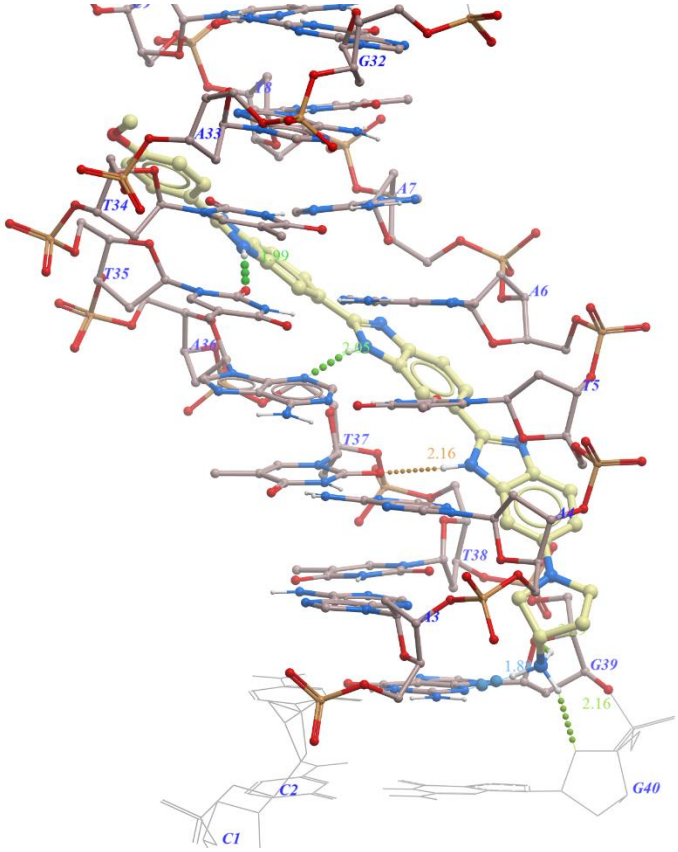


Figure 5.24 Simulated stereo interaction diagram showing tris-benzimidazole molecule bonded to the minor groove of 5′d(CCAATAATCG)2 duplex

Tris-benzimidazole forms hydrogen bonds with O2T37, N3A36, O2T35, N3G39, and O4 (NH) G40.

The simulations inferred from the docked poses of congocidine 2, congocidine 3, and tris-benzimidazole revealed coverage of the embedded ATAA containing hexamer region (A[ATAA]T) was viable for the synthetic DNA sequence 5′d (CCAATAATCG)2.

5.3.12 Free Energy Calculations of Test MGBs to the Duplex 5′d(CGCGATATCGCG)2

Earlier, during biased Monte-Carlo docking studies, the effect of flanking GC base pairs was generally observed to lower the docking score. Compared to the parent compound congocidine, the S-score was observed to be higher for the test tris-benzimidazole and congocidine congeners (where a lower score of −32, implies greater significance). From the test set binding affinity calculations ($\Delta G\ ^{\circ}_{pred}$), the test ligand-DNA was also an improvement on the parent compound congocidine (Table 5.5). As the van der Waal factor improved (where more –ve denotes improved scores) the binding affinity ($\Delta G\ ^{\circ}_{pred}$) increased. Additionally, there was an improvement in the thermal melting temperature for the test ligands, as the thermal melting temperature increased for the test ligands the binding affinity increased thus an improvement in stability. The high binding affinity for the congocidine congeners and the tris-benzimidazole ligands is attributed to better van der Waal interactions from the lengthier test ligands that had additional benzimidazole and N-methylpyrrole rings. Additionally, the flexible test set ligands could be accommodated by one or two GC base pairs during the docking simulation thus improving interactions.

Table 5.5 Free energy calculations for docked complexes to 5′d(CGCGATATCGCG)2

Ligand	$\Delta H°_{el}$[1]	$\Delta H°_{vdw}$[2]	$T\Delta S°_{rt}$[3]	$\Delta G°_{w}$[4]	$\Delta G°_{cbe}$[5]	ΔTm (K)[6]	$\Delta G°_{pred}$[7]
1vtj-Congocidine	−2.9	−28.7	25.5	−12.7	−18.8	14.4	−10.2
Congocidine 2	−6.1	−36.8	26.1	−13.7	−30.5	22.4	−11.7
Congocidine 2	−6.8	−37.0	26.1	−15.8	−33.5	24.4	−12.1
Congocidine 2	−6.0	−37.6	26.1	−14.8	−32.3	23.6	−12.0
Congocidine 3	−6.6	−38.1	26.1	−17.0	−33.7	24.6	−12.4
Congocidine 3	−5.2	−36.8	26.1	−16.4	−32.3	23.6	−11.9
Congocidine 3	−5.3	−38.1	26.2	−17.6	−34.8	25.4	−12.3
Tris-benzimidazole	−3.6	−39.0	26.1	−19.8	−36.2	26.3	−12.4
Tris-benzimidazole	−3.4	−40.6	26.2	−20.1	−37.9	27.5	−12.7
Tris-benzimidazole	−3.7	−36.6	26.0	−18.3	−32.6	23.9	−12.0

Note: [1]$\Delta H°_{el}$: total electrostatics (kcal/mol); [2]$\Delta H°_{vdw}$: total van der Waals (kcal/mol); [3]$T\Delta S°_{rt}$: rotational and translational entropy (kcal/mol); [4]$\Delta G°_{w}$: hydration free energy (kcal/mol); [5]$\Delta G°_{cbe}$: calculated binding energy (kcal/mol); [6]ΔTm: thermal melting temperature kelvins; [7]$\Delta G°_{pred}$: predicted binding affinity (kcal/mol).

5.3.13 Free Energy Calculations of Test Ligands to the Duplex 5′d(CCAATAATCG)2

For the single structure docked complexes of the test reversible MGBs to the duplex 5′d(CCAATAATCG)2, the calculated binding affinities were improved and significantly higher (Table 4.2 and Table 5.6). The binding parameters of the calculated van der Waals forces improved as well as the thermal melting temperature compared to 5′d(GTATATAC)2 and 5′d(CGCGATATCGCG)2. The calculated binding affinities were also relatively higher. Steric interference from the amino group of guanine (c2 -NH_2) was minimal in 5′d(CCAATAATCG)2 compared to 5′d(CGCGATATCGCG)2 hence deeper penetration and good van der Waal contacts (Table 5.6).

Table 5.6 Free energy predictions for docked complexes to 5′d(CCAATAATCG)2

Ligand	$\Delta H°_{el}$[1]	$\Delta H°_{vdw}$[2]	$T\Delta S°_{rt}$[3]	$\Delta G°_{w}$[4]	$\Delta G°_{cbe}$[5]	ΔTm[6]	$\Delta G°_{pred}$[7]
Congocidine 2	−7.4	−43.5	26.1	−16.5	−41.3	29.8	−13.1
Congocidine 2	−8.2	−41.3	26.1	−17.1	−40.6	29.3	−13.0
Congocidine 2	−7.4	−47.0	26.1	−17.4	−46.7	33.4	−13.8
Congocidine 3	−6.3	−45.7	26.1	−19.8	−45.8	32.8	−13.7
Congocidine 3	−6.4	−42.3	26.1	−18.8	−41.4	29.8	−13.1
Congocidine 3	−7.6	−47.6	26.2	−19.9	−49.0	35.1	−14.1
Tris-benzimidazole	−5.4	−39.1	26.1	−22.3	−40.6	29.3	−13.0
Tris-benzimidazole	−5.0	−46.4	26.1	−22.8	−47.7	34.1	−13.9
Tris-benzimidazole	−5.1	−44.1	26.0	−22.8	−45.9	23.9	−13.7

Note:[1]$\Delta H°_{el}$: total electrostatics (kcal/mol);[2]$\Delta H°_{vdw}$: total van der Waals (kcal/mol);[3]$T\Delta S°_{rt}$: rotational and translational entropy (kcal/mol);[4]$\Delta G°_{w}$: hydration free energy (kcal/mol);[5]$\Delta G°_{cbe}$: calculated binding energy(kcal/mol);[6]ΔTm: thermal melting temperature kelvins; [7]$\Delta G°_{pred}$: predicted binding affinity (kcal/mol).

5.3.14 Hoechst 33258 and Tris-benzimidazole had High Liposomal Encapsulation Potential

The LogP calculation by SwissADME revealed low LogP values of −1.53 and −0.67 for congocidine 2 and 3 respectively, implying that these congeners were very soluble (Daina *et al*., 2017) (Table 5.7). The tris-benzimidazole predicted LogP value of 3.45 was high, implying the congener had poor solubility in water. Hoechst 33258 was moderately soluble with a predicted LogP value of 2.56 (Table 5.7). Having a large LogP value has been related to efficacious microencapsulation or the formation of liposomes (Nii & Ishii, 2005). As such, tris-benzimidazole and Hoechst 33258 may be microencapsulated in liposomes and possibly be engineered to target macrophages for specific release to minimize potential toxic effects.

Table 5.7 ADME properties of select minor groove binders

Compounds	Log P	Solubility
Congocidine 1	−1.59	Very Soluble
Congocidine 2	−1.53	Very soluble
Congocidine 3	−0.67	Very soluble
Tris-benzimidazole	3.45	Poorly Soluble
Distamycin	−0.22	Very Soluble
Hoechst 33258	2.56	Moderately soluble

5.4 DISCUSSION

In the docking and simulation experiment to the synthetic DNA construct, significant binding to the TATTT motif analogous to ASFV promoter regions was demonstrated using congocidine congeners and Hoechst 33258. The free energy calculations for the congocidine congeners and Hoechst 33258 to the TATTT motif showed favorable binding affinity. Additionally, from MD simulation studies, congocidine congeners and Hoechst 33258 predominantly span A-T base pairs of TATTT and remained stably embedded in the minor groove. The TATTT motif is a druggable target whose mutation is lethal (García-Escudero & Viñuela, 2000). Therefore, targeting the TATTT motif can be exploited for the potential inhibition of ASFV gene transcription. Several studies have shown NCLDVs have bipartite promoter motifs (García-Escudero & Viñuela, 2000; Oliveira *et al*., 2017, 2018; Rodríguez & Salas, 2013). The late ASFV promoter structure is bipartite composed of late and intermediate promoters which are consistent with the vaccinia virus where the TAAAT motif followed by guanine is the initiator region of the vaccinia virus late promoters, while the intermediate promoters have TAAA upstream of the ORF (Zhilong Yang *et al*., 2012). Inhibiting TBP through promoter motifs by Hoechst 33258 (Yakimovich *et al*., 2017) and distamycin A (Bellorini *et al*., 1995; Broyles *et al*., 2004; Knutson *et al*., 2006) has been shown to effectively stop transcription in the Vaccinia virus. Therefore, similar logic may be inferred in ASFV. For instance, the B646L gene coding the major capsid protein p72 has been shown to have indispensable TATATA and TATTT bipartite motifs as core and initiator regions respectively (García-Escudero & Viñuela, 2000; Rodríguez & Salas, 2013). Deletion/insertion and mutation of the TATTT motif in ASFV are fatal to viral transcription (García-Escudero & Viñuela, 2000). Studies have further shown that copy numbers of TATTT motifs in NCLDVs are

significantly higher than TATATA (Oliveira *et al*., 2018) and thus the TATTT target may be more effective in targeted promoter transcription stagnation. Additionally, the ATATTT motif displayed higher binding affinity and thermal melting temperature compared to TATATA. In retrospect, therefore, the transcription efficiency of genes having this TATTT motif could be targeted through steric interference of transcription factors and melting effectively reducing viral transcription efficiency.

From Monte Carlo-based multiple docking simulation runs involving congocidine 3, congocidine 2, and tris-benzimidazole (Table 5.3) to the ATAT motif of 5′d(CGCGATATCGCG)2, the docking test set S-scores (congocidine 3, congocidine 2 and tris-benzimidazole) were significantly better than the parent compound congocidine, implying the test set were better binders. Congocidine 3, congocidine 2, and tris-benzimidazole showed preferential docking to the central ATAT of the dodecamer sequence, probably due to the minimal steric hindrance within the minor groove. Judging from the docking simulation coverage, no minor groove binders of the test set were observed to preferentially dock to the CGCG motif that flanked the ATAT motif. An ideal binding site of the A/T sequence (AATAAT and ATATTT) without interruptions with G-C base pairs would offer minimal steric hindrance involving the c2-NH_2 group of guanine. Free energy calculations showed AATAAT and ATATTT-dominated complexes were also favorable in allowing appreciable van der Waal contacts that correlated with an increased thermal melting temperature and binding affinity the lengthier ligands had appreciable van der Waal contacts resulting from an extra added benzimidazole and N-methylpyrrole rings along the AT tract of the minor groove in comparison to congocidine and Hoechst 33238. Additionally, bases in a DNA double helix are stacked on top of each other. The stacking interaction is stabilized by electronic

pi-pi interactions of the aromatic rings, this stacking also minimizes the contact of bases to the solvent. Base stacking has favored interactions. The 5′ ApT base-pair step has a good stacking overlap while the 5′ TpA base-pair step has a poor stacking overlap and a wider phosphodiester link (Abu-daya *et al*., 1995; Klug *et al*., 1979). Stacking energy differences have implications on the stability, bendability of DNA, and minor groove width through influencing base-pair parameters, The ApT results in a local bend towards the minor groove, whereas the TpA step bends toward the major groove (Klug *et al*., 1979; Stefl *et al*., 2004). The (5′ [TA][TA][TA]) motif had three TpA units while the (5′AA[TA]AT and 5′A[TA]TTT) motif have one TpA unit. From the calculation observed in the study results, the 5′ ATATTT and 5′AATAAT motifs had better binding affinities compared to TATATA due to the ApT narrower widths which resulted in more van der Waal contacts, nonetheless, it should be noted that a TpA step may also play a role in terms of malleability in having a ligand or a transcription factor being hosted in the minor groove as seen in TATATA with lower binding affinities, such phenomena are observed in TpA, CpG TpG, and CpA (Klug *et al*., 1979).

Finally, it was important to predict targeted therapeutics for macrophages where ASFV replicates predominantly. Liposomes naturally target macrophages (Kelly *et al*., 2011). This natural targeting ability can be utilized in drug delivery involving ASFV. From the ADME predictions, it was observed that tris-benzimidazole and Hoechst 33258 had high LogP values. High LogP values correlate positively with liposomal packaging ability thus tris-benzimidazole and Hoechst 33258 could potentially be packaged in liposomes and engineered to show substantial accumulation in macrophages to reduce potentially toxic effects as a result of specific release (Nii & Ishii, 2005). Congocidine congeners nonetheless, had low LogP and high solubility values. The high solubility

of these congeners might be beneficial in parenteral administration since a drug has to be highly soluble in water to provide a sufficient quantity of active elements in a minor volume of pharmaceutical dosage (Savjani *et al*., 2012). Additionally, congocidine 2 and congocidine 3, have previously been shown to have potential therapeutic properties in Herpes Simplex Virus, with low cytotoxicity in congocidine 2 and no cytotoxicity in congocidine 3 (Bialer *et al*., 1980).

5.5 CONCLUSION

In this section, reversible minor groove binders' congocidine 2, congocidine 3, and Hoechst 33258 were studied targeting the characterized TATTT motif of a synthetic DNA construct. The outcomes of this *in silico* study revealed that reversible minor groove binders can target the TATTT motif similar to those characterized in ASFV which is vital for transcription.

CHAPTER SIX

6.1 GENERAL DISCUSSION

The first part of this thesis attempted to find a transcription factor from the ASFV- uncharacterized proteome. Additionally, an attempt was made to validate if antiviral/s can potentially target TATA-like motifs. A substantial number of studies have shown the existence of putative TATA-like motifs in NCLDVs (Oliveira *et al*., 2017, 2018). Widespread distribution of motifs including the TATTT and TATATA have been shown to play the role of a promoter in asfarvirus (García-Escudero & Viñuela, 2000). The search for a transcriptional factor with the requisite features capable of interacting with these TATA-like motifs has not been carried out (García-Escudero & Viñuela, 2000). Therefore an *in silico* search was applied to find a TBP having inherent features to interact with these motifs. The approach taken to identify the TBP involved the prediction of homologous sequences, modeling of structures, and assignment of function, involving mutually exclusive algorithms.

The results of this study were the significant finding of a novel uncharacterized protein, pB263R, which had features of a TBP (Appendix VIII). This helped to bridge a well-defined knowledge gap in identifying a possible transcriptional factor that could interact with putative TATA-like promoter elements in ASFV a problem initially described by García-Escudero & Viñuela, 2000.

The second part of this study validated if reversible minor groove binders could be exploited in targeting druggable TATATA motifs similar to those observed in ASFV late transcription. Various studies have shown the existence of late promoter motifs (García-Escudero & Viñuela, 2000).

Nonetheless, there is no approved reversible minor groove ligand targeting the identified druggable TATATA target in ASFV. The approaches used in demonstrating if reversible MGBs could be used in targeting the TATATA motifs are Monte carlo docking simulation, calculated binding energy, and molecular dynamic simulations in a dynamic environment. The study examined how a less cytotoxic congocidine congener (congocidine 2), a non-cytotoxic congener (congocidine 3), and a tris-benzimidazole bioisostere, could potentially be used in targeting the synthetic TATATA motif similar to those observed in ASFV that is vital in late gene transcription. From the results, the minor groove binders (congocidine 2, congocidine 3, and a tris-benzimidazole) could be used in targeting 5 TATATA bases. Such hexametric motifs are similar to conserved promoter motifs that have been characterized in ASFV and are responsible for transcription regulation (Appendix VI).

The third part of this study further investigated the possibility of targeting a second TATTT motif that has been characterized and shown to exist in ASFV (Oliveira *et al.,* 2018). Deletion substitution and mutation of the TATTT motif are fatal to transcription (García-Escudero & Viñuela, 2000). No drug has been approved for ASFV treatment (Arabyan *et al*., 2019) and no reversible minor groove binders have been shown to target the essential TATTT motif in ASFV. The study, therefore, validated *in silico* if congocidine congeners and Hoechst 33258 could be used in targeting the TATTT druggable motif. The TATTT motifs are present in ASFV intergenic regions upstream the transcription start site (García-Escudero & Viñuela, 2000). Their motif copy numbers are significantly higher than other identified motifs (Oliveira *et al*., 2018). The approach taken to validate if congocidine congeners and Hoechst 33258 could target the TATTT motif

included Monte Carlo biased docking, free energy predictive calculations, and MD simulation. From the results, the study established that congocidine congeners and Hoechst 33258 could target TATTT motifs.

From the results of this study's first objective, pB263R has the requisite features to bind to TATA elements. TATA-like elements TATATA and TATTT have been characterized in ASFV (García-Escudero & Viñuela, 2000; Oliveira *et al*., 2017, 2018) and their mutation stagnates transcription (García-Escudero & Viñuela, 2000). Minor groove binders can also stagnate transcription by binding to TATA elements (Bellorini *et al*., 1995; Chiang *et al*., 1994). TATA-like elements have been characterized in vaccinia virus, a nucleocytoplasmic large DNA virus that replicates in macrophages, and minor groove binders can stagnate transcription in vaccinia virus late promoters (Broyles *et al*., 2004; Yakimovich *et al*., 2017). From these results, a TBP was identified and minor groove binders capable of separating cytotoxicity from antiviral activity were verified as having the capability to target both TATATA and TATTT promoter motifs in the second and third objectives respectively. It is highly likely the identified MGBs will inhibit transcription in ASFV genotypes by blocking the transcription factor identified and increasing the thermal melting temperature of virus DNA.

The study also established that the minor groove binders could tolerate G-C bases flanking the A-T regions in 1 or 2 base pairs.

6.2 CONCLUSIONS

The three main conclusions and significant original contributions to knowledge were drawn from this study. First, a protein pB263R with features of a TBP was identified from the uncharacterized ASFV ORFs using *in silico* protein structure prediction approaches. Second, from docking, binding affinity studies, and MD simulation studies. It was determined that congocidine 2, congocidine 3, and tris-benzimidazole could be used to target the TATATA motif that is similar to characterized promoter motifs in ASFV, which is crucial in ASFV transcription of multiple late genes. Third, from docking, binding affinity studies, and MD simulation studies. It was determined that congocidine 2, congocidine 3, and Hoechst 33258 are capable of targeting the TATTT motif. The TATTT motif is similar to characterized promoter motifs in ASFV and is essential in the ASFV transcription of multiple genes.

6.3 RECOMMENDATIONS

The following recommendations are drawn from this study:

- X-ray and NMR studies should be exploited to determine the exact mechanism of kinking of TBP with three phenylalanine residues, even though their existence has been shown, there is a deficit of literature showing the exact kinking mechanism of TBPs with three phenylalanine residues.
- Liposomal encapsulation of tris-benzimidazole and Hoechst 33258 should be carried out.
- Testing the efficacy of Hoechst 33258, congocidine 2, congocidine 3, and tris-benzimidazole in targeting ASFV infected porcine macrophages *in vitro* and *in vivo* should be carried out.

REFERENCES

Abagyan, R., Totrov, M., & Kuznetsov, D. (1994). ICM-A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry*, *15*(5), 488–506. https://doi.org/10.1002/jcc.540150503

Abrescia, N. G., Malinina, L., Fernandez, L. G., Huynh-Dinh, T., Neidle, S., & Subirana, J. A. (1999). Structure of the oligonucleotide d(CGTATATACG) as a site-specific complex with nickel ions. *Nucleic Acids Research*, *27*(7), 1593–1599. https://doi.org/10.1093/nar/27.7.1593

Abu-daya, A., Brown, P. M., & Fox, K. R. (1995). DNA sequence preferences of several AT-selective minor groove binding ligands. *Nucleic Acids Research*, *23*(17), 3385–3392. https://doi.org/10.1093/nar/23.17.3385

Abu-daya, A., & Fox, K. R. (1997). Interaction of minor groove binding ligands with long AT tracts. *Nucleic Acids Research*, *25*(24), 4962–4969. https://doi.org/10.1093/nar/25.24.4962

Achenbach, J. E., Gallardo, C., Nieto-Pelegrín, E., Rivera-Arroyo, B., Degefa-Negi, T., Arias, M., Jenberie, S., Mulisa, D. D., Gizaw, D., Gelaye, E., Chibssa, T. R., Belaye, A., Loitsch, A., Forsa, M., Yami, M., Diallo, A., Soler, A., Lamien, C. E., & Sánchez-Vizcaíno, J. M. (2017). Identification of a New Genotype of African Swine Fever Virus in Domestic Pigs from Ethiopia. *Transboundary and Emerging Diseases*, *64*(5), 1393–1404. https://doi.org/10.1111/tbed.12511

Afonso, C. L., Piccone, M. E., Zaffuto, K. M., Neilan, J., Kutish, G. F., Lu, Z., Balinsky, C. A., Gibb, T. R., Bean, T. J., Zsak, L., & Rock, D. L. (2004). African swine fever virus multigene family 360 and 530 genes affect host interferon response. *Journal of Virology*, *78*(4), 1858–1864. https://doi.org/10.1128/JVI.78.4.1858

Ahsan, F., Rivas, I. P., Khan, M. A., & Torres Suárez, A. I. (2002). Targeting to macrophages: Role of physicochemical properties of particulate carriers - Liposomes and microspheres -

On the phagocytosis by macrophages. *Journal of Controlled Release*, *79*(1–3), 29–40. https://doi.org/10.1016/S0168-3659(01)00549-1

Alcami, A., Angulo, A., & Vinuela, E. (1993). Mapping and sequence of the gene encoding the African swine fever virion protein of M(r) 11500. *Journal of General Virology*, *74*(11), 2317–2324. https://doi.org/10.1099/0022-1317-74-11-2317

Alder, B. J., & Wainwright, T. E. (1959). Studies in molecular dynamics. I. General method. *The Journal of Chemical Physics*, *31*(2), 459–466. https://doi.org/10.1063/1.1730376

Alder, B. J., & Wainwright, T. E. (1960). Studies in molecular dynamics. II. Behavior of a small number of elastic spheres. *The Journal of Chemical Physics*, *33*(5), 1439–1451. https://doi.org/10.1063/1.1731425

Alejo, A., Andrés, G., & Salas, M. L. (2003). African Swine Fever Virus Proteinase Is Essential for Core Maturation and Infectivity. *Journal of Virology*, *77*(10), 5571–5577. https://doi.org/10.1128/jvi.77.10.5571-5577.2003

Alejo, A., Matamoros, T., Guerra, M., & Andrés, G. (2018). A Proteomic Atlas of the African Swine Fever Virus Particle. *Journal of Virology*, *92*(23), 1–18. https://doi.org/10.1128/JVI.01293-18

Alejo, A., Yáñez, R. J., Rodríguez, J. M., Viñuela, E., & Salas, M. L. (1997). African swine fever virus trans-prenyltransferase. *Journal of Biological Chemistry*, *272*(14), 9417–9423. https://doi.org/10.1074/jbc.272.14.9417

Almazán, F., Murguía, J. R., Rodríguez, J. M., de la Vega, I., & Viñuela, E. (1995). A set of African swine fever virus tandem repeats shares similarities with SAR-like sequences. *The Journal of General Virology*, *76 (Pt 4)*(4), 729–740. https://doi.org/10.1099/0022-1317-76-4-729

Almazán, F., Rodríguez, J. M., Angulo, A., Viñuela, E., & Rodriguez, J. F. (1993). Transcriptional mapping of a late gene coding for the p12 attachment protein of African swine fever virus. *Journal of Virology*, *67*(1), 553–556.

https://doi.org/10.1128/JVI.67.1.553-556.1993

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Alva, V., Nam, S.-Z., Söding, J., & Lupas, A. N. (2016). The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Research*, *44*(W1), W410-5. https://doi.org/10.1093/nar/gkw348

Andreani, J., Khalil, J. Y. B., Sevvana, M., Benamar, S., Di Pinto, F., Bitam, I., Colson, P., Klose, T., Rossmann, M. G., Raoult, D., & La Scola, B. (2017). Pacmanvirus, a new giant icosahedral virus at the crossroads between asfarviridae and faustoviruses. *Journal of Virology*, *91*(14), 1–11. https://doi.org/10.1128/jvi.00212-17

Anfinsen, C. (1973). Principles that govern the folding of protein chains. *Science*, *181*(4096), 223–230.

Arabyan, E., Kotsynyan, A., Hakobyan, A., & Zakaryan, H. (2019). Antiviral agents against African swine fever virus. *Virus Research*, *270*, 197669. https://doi.org/10.1016/j.virusres.2019.197669

Arias, M., de la Torre, A., Dixon, L., Gallardo, C., Jori, F., Laddomada, A., Martins, C., Parkhouse, R. M., Revilla, Y., Rodriguez, F. and J.-M., & Sanchez-Vizcaino. (2017). Approaches and perspectives for development of African swine fever virus vaccines. *Vaccines*, *5*(4), 35. https://doi.org/10.3390/vaccines5040035

Aymami, J., Nunn, C. M., & Neidle, S. (1999). DNA minor groove recognition of a non-self-complementary AT-rich sequence by a tris-benzimidazole ligand. *Nucleic Acids Research*,

27(13), 2691–2698. https://doi.org/10.1093/nar/27.13.2691

Backer, J. A., Vrancken, R., Neyts, J., & Goris, N. (2013). The potential of antiviral agents to control classical swine fever: A modelling study. *Antiviral Research*, *99*(3), 245–250. https://doi.org/10.1016/j.antiviral.2013.06.013

Bajrai, L. H., Benamar, S., Azhar, E. I., Robert, C., Levasseur, A., Raoult, D., & La Scola, B. (2016). Kaumoebavirus, a new virus that clusters with Faustoviruses and Asfarviridae. *Viruses*, *8*(11). https://doi.org/10.3390/v8110278

Basta, S., Gerber, H., Schaub, A., Summerfield, A., & McCullough, K. C. (2010). Cellular processes essential for African swine fever virus to infect and replicate in primary macrophages. *Veterinary Microbiology*, *140*(1–2), 9–17. https://doi.org/10.1016/j.vetmic.2009.07.015

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Bye-AJee, H., Cowley, A., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Castro, L. G., Figueira, L., … Zhang, J. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158–D169. https://doi.org/10.1093/nar/gkw1099

Battey, J. N. D., Kopp, J., Bordoli, L., Read, R. J., Clarke, N. D., & Schwede, T. (2007). Automated server predictions in CASP7. *Proteins*, *69 Suppl 8*(SUPPL. 8), 68–82. https://doi.org/10.1002/prot.21761

Becker, Y., Asher, Y., & Zakay-Rones, Z. (1972). Congocidine and distamycin A, antipoxvirus antibiotics. *Antimicrobial Agents and Chemotherapy*, *1*(6), 483–488. https://doi.org/10.1128/AAC.1.6.483

Beerman, T. A., Woynarowski, J. M., Sigmund, R. D., Gawron, L. S., Rao, K. E., & Lown, J. W. (1991). Netropsin and bis-netropsin analogs as inhibitors of the catalytic activity of mammalian DNA topoisomerase II and topoisomerase cleavable complexes. *BBA - Gene Structure and Expression*, *1090*(1), 52–60. https://doi.org/10.1016/0167-4781(91)90036-L

Bellorini, M., Moncollin, V., D'lncalci, M., Mongelli, N., & Mantovani, R. (1995). Distamycin A and tallimustine inhibit TBP binding and basal in vitro transcription. *Nucleic Acids Research*, *23*(10), 1657–1663. https://doi.org/10.1093/nar/23.10.1657

Beltran-Alcrudo, D., Falco, J. R., Raizman, E., & Dietze, K. (2019). Transboundary spread of pig diseases: the role of international trade and travel. *BMC Veterinary Research*, *15*(1), 64. https://doi.org/10.1186/s12917-019-1800-5

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Bialer, M., Yagen, B., & Mechoulam, R. (1978). A total synthesis of distamycin a, an antiviral antibiotic. *Tetrahedron*, *34*(15), 2389–2391. https://doi.org/10.1016/0040-4020(78)89056-5

Bialer, M., Yagen, B., Mechoulam, R., & Becker, Y. (1980). Structure-activity relationships of pyrrole amidine antiviral antibiotics. 2. Preparation of mono- and tripyrrole derivatives of congocidine. *Journal of Medicinal Chemistry*, *23*(10), 1144–1148. http://www.ncbi.nlm.nih.gov/pubmed/6252323

Bignon, E., Chan, C. H., Morell, C., Monari, A., Ravanat, J. L., & Dumont, E. (2017). Molecular dynamics insights into polyamine–DNA binding modes: Implications for cross-link selectivity. *Chemistry - A European Journal*, *23*(52), 12845–12852. https://doi.org/10.1002/chem.201702065

Bishop, R. P., Fleischauer, C., de Villiers, E. P., Okoth, E. A., Arias, M., Gallardo, C., & Upton, C. (2015). Comparative analysis of the complete genome sequences of Kenyan African swine fever virus isolates within p72 genotypes IX and X. *Virus Genes*, *50*(2), 303–309. https://doi.org/10.1007/s11262-014-1156-7

Boinas, F. S., Wilson, A. J., Hutchings, G. H., Martins, C., & Dixon, L. J. (2011). The persistence of African swine fever virus in field-infected Ornithodoros erraticus during the ASF endemic period in Portugal. *PLoS ONE*, *6*(5).

https://doi.org/10.1371/journal.pone.0020383

Borca, M. V., O'Donnell, V., Holinka, L. G., Ramírez-Medina, E., Clark, B. A., Vuono, E. A., Berggren, K., Alfano, M., Carey, L. B., Richt, J. A., Risatti, G. R., & Gladue, D. P. (2018). The L83L ORF of African swine fever virus strain Georgia encodes for a non-essential gene that interacts with the host protein IL-1β. *Virus Research*, *249*(March), 116–123. https://doi.org/10.1016/j.virusres.2018.03.017

Bostock-Smith, C. E., & Searle, M. S. (1999). DNA minor groove recognition by bis-benzimidazole analogues of Hoechst 33258: Insights into structure-DNA affinity relationships assessed by fluorescence titration measurements. *Nucleic Acids Research*, *27*(7), 1619–1624. https://doi.org/10.1093/nar/27.7.1619

Bowers, K. J., Chow, D. E., Xu, H., Dror, R. O., Eastwood, M. P., Gregersen, B. A., Klepeis, J. L., Kolossvary, I., Moraes, M. A., Sacerdoti, F. D., Salmon, J. K., Shan, Y., & Shaw, D. E. (2006). Scalable algorithms for molecular dynamics simulations on commodity clusters. *ACM/IEEE SC 2006 Conference (SC'06)*, 43–43. https://doi.org/10.1109/SC.2006.54

Bowie, J. U., Lüthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science (New York, N.Y.)*, *253*(5016), 164–170. https://doi.org/10.1126/science.1853201

Bradley, P., Misura, K. M. S., & Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science (New York, N.Y.)*, *309*(5742), 1868–1871. https://doi.org/10.1126/science.1113801

Broyles, S. S., Kremer, M., & Knutson, B. A. (2004). Antiviral activity of distamycin A against vaccinia virus is the result of inhibition of postreplicative mRNA synthesis. *Journal of Virology*, *78*(4), 2137–2141. https://doi.org/10.1128/jvi.78.4.2137-2141.2004

Cackett, G., Matelska, D., Sýkora, M., Portugal, R., Malecki, M., Bähler, J., Dixon, L., & Werner, F. (2020). The African Swine Fever Virus Transcriptome. *Journal of Virology*, *94*(9). https://doi.org/10.1128/JVI.00119-20

Cackett, G., Sýkora, M., & Werner, F. (2020). Transcriptome view of a killer: African swine fever virus. *Biochemical Society Transactions*, *48*(4), 1569–1581. https://doi.org/10.1042/BST20191108

Carrillo, C., Borca, M. V, Afonso, C. L., Onisk, D. V, & Rock, D. L. (1994). Long-term persistent infection of swine monocytes/macrophages with African swine fever virus. *Journal of Virology*, *68*(1), 580–583.

Carrondo, M. A., Coll, M., Aymami, J., Wang, A. H., van der Marel, G. A., van Boom, J. H., & Rich, A. (1989). Binding of a Hoechst dye to d(CGCGATATCGCG) and its influence on the conformation of the DNA fragment. *Biochemistry*, *28*(19), 7849–7859. https://doi.org/10.1021/bi00445a047

Chacón, M. R., Almazán, F., Nogal, M. L., Viñuela, E., & Rodríguez, J. F. (1995). The African swine fever virus IAP homolog is a late structural polypeptide. *Virology*, *214*(2), 670–674. https://doi.org/10.1006/viro.1995.0083

Chandonia, J. M., Fox, N. K., & Brenner, S. E. (2019). SCOPe: Classification of large macromolecular structures in the structural classification of proteins - Extended database. *Nucleic Acids Research*, *47*(D1), D475–D481. https://doi.org/10.1093/nar/gky1134

Chapman, Tcherepanov, V., Upton, C., & Dixon, L. K. (2008). Comparison of the genome sequences of non-pathogenic and pathogenic African swine fever virus isolates. *The Journal of General Virology*, *89*(Pt 2), 397–408. https://doi.org/10.1099/vir.0.83343-0

Chiang, S. Y., Welch, J., Rauscher, F. J., & Beerman, T. A. (1994). Effects of minor groove binding drugs on the interaction of TATA box binding protein and TFIIA with DNA. *Biochemistry*, *33*(23), 7033–7040. https://doi.org/10.1021/bi00189a003

Coelho, J., Ferreira, F., Martins, C., & Leitão, A. (2016). Functional characterization and inhibition of the type II DNA topoisomerase coded by African swine fever virus. *Virology*, *493*, 209–216. https://doi.org/10.1016/j.virol.2016.03.023

Coll, M., Aymami, J., van der Marel, G. A., van Boom, J. H., Rich, A., & Wang, A. H. J. (1989).

Molecular structure of the netropsin-d(CGCGATATCGCG) complex: DNA conformation in an alternating AT segment. *Biochemistry*, *28*(1), 310–320. https://doi.org/10.1021/bi00427a042

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. (1995). A second generation force field for the simulation of proteins , nucleic acids , and organic molecules. *Journal of the American Chemical Society*, *117*(19), 5179–5197. https://doi.org/10.1021/ja00124a002

Cortesi, R., Romagnoli, R., Drechsler, M., Menegatti, E., Zaid, A. N., Ravani, L., & Esposito, E. (2010). Liposomes- and ethosomes-associated distamycins: a comparative study. *Journal of Liposome Research*, *20*(4), 277–285. https://doi.org/10.3109/08982100903443057

Cortesi, R., Romagnoli, R., Menegatti, E., Esposito, E., Cervellati, F., & Nastruzzi, C. (2004). Liposomes containing distamycins: preparation, characterization and antiproliferative activity. *Drug Delivery*, *11*(2), 83–88. https://doi.org/10.1080/10717540490280516

Costard, S., Wieland, B., de Glanville, W., Jori, F., Rowlands, R., Vosloo, W., Roger, F., Pfeiffer, D. U., & Dixon, L. K. (2009). African swine fever: how can global spread be prevented? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1530), 2683–2696. https://doi.org/10.1098/rstb.2009.0098

Daina, A., Michielin, O., & Zoete, V. (2017). SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific Reports*, *7*(October 2016), 42717. https://doi.org/10.1038/srep42717

Dixon, Chapman, D. A. G., Netherton, C. L., & Upton, C. (2013). African swine fever virus replication and genomics. *Virus Research*, *173*(1), 3–14. https://doi.org/10.1016/j.virusres.2012.10.020

Dixon, L., Abrams, C. ., Bowick, G., Goatley, L., Kay-Jackson, P., Chapman, D., Liverani, E., Nix, R., Silk, R., & Zhang, F. (2004). African swine fever virus proteins involved in evading host defence systems. *Veterinary Immunology and Immunopathology*, *100*(3–4),

117–134. https://doi.org/10.1016/j.vetimm.2004.04.002

Dixon, L. K., Islam, M., Nash, R., & Reis, A. L. (2019). African swine fever virus evasion of host defences. *Virus Research*, *266*, 25–33. https://doi.org/10.1016/j.virusres.2019.04.002

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, *14*(9), 755–763. https://doi.org/10.1093/bioinformatics/14.9.755

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Elena, S., Daniel, P., & Revilla, Y. (2019). Development of vaccines against African swine fever virus. *Virus Research*. https://doi.org/10.1016/j.virusres.2019.03.022

Fabregas, J., García, D., Fernandez-Alonso, M., Rocha, A. I., Gómez-Puertas, P., Escribano, J. M., Otero, A., & Coll, J. M. (1999). In vitro inhibition of the replication of haemorrhagic septicaemia virus (VHSV) and African swine fever virus (ASFV) by extracts from marine microalgae. *Antiviral Research*, *44*(1), 67–73. https://doi.org/10.1016/s0166-3542(99)00049-2

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., & Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, *44*(D1), D279–D285. https://doi.org/10.1093/nar/gkv1344

Fox, N. K., Brenner, S. E., & Chandonia, J.-M. (2014). SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, *42*(Database issue), D304-9. https://doi.org/10.1093/nar/gkt1240

Freitas, F. B., Frouco, G., Martins, C., Leitão, A., & Ferreira, F. (2016). In vitro inhibition of African swine fever virus-topoisomerase II disrupts viral replication. *Antiviral Research*, *134*, 34–41. https://doi.org/10.1016/j.antiviral.2016.08.021

Frouco, G., Freitas, F. B., Martins, C., & Ferreira, F. (2017). Sodium phenylbutyrate abrogates African swine fever virus replication by disrupting the virus-induced hypoacetylation status of histone H3K9/K14. *Virus Research*, *242*(September), 24–29. https://doi.org/10.1016/j.virusres.2017.09.009

Frouco, G., Freitas, F., Coelho, J., Leitão, A., Martins, C., & Ferreira, F. (2017). DNA-Binding properties of African swine fever virus pA104R, a histone-like protein involved in viral replication and transcription. *Journal of Virology*, *91*(12), 1–14. https://doi.org/10.1128/JVI.02498-16

Fuchs, J. E., Spitzer, G. M., Javed, A., Biela, A., Kreutz, C., Wellenzohn, B., & Liedl, K. R. (2011). Minor groove binders and drugs targeting proteins cover complementary regions in chemical shape space. *Journal of Chemical Information and Modeling*, *51*(9), 2223–2232. https://doi.org/10.1021/ci200237c

Galindo, Hernáez, B., Berná, J., Fenoll, J., Cenis, J. L., Escribano, J. M., & Alonso, C. (2011). Comparative inhibitory activity of the stilbenes resveratrol and oxyresveratrol on African swine fever virus replication. *Antiviral Research*, *91*(1), 57–63. https://doi.org/10.1016/j.antiviral.2011.04.013

Galindo, Immaculada, Almazán, F., Bustos, M. J., Viñuela, E., & Carrascosa, A. L. (2000). African swine fever virus EP153R open reading frame encodes a glycoprotein involved in the hemadsorption of infected cells. *Virology*, *266*(2), 340–351. https://doi.org/10.1006/viro.1999.0080

Galindo, Inmaculada, & Alonso, C. (2017). African swine fever virus: A review. *Viruses*, *9*(5), 103. https://doi.org/10.3390/v9050103

Galindo, Inmaculada, Viñuela, E., & Carrascosa, A. L. (2000). Characterization of the African swine fever virus protein p49: A new late structural polypeptide. *Journal of General Virology*, *81*(1), 59–65. https://doi.org/10.1099/0022-1317-81-1-59

García-escudero, R., & Viñuela, E. (2000). Structure of African Swine Fever Virus Late

Promoters : Requirement of a TATA Sequence at the Initiation Region Structure of African Swine Fever Virus Late Promoters *: Requirement of a TATA Sequence at the Initiation Region*. *74*(17). https://doi.org/10.1128/JVI.74.17.8176-8182.2000.Updated

García-Escudero, R., & Viñuela, E. (2000). Structure of African swine fever virus late promoters: requirement of a TATA sequence at the initiation region. *Journal of Virology*, *74*(17), 8176–8182. https://doi.org/10.1128/jvi.74.17.8176-8182.2000

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., & Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*, *31*(13), 3784–3788. https://doi.org/10.1093/nar/gkg563

Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, *32*(90001), 258D – 261. https://doi.org/10.1093/nar/gkh036

Genheden, S., & Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*, *10*(5), 449–461. https://doi.org/10.1517/17460441.2015.1032936

Gladue, D. P., & Borca, M. V. (2022). Recombinant ASF Live Attenuated Virus Strains as Experimental Vaccine Candidates. *Viruses*, *14*(5). https://doi.org/10.3390/v14050878

Hakizimana, J. N., Yona, C., Kamana, O., Nauwynck, H., & Misinzo, G. (2021). African Swine Fever Virus Circulation between Tanzania and Neighboring Countries: A Systematic Review and Meta-Analysis. *Viruses*, *13*(2). https://doi.org/10.3390/v13020306

Hakobyan, A., Arabyan, E., Avetisyan, A., Abroyan, L., Hakobyan, L., & Zakaryan, H. (2016). Apigenin inhibits African swine fever virus infection in vitro. *Archives of Virology*, *161*(12), 3445–3453. https://doi.org/10.1007/s00705-016-3061-y

Hakobyan, A., Arabyan, E., Kotsinyan, A., Karalyan, Z., Sahakyan, H., Arakelov, V., Nazaryan, K., Ferreira, F., & Zakaryan, H. (2019). Inhibition of African swine fever virus infection by genkwanin. *Antiviral Research*, *167*, 78–82. https://doi.org/10.1016/j.antiviral.2019.04.008

Hänggi, M., Bannwarth, W., & Stunnenberg, H. G. (1986). Conserved TAAAT motif in vaccinia virus late promoters: overlapping TATA box and site of transcription initiation. *The EMBO Journal*, *5*(5), 1071–1076. http://www.ncbi.nlm.nih.gov/pubmed/3013615

Harshman, K. D., & Dervan, P. B. (1985). Molecular recognition of B-DNA by Hoechst 33258. *Nucleic Acids Research*, *13*(13), 4825–4835. https://doi.org/10.1093/nar/13.13.4825

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(22), 10915–10919. https://doi.org/10.1073/pnas.89.22.10915

Hernáez, B., Guerra, M., Salas, M. L., & Andrés, G. (2016). African Swine Fever Virus Undergoes Outer Envelope Disruption, Capsid Disassembly and Inner Envelope Fusion before Core Release from Multivesicular Endosomes. *PLoS Pathogens*, *12*(4), e1005595. https://doi.org/10.1371/journal.ppat.1005595

Hernáez, B., Tarragó, T., Giralt, E., Escribano, J. M., & Alonso, C. (2010). Small peptide inhibitors disrupt a high-affinity interaction between cytoplasmic dynein and a viral cargo protein. *Journal of Virology*, *84*(20), 10792–10801. https://doi.org/10.1128/JVI.01168-10

Hlaka, L., Rosslee, M.-J., Ozturk, M., Kumar, S., Parihar, S. P., Brombacher, F., Khalaf, A. I., Carter, K. C., Scott, F. J., Suckling, C. J., & Guler, R. (2017). Evaluation of minor groove binders (MGBs) as novel anti-mycobacterial agents and the effect of using non-ionic surfactant vesicles as a delivery system to improve their efficacy. *The Journal of Antimicrobial Chemotherapy*, *72*(12), 3334–3341. https://doi.org/10.1093/jac/dkx326

Holm, L., & Rosenström, P. (2010). Dali server: conservation mapping in 3D. *Nucleic Acids Research*, *38*(Web Server issue), W545-9. https://doi.org/10.1093/nar/gkq366

Huang, Y. J., Mao, B., Aramini, J. M., & Montelione, G. T. (2014). Assessment of template-based protein structure predictions in CASP10. *Proteins: Structure, Function and*

Bioinformatics, *82*(SUPPL.2), 43–56. https://doi.org/10.1002/prot.24488

Huntley, R. P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M. J., & O'Donovan, C. (2015). The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Research*, *43*(D1), D1057–D1063. https://doi.org/10.1093/nar/gku1113

Hurtado, C., Bustos, M. J., Sabina, P., Nogal, M. L., Granja, A. G., González, M. E., Gónzalez-Porqué, P., Revilla, Y., & Carrascosa, A. L. (2008). Antiviral activity of lauryl gallate against animal viruses. *Antiviral Therapy*, *13*(7), 909–917. http://www.ncbi.nlm.nih.gov/pubmed/19043925

Iyer, L. ., Balaji, S., Koonin, E. V., & Aravind, L. (2006). Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Research*, *117*(1), 156–184. https://doi.org/10.1016/j.virusres.2006.01.009

Jian Zhang, Liang, Y., & Yang, Z. (2011). Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure (London, England : 1993)*, *19*(12), 1784–1795. https://doi.org/10.1016/j.str.2011.09.022

Jianyi, Y., Roy, A., & Yang, Z. (2013a). BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Research*, *41*(D1), 1096–1103. https://doi.org/10.1093/nar/gks966

Jianyi, Y., Roy, A., & Yang, Z. (2013b). Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, *29*(20), 2588–2595. https://doi.org/10.1093/bioinformatics/btt447

Jianyi, Y., Yan, R., Roy, A., Xu, D., Poisson, J., & Yang, Z. (2014). The I-TASSER Suite: protein structure and function prediction. *Nature Methods*, *12*(1), 7–8. https://doi.org/10.1038/nmeth.3213

Jianyi, Y., & Yang, Z. (2015). Protein Structure and Function Prediction Using I-TASSER. *Current Protocols in Bioinformatics*, *52*(1), 5.8.1-5.8.15. https://doi.org/10.1002/0471250953.bi0508s52

Jinrui, X., & Yang, Z. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics (Oxford, England)*, *26*(7), 889–895. https://doi.org/10.1093/bioinformatics/btq066

Jori, F., & Bastos, A. D. S. (2009). Role of wild suids in the epidemiology of african swine fever. *EcoHealth*, *6*(2), 296–310. https://doi.org/10.1007/s10393-009-0248-7

Juo, Z. S., Chiu, T. K., Leiberman, P. M., Baikalov, I., Berk, A. J., & Dickerson, R. E. (1996). How proteins recognize the TATA box. *Journal of Molecular Biology*, *261*(2), 239–254. https://doi.org/10.1006/jmbi.1996.0456

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, *32*(5), 922–923. https://doi.org/10.1107/S0567739476001873

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., & Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. *Nature Protocols*, *7*(8), 1511–1522. https://doi.org/10.1038/nprot.2012.085

Kelly, C., Jefferies, C., & Cryan, S.-A. (2011). Targeted liposomal drug delivery to monocytes and macrophages. *Journal of Drug Delivery*, *2011*, 727241. https://doi.org/10.1155/2011/727241

Keßler, C., Forth, J. H., Keil, G. M., Mettenleiter, T. C., Blome, S., & Karger, A. (2018). The intracellular proteome of African swine fever virus. *Scientific Reports*, *8*(1), 1–9. https://doi.org/10.1038/s41598-018-32985-z

Khan, G. S., Shah, A., Zia-ur-Rehman, & Barker, D. (2012). Chemistry of DNA minor groove binding agents. *Journal of Photochemistry and Photobiology. B, Biology*, *115*, 105–118. https://doi.org/10.1016/j.jphotobiol.2012.07.003

Kihara, D., & Skolnick, J. (2003). The PDB is a covering set of small protein structures. *Journal of Molecular Biology*, *334*(4), 793–802. https://doi.org/10.1016/j.jmb.2003.10.027

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S.,

Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., & Bryant, S. H. (2016). PubChem Substance and Compound databases. *Nucleic Acids Research*, *44*(D1), D1202-13. https://doi.org/10.1093/nar/gkv951

Kinch, L. N., Li, W., Schaeffer, R. D., Dunbrack, R. L., Monastyrskyy, B., Kryshtafovych, A., & Grishin, N. V. (2016). CASP 11 target classification. *Proteins*, *84 Suppl 1*(10), 20–33. https://doi.org/10.1002/prot.24982

King, K., Chapman, D., Argilaguet, J. M., Fishbourne, E., Hutet, E., Cariolet, R., Hutchings, G., Oura, C. A. L., Netherton, C. L., Moffat, K., Taylor, G., Le Potier, M.-F., Dixon, L. K., & Takamatsu, H.-H. (2011). Protection of European domestic pigs from virulent African isolates of African swine fever virus by experimental immunisation. *Vaccine*, *29*(28), 4593–4600. https://doi.org/10.1016/j.vaccine.2011.04.052

Kinyanyi, D., Obiero, G., Obiero, G. F. O., Amwayi, P., Mwaniki, S., & Wamalwa, M. (2018). In silico structural and functional prediction of African swine fever virus protein-B263R reveals features of a TATA-binding protein. *PeerJ*, *2018*(2), e4396. https://doi.org/10.7717/peerj.4396

Klug, A., Jack, A., Viswamitra, M. A., Kennard, O., Shakked, Z., & Steitz, T. A. (1979). A hypothesis on a specific sequence-dependent conformation of DNA and its relation to the binding of the lac-repressor protein. *Journal of Molecular Biology*, *131*(4), 669–680. https://doi.org/10.1016/0022-2836(79)90196-7

Knudsen, M., & Wiuf, C. (2010). Genome Databases The CATH database. *Human Genomics*, *4*(3), 207–212.

Knutson, B. A., Liu, X., Oh, J., & Broyles, S. S. (2006). Vaccinia virus intermediate and late promoter elements are targeted by the TATA-binding protein. *Journal of Virology*, *80*(14), 6784–6793. https://doi.org/10.1128/JVI.02705-05

Kopka, M. L., Yoon, C., Goodsell, D., Pjura, P., & Dickerson, R. E. (1985). The molecular origin of DNA-drug specificity in netropsin and distamycin. *Proceedings of the National*

Academy of Sciences of the United States of America, *82*(5), 1376–1380. https://doi.org/10.1073/pnas.82.5.1376

Koster, M. J. E., Snel, B., & Timmers, H. T. M. (2015). Genesis of chromatin and transcription dynamics in the origin of species. *Cell*, *161*(4), 724–736. https://doi.org/10.1016/j.cell.2015.04.033

Krieger, F., Möglich, A., & Kiefhaber, T. (2005). Effect of proline and glycine residues on dynamics and barriers of loop formation in polypeptide chains. *Journal of the American Chemical Society*, *127*(10), 3346–3352. https://doi.org/10.1021/ja042798i

Kryshtafovych, A., Krysko, O., Daniluk, P., Dmytriv, Z., & Fidelis, K. (2009). Protein structure prediction center in CASP8. *Proteins: Structure, Function and Bioinformatics*, *77*(SUPPL. 9), 5–9. https://doi.org/10.1002/prot.22517

Lam, P. C.-H., Abagyan, R., & Totrov, M. (2018). Ligand-biased ensemble receptor docking (LigBEnD): a hybrid ligand/receptor structure-based approach. *Journal of Computer-Aided Molecular Design*, *32*(1), 187–198. https://doi.org/10.1007/s10822-017-0058-x

León, P. De, Bustos, M. J., & Carrascosa, A. L. (2012). Laboratory methods to study African swine fever virus. *Virus Research*. https://doi.org/10.1016/j.virusres.2012.09.013

Levitt, M., & Gerstein, M. (1998). A unified statistical framework for sequence comparison and. *PNAS*, *95*(May), 5913–5920.

Li, Y., & Zhang, Y. (2009). REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. *Proteins: Structure, Function, and Bioinformatics*, *76*(3), 665–676. https://doi.org/10.1002/prot.22380

Liu, Q., Ma, B., Qian, N., Zhang, F., Tan, X., Lei, J., & Xiang, Y. (2019). Structure of the African swine fever virus major capsid protein p72. *Cell Research*, *29*(11), 953–955. https://doi.org/10.1038/s41422-019-0232-x

Lombardi, P., & Crisanti, A. (1997). Antimalarial activity of synthetic analogues of distamycin.

Pharmacology & Therapeutics, *76*(1–3), 125–133. https://doi.org/10.1016/s0163-7258(97)00095-8

Lovell, S. C., Davis, I. W., Arendall, W. B., de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S., & Richardson, D. C. (2003). Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins*, *50*(3), 437–450. https://doi.org/10.1002/prot.10286

Lown, J. W., Krowicki, K., Balzarini, J., Newman, R. A., & De Clercq, E. (1989). Novel linked antiviral and antitumor agents related to netropsin and distamycin: synthesis and biological evaluation. *Journal of Medicinal Chemistry*, *32*(10), 2368–2375. https://doi.org/10.1021/jm00130a024

Ma, J., Wang, S., Zhao, F., & Xu, J. (2013). Protein threading using context-specific alignment potential. *Bioinformatics (Oxford, England)*, *29*(13), i257-65. https://doi.org/10.1093/bioinformatics/btt210

Manning, G. S. (1978). The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. *Quarterly Reviews of Biophysics*, *11*(2), 179–246. https://doi.org/10.1017/S0033583500002031

Mariani, V., Kiefer, F., Schmidt T. Haas, J., & Schwede, T. (2011). Assessment of template based protein structure predictions in CASP9. *Proteins: Struct., Funct., Bioinf.*, *79*(S10), 37–58. https://doi.org/10.1002/prot.23177

McAndrew, M. B., Read, M., Sims, P. F., & Hyde, J. E. (1993). Characterisation of the gene encoding an unusually divergent TATA-binding protein (TBP) from the extremely A+T-rich human malaria parasite Plasmodium falciparum. *Gene*, *124*(2), 165–171. https://doi.org/10.1016/0378-1119(93)90390-o

Mondal, M., Mukherjee, S., & Bhattacharyya, D. (2014). Contribution of phenylalanine side chain intercalation to the TATA-box binding protein–DNA interaction: molecular dynamics and dispersion-corrected density functional theory studies. *Journal of Molecular Modeling*, *20*(11), 1–11. https://doi.org/10.1007/s00894-014-2499-7

Montgomery, R. E. (1921). On A Form of Swine Fever Occurring in British East Africa (Kenya Colony). *Journal of Comparative Pathology and Therapeutics*, *34*, 159–191. https://doi.org/10.1016/S0368-1742(21)80031-4

Mottola, C., Freitas, F. B., Simões, M., Martins, C., Leitão, A., & Ferreira, F. (2013). In vitro antiviral activity of fluoroquinolones against African swine fever virus. *Veterinary Microbiology*, *165*(1–2), 86–94. https://doi.org/10.1016/j.vetmic.2013.01.018

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2018). Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins*, *86 Suppl 1*(October 2018), 7–15. https://doi.org/10.1002/prot.25415

Narayanan, B., Westbrook, J., Ghosh, S., Petrov, A. I., Sweeney, B., Zirbel, C. L., Leontis, N. B., & Berman, H. M. (2014). The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Research*, *42*(Database issue), D114-22. https://doi.org/10.1093/nar/gkt980

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Nelson, S. M., Ferguson, L. R., & Denny, W. A. (2007). Non-covalent ligand/DNA interactions: Minor groove binding agents. *Mutation Research*, *623*(1–2), 24–40. https://doi.org/10.1016/j.mrfmmm.2007.03.012

Neves, M. A. C., Totrov, M., & Abagyan, R. (2012). Docking and scoring with ICM: the benchmarking results and strategies for improvement. *Journal of Computer-Aided Molecular Design*, *26*(6), 675–686. https://doi.org/10.1007/s10822-012-9547-0

Nii, T., & Ishii, F. (2005). Encapsulation efficiency of water-soluble and insoluble drugs in liposomes prepared by the microencapsulation vesicle method. *International Journal of Pharmaceutics*, *298*(1), 198–205. https://doi.org/10.1016/j.ijpharm.2005.04.029

Nikolov, D. B., & Burley, S. K. (1994). 2.1 A resolution refined structure of a TATA box-binding protein (TBP). *Nature Structural Biology*, *1*(9), 621–637.

https://doi.org/10.1038/nsb0994-621

Njau, E. P., Machuka, E. M., Cleaveland, S., Shirima, G. M., Kusiluka, L. J., Okoth, E. A., & Pelle, R. (2021). African Swine Fever Virus (ASFV): Biology, Genomics and Genotypes Circulating in Sub-Saharan Africa. *Viruses*, *13*(11), 1–17. https://doi.org/10.3390/v13112285

O'Donnell, V., Holinka, L. G., Krug, P. W., Gladue, D. P., Carlson, J., Sanford, B., Alfano, M., Kramer, E., Lu, Z., Arzt, J., Reese, B., Carrillo, C., Risatti, G. R., & Borca, M. V. (2015). African Swine Fever Virus Georgia 2007 with a Deletion of Virulence-Associated Gene 9GL (B119L), when Administered at Low Doses, Leads to Virus Attenuation in Swine and Induces an Effective Protection against Homologous Challenge. *Journal of Virology*, *89*(16), 8556–8566. https://doi.org/10.1128/JVI.00969-15

O'Donnell, V., Holinka, L. G., Sanford, B., Krug, P. W., Carlson, J., Pacheco, J. M., Reese, B., Risatti, G. R., Gladue, D. P., & Borca, M. V. (2016). African swine fever virus Georgia isolate harboring deletions of 9GL and MGF360/505 genes is highly attenuated in swine but does not confer protection against parental virus challenge. *Virus Research*, *221*, 8–14. https://doi.org/10.1016/j.virusres.2016.05.014

Oliveira, Andrade, A., Rodrigues, R., Arantes, T., Boratto, P., Silva, L., Dornas, F., Trindade, G., Drumond, B., La Scola, B., Kroon, E., & Abrahão, J. (2017). Promoter Motifs in NCLDVs: An Evolutionary Perspective. *Viruses*, *9*(1), 16. https://doi.org/10.3390/v9010016

Oliveira, de Aquino, I. L. M., Luiz, A. P. M. F., & Abrahão, J. S. (2018). Putative promoter motif analyses reinforce the evolutionary relationships among faustoviruses, kaumoebavirus, and asfarvirus. *Frontiers in Microbiology*, *9*(MAY), 1041. https://doi.org/10.3389/fmicb.2018.01041

Oliveros, M., García-Escudero, R., Alejo, A., Viñuela, E., Salas, M. L., & Salas, J. (1999). African swine fever virus dUTPase is a highly specific enzyme required for efficient replication in swine macrophages. *Journal of Virology*, *73*(11), 8934–8943. https://doi.org/10.1128/JVI.73.11.8934-8943.1999

Onzere, C. K., Bastos, A. D., Okoth, E. A., Lichoti, J. K., Bochere, E. N., Owido, M. G., Ndambuki, G., Bronsvoort, M., & Bishop, R. P. (2018). Multi-locus sequence typing of African swine fever viruses from endemic regions of Kenya and Eastern Uganda (2011–2013) reveals rapid B602L central variable region evolution. *Virus Genes*, *54*(1), 111–123. https://doi.org/10.1007/s11262-017-1521-4

Ortiz, A. R., Strauss, C. E. M., & Olmea, O. (2009). MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. *Protein Science*, *11*(11), 2606–2621. https://doi.org/10.1110/ps.0215902

Patel, S. R., Kvols, L. K., Rubin, J., O'Connell, M. J., Edmonson, J. H., Ames, M. M., & Kovach, J. S. (1991). Phase I-II study of pibenzimol hydrochloride (NSC 322921) in advanced pancreatic carcinoma. *Investigational New Drugs*, *9*(1), 53–57. https://doi.org/10.1007/BF00194545

Patikoglou, G. A., Kim, J. L., Sun, L., Yang, S. H., Kodadek, T., & Burley, S. K. (1999). TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes & Development*, *13*(24), 3217–3230. https://doi.org/10.1101/gad.13.24.3217

Pérez-núñez, D., Sunwoo, S., Sánchez, E. G., Haley, N., García-belmonte, R., Nogal, M., Morozov, I., Madden, D., Gaudreault, N. N., Mur, L., Shivanna, V., Richt, J. A., & Revilla, Y. (2019). Evaluation of a viral DNA-protein immunization strategy against African swine fever in domestic pigs. *Veterinary Immunology and Immunopathology*, *208*(October 2018), 34–43. https://doi.org/10.1016/j.vetimm.2018.11.018

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612. https://doi.org/10.1002/jcc.20084

Pikalo, J., Zani, L., Hühr, J., Beer, M., & Blome, S. (2019). Pathogenesis of African swine fever in domestic pigs and European wild boar – Lessons learned from recent animal trials. *Virus*

Research, *271*(March), 197614. https://doi.org/10.1016/j.virusres.2019.04.001

Plimpton, S. (1995). Fast Parallel Algorithms for Short – Range Molecular Dynamics. *117*(June 1994), 1–42.

Portugal, J., & Waring, M. J. (1987). Comparison of binding sites in DNA for berenil, netropsin and distamycin: A footprinting study. *European Journal of Biochemistry*, *167*(2), 281–289. https://doi.org/10.1111/j.1432-1033.1987.tb13334.x

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *35*(SUPPL. 1), 501–504. https://doi.org/10.1093/nar/gkl842

Ramachandran, G. N., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, *7*(1), 95–99. https://doi.org/10.1016/s0022-2836(63)80023-6

Raush, E., Totrov, M., Marsden, B. D., & Abagyan, R. (2009). A new method for publishing three-dimensional content. *PloS One*, *4*(10), e7394. https://doi.org/10.1371/journal.pone.0007394

Reis, Abrams, C. C., Goatley, L. C., Netherton, C., Chapman, D. G., Sanchez-Cordon, P., & Dixon, L. K. (2016). Deletion of African swine fever virus interferon inhibitors from the genome of a virulent isolate reduces virulence in domestic pigs and induces a protective response. *Vaccine*, *34*(39), 4698–4705. https://doi.org/10.1016/j.vaccine.2016.08.011

Ren, J., Jenkins, T. C., & Chaires, J. B. (2000). Energetics of DNA intercalation reactions. *Biochemistry*, *39*(29), 8439–8447. https://doi.org/10.1021/bi000474a

Reteno, D. G., Benamar, S., Khalil, J. B., Andreani, J., Armstrong, N., Klose, T., Rossmann, M., Colson, P., Raoult, D., & La Scola, B. (2015). Faustovirus, an asfarvirus-related new lineage of giant viruses infecting amoebae. *Journal of Virology*, *89*(13), 6585–6594. https://doi.org/10.1128/JVI.00115-15

Revilla, Y., Pérez-Núñez, D., & Richt, J. A. (2018). African Swine Fever Virus Biology and Vaccine Approaches. In *Advances in Virus Research* (Vol. 100, pp. 41–74). https://doi.org/10.1016/bs.aivir.2017.10.002

Rock, Daniel.L. (2017). Challenges for African swine fever vaccine development“ perhaps the end of the beginning.” *Veterinary Microbiology*, *206*, 52–58. https://doi.org/10.1016/j.vetmic.2016.10.003

Rock, Daniel L. (2021). Thoughts on African Swine Fever Vaccines. *Viruses*, *13*(5). https://doi.org/10.3390/v13050943

Rodriguez, F., Fernandez, A., Perez, J., Mulas, J. M. De, Sierra, M. A., & Jover, A. (1996). African swine fever : morphopathology of a viral haemorrhagic disease. *Veterinary Records*, 249–255.

Rodríguez, J., García-Escudero, R., Salas, M., & Andrés, G. (2004). African swine fever virus structural protein p54 is essential for the recruitment of envelope precursors to assembly sites. *Journal of Virology*, *78*(8), 4299–1313. https://doi.org/10.1128/jvi.78.8.4299-4313.2004

Rodríguez, J., Yáñez, R., Almazán, F., Viñuela, E., & Rodriguez, J. (1993). African swine fever virus encodes a CD2 homolog responsible for the adhesion of erythrocytes to infected cells. *Journal of Virology*, *67*(9), 5312–5320. https://doi.org/10.1128/JVI.67.9.5312-5320.1993

Rodríguez, Moreno, L. T., Alejo, A., Lacasta, A., Rodríguez, F., & Salas, M. L. (2015). Genome Sequence of African Swine Fever Virus BA71, the Virulent Parental Strain of the Nonpathogenic and Tissue-Culture Adapted BA71V. *PloS One*, *10*(11), e0142889. https://doi.org/10.1371/journal.pone.0142889

Rodríguez, & Salas. (2013). African swine fever virus transcription. *Virus Research*, *173*(1), 15–28. https://doi.org/10.1016/j.virusres.2012.09.014

Rodríguez, Salas, M., & Viñuela, E. (1996). Intermediate class of mRNAs in African swine fever virus. *Journal of Virology*, *70*(12), 8584–8589. https://doi.org/10.1128/JVI.70.12.8584-

8589.1996

Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods in Enzymology*, *383*(2003), 66–93. https://doi.org/10.1016/S0076-6879(04)83004-0

Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*, *5*(4), 725–738. https://doi.org/10.1038/nprot.2010.5

Roy, A., Yang, J., & Zhang, Y. (2012). COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research*, *40*(Web Server issue), W471-7. https://doi.org/10.1093/nar/gks372

Ruvalcaba-Salazar, O. K., del Carmen Ramírez-Estudillo, M., Montiel-Condado, D., Recillas-Targa, F., Vargas, M., & Hernández-Rivas, R. (2005). Recombinant and native Plasmodium falciparum TATA-binding-protein binds to a specific TATA box element in promoter regions. *Molecular and Biochemical Parasitology*, *140*(2), 183–196. https://doi.org/10.1016/j.molbiopara.2005.01.002

Sali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, *234*(3), 779–815. https://doi.org/10.1006/jmbi.1993.1626

Sánchez-Cordón, P. J., Chapman, D., Jabbar, T., Reis, A. L., Goatley, L., Netherton, C. L., Taylor, G., Montoya, M., & Dixon, L. (2017). Different routes and doses influence protection in pigs immunised with the naturally attenuated African swine fever virus isolate OURT88/3. *Antiviral Research*, *138*, 1–8. https://doi.org/10.1016/j.antiviral.2016.11.021

Sánchez-Vizcaíno, J. M., Mur, L., Bastos, A. D. S., Penrith, M. L., & Penrith, M. L. (2015). New insights into the role of ticks in African swine fever epidemiology. *Revue Scientifique et Technique (International Office of Epizootics)*, *34*(2), 503–511. https://doi.org/10.20506/rst.34.2.2375

Sánchez-Vizcaíno, J. M., Mur, L., Gomez-Villamandos, J. C., & Carrasco, L. (2015). An update on the epidemiology and pathology of African swine fever. *Journal of Comparative Pathology*, *152*(1), 9–21. https://doi.org/10.1016/j.jcpa.2014.09.003

Sánchez-Vizcaíno, J. M., Mur, L., & Martínez-López, B. (2012). African swine fever: an epidemiological update. *Transboundary and Emerging Diseases*, *59 Suppl 1*(SUPPL. 1), 27–35. https://doi.org/10.1111/j.1865-1682.2011.01293.x

Sánchez-Vizcaíno, J. M., Laddomada, A., & Arias, M. L. (2019). African Swine Fever Virus. In *Diseases of Swine* (pp. 443–452). Wiley. https://doi.org/10.1002/9781119350927.ch25

Sánchez, E. G., Pérez-Núñez, D., & Revilla, Y. (2017). Mechanisms of Entry and Endosomal Pathway of African Swine Fever Virus. *Vaccines*, *5*(4), 1–13. https://doi.org/10.3390/vaccines5040042

Sánchez, E. G., Riera, E., Nogal, M., Gallardo, C., Fernández, P., Bello-Morales, R., López-Guerrero, J. A., Chitko-McKown, C. G., Richt, J. A., & Revilla, Y. (2017). Phenotyping and susceptibility of established porcine cells lines to African Swine Fever Virus infection and viral production. *Scientific Reports*, *7*(1), 10369. https://doi.org/10.1038/s41598-017-09948-x

Savjani, K. T., Gajjar, A. K., & Savjani, J. K. (2012). Drug solubility: importance and enhancement techniques. *ISRN Pharmaceutics*, *2012*. https://doi.org/10.5402/2012/195727

Schulz-Gasch, T., & Stahl, M. (2004). Scoring functions for protein-ligand interactions: A critical perspective. *Drug Discovery Today: Technologies*, *1*(3), 231–239. https://doi.org/10.1016/j.ddtec.2004.08.004

Schulz, K., Staubach, C., & Blome, S. (2017). African and classical swine fever : similarities , differences and epidemiological consequences. *Veterinary Research*, 1–13. https://doi.org/10.1186/s13567-017-0490-x

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S.,

Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2019). Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins*, *87*(12), 1141–1148. https://doi.org/10.1002/prot.25834

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7

Shahbaaz, M., Nkaule, A., & Christoffels, A. (2019). Designing novel possible kinase inhibitor derivatives as therapeutics against Mycobacterium tuberculosis: An in silico study. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-40621-7

Shaikh, S. A., Ahmed, S. R., & Jayaram, B. (2004). A molecular thermodynamic view of DNA-drug interactions: a case study of 25 minor-groove binders. *Archives of Biochemistry and Biophysics*, *429*(1), 81–99. https://doi.org/10.1016/j.abb.2004.05.019

Shaikh, S. A., & Jayaram, B. (2007). A swift all-atom energy-based computational protocol to predict DNA-ligand binding affinity and ΔTm. *Journal of Medicinal Chemistry*, *50*(9), 2240–2244. https://doi.org/10.1021/jm060542c

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, *33*(Web Server issue), W244-8. https://doi.org/10.1093/nar/gki408

Spacková, N., Cheatham, T. E., Ryjácek, F., Lankas, F., Van Meervelt, L., Hobza, P., & Sponer, J. (2003). Molecular dynamics simulations and thermodynamics analysis of DNA-drug

complexes. Minor groove binding between 4',6-diamidino-2-phenylindole and DNA duplexes in solution. *Journal of the American Chemical Society*, *125*(7), 1759–1769. https://doi.org/10.1021/ja025660d

Spitzer, G. M., Wellenzohn, B., Markt, P., Kirchmair, J., Langer, T., & Liedl, K. R. (2009). Hydrogen-bonding patterns of minor groove-binder - DNA complexes reveal criteria for discovery of new scaffolds. *Journal of Chemical Information and Modeling*, *49*(4), 1063–1069. https://doi.org/10.1021/ci800455f

Spolar, R. S., Hat, J.-H., & Thomas Record, A. M. (1989). Hydrophobic effect in protein folding and other noncovalent processes involving proteins (heat capacity differences/hydrocarbon transfer processes). *Biochemistry*, *86*(November), 8382–8385. http://www.pnas.org/content/86/21/8382.full.pdf

Srivastava, H. K., Chourasia, M., Kumar, D., & Sastry, G. N. (2011). Comparison of computational methods to model DNA minor groove binders. *Journal of Chemical Information and Modeling*, *51*(3), 558–571. https://doi.org/10.1021/ci100474n

Staroverov, S. A., Pristensky, D. V, Yermilov, D. N., Gabalov, K. P., Zhemerichkin, D. A., Sidorkin, V. A., Shcherbakov, A. A., Shchyogolev, S. Y., & Dykman, L. A. (2008). The effectivity analysis of accumulation of liposomal, micellar, and water-soluble forms of diminazene in cells and in organs. *Drug Delivery*, *13*(5), 351–355. https://doi.org/10.1080/10717540500459084

Stefanovic, S., Windsor, M., Nagata, K., Inagaki, M., & Wileman, T. (2005). Vimentin rearrangement during African swine fever virus infection involves retrograde transport along microtubules and phosphorylation of vimentin by calcium calmodulin kinase II. *Journal of Virology*, *79*(18), 11766–11775. https://doi.org/10.1128/JVI.79.18.11766-11775.2005

Stefl, R., Wu, H., Ravindranathan, S., Sklenár, V., & Feigon, J. (2004). DNA A-tract bending in three dimensions: solving the dA4T4 vs. dT4A4 conundrum. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(5), 1177–1182.

https://doi.org/10.1073/pnas.0308143100

Stockert, J. C. (2014). Predictive binding geometry of ligands to DNA minor groove: isohelicity and hydrogen-bonding pattern. *Methods in Molecular Biology (Clifton, N.J.)*, *1094*(1094), 1–12. https://doi.org/10.1007/978-1-62703-706-8_1

Sur, J. H. (2019). How far can African swine fever spread? *Journal of Veterinary Science*, *20*(4), e41. https://doi.org/10.4142/jvs.2019.20.e41

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, *30*(12), 2725–2729. https://doi.org/10.1093/molbev/mst197

Tang, J., Fu, H., Kuang, Q., Zhang, L., Zhang, Q., Liu, Y., Ran, R., Gao, H., Zhang, Z., & He, Q. (2014). Liposomes co-modified with cholesterol anchored cleavable PEG and octaarginines for tumor targeted drug delivery. *Journal of Drug Targeting*, *22*(4), 313–326. https://doi.org/10.3109/1061186X.2013.875029

Teng, M. K., Usman, N., Frederick, C. A., & Wang, A. H.-J. (1988). The molecular structure of the complex of Hoechst 33258 and the DNA dodecamer d(CGCGAATTCGCG). *Nucleic Acids Research*, *16*(6), 2671–2690. https://doi.org/10.1093/nar/16.6.2671

Trent, J. O., Clark, G. R., Kumar, A., Wilson, W. D., Boykin, D. W., Hall, J. E., Tidwell, R. R., Blagburn, B. L., & Neidle, S. (1996). Targeting the minor groove of DNA: crystal structures of two complexes between furan derivatives of berenil and the DNA dodecamer d(CGCGAATTCGCG)2. *Journal of Medicinal Chemistry*, *39*(23), 4554–4562. https://doi.org/10.1021/jm9604484

Tsai, F. T. F., & Sigler, P. B. (2000). Structural basis of preinitiation complex assembly on human Pol II promoters. *EMBO Journal*, *19*(1), 25–36. https://doi.org/10.1093/emboj/19.1.25

Turlewicz-Podbielska, H., Kuriga, A., Niemyjski, R., Tarasiuk, G., & Pomorska-Mól, M. (2021). African Swine Fever Virus as a Difficult Opponent in the Fight for a Vaccine-Current Data.

Viruses, *13*(7), 1–20. https://doi.org/10.3390/v13071212

Utnik-Banaś, K., Schwarz, T., Szymanska, E. J., Bartlewski, P. M., & Satoła, Ł. (2022). Scrutinizing pork price volatility in the european union over the last decade. *Animals*, *12*(1). https://doi.org/10.3390/ani12010100

Vrancken, R., Haegeman, A., Paeshuyse, J., Puerstinger, G., Rozenski, J., Wright, M., Tignon, M., Le Potier, M. F., Neyts, J., & Koenen, F. (2009). Proof of concept for the reduction of classical swine fever infection in pigs by a novel viral polymerase inhibitor. *Journal of General Virology*, *90*(6), 1335–1342. https://doi.org/10.1099/vir.0.008839-0

Wang, X., Yang, M., Zhang, L., Yao, T., Chen, C., Mao, L., Wang, Y., & Wu, J. (2014). Design of novel bis-benzimidazole derivatives as DNA minor groove binding agents. *Chinese Chemical Letters*, *25*(4), 589–592. https://doi.org/10.1016/j.cclet.2014.01.019

Wang, Y., Kang, W., Yang, W., Zhang, J., Li, D., & Zheng, H. (2021). Structure of African Swine Fever Virus and Associated Molecular Mechanisms Underlying Infection and Immunosuppression: A Review. *Frontiers in Immunology*, *12*(September), 715582. https://doi.org/10.3389/fimmu.2021.715582

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189–1191. https://doi.org/10.1093/bioinformatics/btp033

Wiederstein, M., & Sippl, M. J. (2007). ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research*, *35*(SUPPL.2), 407–410. https://doi.org/10.1093/nar/gkm290

Wu, S., & Yang, Z. (2007). LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research*, *35*(10), 3375–3382. https://doi.org/10.1093/nar/gkm251

Wu, S., & Yang, Z. (2008). A comprehensive assessment of sequence-based and template- based methods for protein contact prediction. *Bioinformatics*, *24*(7), 924–931. https://doi.org/10.1093/bioinformatics/btn069.A

Xu, D., & Yang, Z. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function and Bioinformatics*, *80*(7), 1715–1735. https://doi.org/10.1002/prot.24065

Xu, D., & Yang, Z. (2013). Toward optimal fragment generations for ab initio protein structure assembly. *Proteins*, *81*(2), 229–239. https://doi.org/10.1002/prot.24179

Yakimovich, A., Huttunen, M., Zehnder, B., Coulter, L. J., Gould, V., Schneider, C., Kopf, M., McInnes, C. J., Greber, U. F., & Mercer, J. (2017). Inhibition of Poxvirus Gene Expression and Genome Replication by Bisbenzimide Derivatives. *Journal of Virology*, *91*(18), 1–15. https://doi.org/10.1128/jvi.00838-17

Yáñez, R., Rodríguez, J., Nogal, M., Yuste, L., Enríquez, C., Rodriguez, J., & Viñuela, E. (1995). Analysis of the Complete Nucleotide Sequence of African Swine Fever Virus. *Virology*, *208*, 249–278. https://doi.org/10.1006/viro.1995.1149

Yang, Zhang. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, *9*, 40. https://doi.org/10.1186/1471-2105-9-40

Yang, Zhang, Hubner, I. A., Arakaki, A. K., Shakhnovich, E., & Skolnick, J. (2006). On the origin and highly likely completeness of single-domain protein structures. *Proceedings of the National Academy of Sciences*, *103*(8), 2605–2610. https://doi.org/10.1073/pnas.0509379103

Yang, Zhang, & Skolnick, J. (2004). SPICKER : A Clustering Approach to Identify Near-Native Protein Folds.

Yang, Zhang, & Skolnick, J. (2005a). The protein structure prediction problem could be solved using the current PDB library. *102*(4), 1029–1034.

Yang, Zhang, & Skolnick, J. (2005b). TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–2309. https://doi.org/10.1093/nar/gki524

Yang, Zhilong, Martens, C. A., Bruno, D. P., Porcella, S. F., & Moss, B. (2012). Pervasive Initiation and 3′-End Formation of Poxvirus Postreplicative RNAs. *Journal of Biological Chemistry*, *287*(37), 31050–31060. https://doi.org/10.1074/jbc.M112.390054

Yoon, C., Privé, G. G., Goodsell, D. S., & Dickerson, R. E. (1988). Structure of an alternating-B DNA helix and its relationship to A-tract DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(17), 6332–6336. https://doi.org/10.1073/pnas.85.17.6332

Zakrzewska, K., Lavery, R., & Pullman, B. (1983). Theoretical studies of the selective binding to DNA of two non-intercalating ligands: netropsin and SN 18071. *Nucleic Acids Research*, *11*(24), 8825–8839. https://doi.org/10.1093/nar/11.24.8825

Zhang, Y., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, *57*(4), 702–710. https://doi.org/10.1002/prot.20264

Ziem, B., Rahn, J., Donskyi, I., Silberreis, K., Cuellar, L., Dernedde, J., Keil, G., Mettenleiter, T. C., & Haag, R. (2017). Polyvalent 2D Entry Inhibitors for Pseudorabies and African Swine Fever Virus. *Macromolecular Bioscience*, *17*(6), 1–9. https://doi.org/10.1002/mabi.201600499

Zimmermann, L., Stephens, A., Nam, S. Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A. N., & Alva, V. (2018). A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *Journal of Molecular Biology*, *430*(15), 2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007

Zsak, L., Caler, E., Lu, Z., Kutish, G. F., Neilan, J. G., & Rock, D. L. (1998). A nonessential African swine fever virus gene UK is a significant virulence determinant in domestic swine. *Journal of Virology*, *72*(2), 1028–1035.

APPENDICES

Appendix I: Uncharacterized proteins of ASFV Ba71V

>sp|Q65134|VF69R_ASFB7 Uncharacterized membrane protein X69R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=BA71V-018 PE=3 SV=1
MLLYIVIIVACIISKLVPNEYWAIHLFFIIMIFMVYMYEKLDIHQKYQFWNYTMSGLSGH
NVQVICKCY
>sp|Q65180|VF312_ASFB7 Uncharacterized protein CP312R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-96 PE=3 SV=1
MLLVKMTTHIFHADDLLQALQQAKAEKNFSSVFSLDWDKLRTAKRNTTVKYVTVNVIVKG
KKAPLMFNFQNEKHVGTIPPSTDEEVIRMNAENPKFLVKKRDRDPCLQFNKYKISPPLED
DGLTVKKNEQGEEIYPGDEEKSKLFQIIELLEEAFEDAVQKGPEAMKTKHVIKLIQRKIS
NSAVKNADKPLPNPIARIRIKINPATSILTPILLDKNKPITLQNGKTSFEELKDEDGVKA
NPDNIHKLIESHSMHDGIINARSICISNMGISFPLCLEMGVVKVFEKNNGIDVNSIYGSD
DISTLVNQIAIA
>sp|Q89907|VF339_ASFB7 Uncharacterized protein D339L OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-105 PE=3 SV=1
MIDQKIFETTLNIDDPTNFCTNVEAHLLKELENIYVGKCFKNSFILNITGVIQRSPCFIM
RTNNSGRGYMHVRFSAVVSYLNAFDLIAAVKIIKNDSNIILGESLLTEPVTIVIPSSESQ
NNVAEVGQIVPVQLANSSVYYIPGRQQASATGSIFIPKHTFSVYHVQEELTQEQALNLTK
LVNIIEMLLESRSKKDFKQICFFEKLYYTYSISSDEILDLKIWKGPKGKEMSRLKPCNVL
SFLYDALKNKSSSLGFWARPPNLLKSSPLAYQQDQNSFNATELPIICSAEVMFVTLLKEI
INYLQFMNDLCDTFNNEQLIKRHENIWMLIEQRKIGHDF
>sp|Q65153|VF448_ASFB7 Uncharacterized protein M448R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-061 PE=3 SV=1
MSNESFPETLENLLSTLQTKQQNAIQSEVIEWLHSFCETFHLKIHCHKQFIPSGEKKRAK
IPAQEIQGNTQPSHHVHRVVLSRAQPVKAQESLLTTMCNGLVLDANTWTCLAIPPPAPFQ
QATRQVQHFYRNNFYEVVPIQDGTLLTIYYWDDPEHGPSWCLASTHGYDVSNYCWIGDKT
FAELVYELLQQHSTCDVTLEKNKTRGTRLFFNNLNPDYCYTIGIRHHNLQPLIYDPQNIW
AIQSTNLKTLKTVYPEYYGYIGIPGIQSQVPELPQYDLPYLIRSYKTAMNQAKNAIKNGK
KDKEYFNYGYLLISRAPAITKSTSNVLLKSPLLVFLQKSVYQKKHNISNSQRLEFIILQN
YLMQHFRDHFIALFPQYISYYTKYQNMLNMIIHSIATKDKDHPFAGAVVKKVLEDIENAE
NIIDHTTIQNYAHQSKYAMLYLSIISHF
>sp|A0A097SRV6|L57L_ASFB7 Uncharacterized protein L57L OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=L57L PE=4 SV=1
MDDTLPKQMTPTDTSPLKEEQAHCNNKTLENQPKNINDNKCTDSQNTDLQNTEPSKV

>sp|Q65156|VF717_ASFB7 Uncharacterized protein C717R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-064 PE=3 SV=1
MTKLAQWMFEQYVKDLNLKNRGSPSFRKWLTLQPSLLRYSGVMRANAFDILKYGYPMQQS
GYTVATLEIHFKNIRSSFANIYWNRDSEEPEYVCCCATYQSHDGEYRYRFVWYQPFIEAY

NAIEAALDPLETIILNLIAARDLDFVVHIFPYNKGHEDYLASTQLILKIFIATLLMDILR
IKDNTLDVHLNSDYIIVMERLWPHIKDAIEHFFEAHKDLLGYLIAFRNGGNFAGSLRPSC
GQKIVPLTIREVLQMNDINLAVWREVFIMQECSDLVINGIAPCFPIFNTWTYLQGINQIF
FENTSLQEKFKKDFIARELSKEIIKGQKTLNDKEFKKLSLHQIQYMESFLLMSDVAIMIT
TEYVGYTLQSLPGIISRSSYLSPIVKNILMDEDSFMSLLFDLCYGAYVLHKKENVIHADL
HLNNMTYYHFNPTSFTDRNKPGKYTLKVKNPVIAFITGPKVETETYVFKHIDGFGCIIDF
SRAIMGPNHAIKLERQYGLAFVNTFYRNQSEHILKVLRYYFPEMLTNRENEIQGVILSNF
NFFFNSITAIDFYAIARNLRSMLSLDYLHTSEVKRNVEISQTFLDTCQFLEEKAVEFLFK
NLHTVLSGKPVEKTAGDVLLPIVFKKFLYPNIPKNILRSFTVIDVYNYNNIKRYSGKAIQ
TFPPWAQTKEILTHAEGRTFEDIFPRGELVFKKAYAENNHLDKILQRIREQLANENL
>sp|Q65213|VF96_ASFB7 Uncharacterized protein DP96R OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-155
PE=3 SV=1
MSTHDCFSKEKPVDMNDISEKSSVVDNAPEKPAGANHIPEKSAREMTSSEWIAEYWKGIK
RGNDVPCCCPRKMTSADKKFSVFGKGSLIRSIQKNN
>sp|Q65190|VF240_ASFB7 Uncharacterized protein H240R OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-120
PE=3 SV=1
MAVNIIATRAAPKMASKKEHQYCLLDSQEKRHGHYPFSFELKPYGQTGANIIGVQGSLTH
VIKMTVFPFMIPFPLQKTHIDDFIGGRIYLFFKELDMQAVSDVNGMQYHFEFKVVPVSPN
QVELLPVNNKYKFTYAIPVVQYLTPIFYDLSGPLDFPLDTLSVHVDSLSNHIQLPIQNHN
LTTGDRVFISGYKHLQTIELCKNNKIFIKYIPPLSSEKIKLYIPKNRIRIPLYFKSLKNV
>sp|Q65157|VF122_ASFB7 Uncharacterized protein C122R OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-065
PE=3 SV=3
MKICKACSSCMVRTYVDGNIIFRCSCGESVQGDSQNLLVSSKVYHTGEMEDKYKIFIKNA
PFDPTNCQIKKDCPNCHLDYLTQICIGSQKIIILVCRCGYMSNRG

>sp|Q65182|VF345_ASFB7 Uncharacterized protein D345L OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-109
PE=3 SV=1
METFVRLFKDSPQQRSDAWHAIRRTQVGGSDLASVLGLNPYKSYYIILAEKANLFKKNLN
RAACSWGTLFERVSKDLLELFCQTTVIGDNIHIDGTYLGYPGHSNSPDGFCHLTLGYTQQ
SWEIKTIFNNVRYEATKRIPVLVEIKSPFNRKIKNSVPSYYMPQIQSGLALSPPISMGIY
VEAMFRVCGIHQLGSNNETNTDIHPPESMLPLAWGIITICSTQEHTEAPQDFGTLDAETF
RQLLETLYQKDQYTIHYSMPYETACPEMPNVVGYFGWKVFIFQIIPVMKHPQFLKDKYPI
IQQFLRDLHTIKASPSPMETYEKICCSEESALSTEDIDNFTDMLT
>sp|Q65149|VF152_ASFB7 Uncharacterized protein EP152R OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-056
PE=3 SV=1
MYSILIACLVLLLCLVIYVGHRADHARKYLEGMWHGDPVFLKQSGLQSFYLYIQPGHTCF
FSIVNKNGEKLMETKIPCTITNKIYMFFKPIFEFHVVMEDIHRYLPKQFNFLLDSAEGKL
ILENNHVIYAVLYKDNFATALGKTVEKYITQN
>sp|Q65197|VF146_ASFB7 Uncharacterized protein E146L OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-129
PE=3 SV=1
MGGTTDFVLSITIVLVILIIIAFIWYNFTGWSPFKYSKGNTVTFKTPDESSIAYMRFRNC
IFTFTDPKGSLHSIDVTEVLNNMAKGFRDAQNPPSSFTLGGHCQAPLNAFSFVLPGVNDR
ATVATADDAKKWENCDATLTGLQRII

>sp|Q65174|VF175_ASFB7 Uncharacterized protein B175L OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-085 PE=3 SV=1
METNCPNILYLSGITIEECLQTKKTATDTLNTNDDEAEVEKKLPSVFTTVSKWVTHSSFK
CWTCHLYFKTVPKFVPTYMRENERGEIEMGVLGNFCSFSCAASYVDVHYTEPKRWEAREL
LNMLYRFFTSQWISYIKPAPSYTMRKEYGGKLSEEAFISELHTLEESISSKHIFI

>sp|Q65195|VF423_ASFB7 Uncharacterized protein E423R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-127 PE=3 SV=1
MLWRNEITEFMDQLSKYSQEILKTFKQLRPSEYKQYNEFLTQVTPLLQKTSEKIPELVDH
IFNYLDNVEKICELLVNASSIIISSKIREQVKHGMSFSYKADLDSLADILSQKQYVLMHL
SKNIAAEYFNTCLNQGKSKLDLKAASVFYSSRPRTASSAELYRKMLYAYGSPQEINYYTE
KARNKTLDVEESDSMAIIERTARHNLSLMHPLEAMGLTFGATNTDADPEDLKDKTVINLT
LPQATESITYHLKSLMQLKKVSTASGLNTNILKAFDNIISTPVKKNKMASKLAPGMDVVF
TSDNGKTFFTKNILSKNMLAGPKERVFAYNNLISNLNNSCFIQNHNDFLRQQDSWPFYDA
HNFTNKFLMQPIFSGQTRPRLQGAMEAAHVETHLTAFLQSIQPSRPQDPSVLASPKLSAL
ILN
>sp|Q65205|VF267_ASFB7 Uncharacterized protein I267L OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-137 PE=3 SV=1
MLLVLIDVDGFMGQLYNENGTQTILIPREVVIFYWEKNTASKILQLFFHGGIDPIFEKIN
QRSFSFQSRHIHHFTLDESPLPNSIALPTDTLQAFKAGKKMIFQHLVKITKDHEQILLLH
KGGPEGEWVRSFNIPNATVQNLNDLCCPSVEKLVLKKRDYISSSIGCPKHIQGSNHCPVF
ECHVLFKWIQENTSIVQGVLKRPSLPYEEAVLFIEHRINMVDNHPFKKDSVKQNQKKKNW
IATQFVQHGIYVDNGILSKIYNKYSLF

>sp|A0A097SRX8|J64R_ASFB7 Uncharacterized protein J64R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=J64R PE=4 SV=1
MLLYIVIIVAYVSYKLVPKQYWPILMFMAYMVYTHEKLDINERSGFWKYIIAKLFRCHGC
EICK
>sp|Q89667|KP86R_ASFB7 Uncharacterized protein KP86R/DP86L OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=BA71V-001 PE=4 SV=1
MGINQSQHSMGVNQSHHFMGVNQSHHFMGINQSQHSMGVNQHIIGDSYLISCMVVAPCCL
KACMVLAPCCLKACMVLAPCCLKAAR
>sp|Q89768|5054R_ASFB7 Protein MGF 505-4R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=BA71R-027 PE=3 SV=1
MFSLQDICRKYLFQLPDSFDEYTLQVLGLYWEKHGSLQRIRKDAVFVQRNLIISINEALR
IAASEGNGRVVKLLLSWEGNFHYVIIGALEGDHYDLIHKYGSQIEDYHMILSSIHNANTF
EKCHELSNCDMWCLIQNAIKYNMLPILQKHRNILTHEGENQELFEMACEEQKYDIVLWIG
QTLMLNEPEFIFDIAFERIDFSLLTMGYSLLFNNKMSSIDIHDEEDLISLLTEHLEKAAT
KGCFFFMLETLKHGGNVNMAVLSKAVEYNHRKILDYFIRQKCLSRKDIEKLLLVAISNSA
SKKTLNLLLSYLNHSVKNIIGKIVQSVLKNGDFTIIIFLKKKKINLVEPALIGFINYYYS
YCFLEQFIHEFDIRPEKMIKMAARKGKLNMIIEFLNEKYVHKDDLGAIFKFLKNLVCTMK
HKKGKETLIVLIHKIYQVIQLETKEKFKLLRFYVMHDATIQFISMYKDCFNLAGFKPFLL
ECLDIAIKKNYPDMIRNIETLLKCE

>sp|Q65214|DP60R_ASFB7 Uncharacterized protein DP60R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=BA71V-158 PE=4 SV=1
MSSIWPPQKKVFTVGFITGGVTPVMVSFVWPAAQPQKKINYSRKKKYFRPRSFYKKNVSF
>sp|Q65200|E248R_ASFB7 Protein E248R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-132 PE=1 SV=1
MGGSTSKNSFKNTTNIISNSIFNQMQNCISMLDGKNYIGVFGDGNILNHVFQDLNLSLDT
SCVQKHVNEENFITNLSNQITQNLKDQEVALTQWMDAGHHDQKTDIEENIKVNLTTTLIQ
NCVSSLSGMNVLVVKGNGNIVENATQKQSQQIISNCLQGSKQAIDTTTGITNTVNQYSHY
TSKNFFDFIADAISAVFKNIMVAAVVIVLIIVGFIAVFYFLHSRHRHEEEEEAEPLISNK
VLKNAAVS
>sp|Q65193|VF184_ASFB7 Uncharacterized protein E184L OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-125 PE=3 SV=1
MKTFITCTSVKNYFRQHLKTNQRISSELISYVCTILNHICHQYLQNPQAQEEEWFALIKE
LPIIKDGLSKEERFFSSGVKHFLHEYKITPENQEKFQKMLNAITEQLMSRLCKVFSIMIQ
RQGFLKTQTLMYSHLFTILSILMVADNLYGEQDPTEFFSLIIEQTKTIKKKKKSSSEEEE
SHEE

>sp|Q65160|VF315_ASFB7 Uncharacterized protein C315R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-068 PE=3 SV=1
MDALLKEIEKLSQPSLQKENNDVCDLCFMQMKKISNYQLLCEECGQLKDWFEPEYNEKFT
VYSRLKIVGANSSYHQRDLDKANSSDYSSLQFHHILEELKSLNVKYMDAGQKPFPIQVLK
ETAHSYNQVQQHRVIRSITKLQILASILRSICLKLNIACTVADAARFTQLNTKGISRGMD
LLRSLFVDNKITLNVDLNPIDSFINSTYSALQIKQIHQELQEENVYNLKEIVKSFILYAD
EKNIGVDLNRRTVVIATMYNVLRRAYYPIEIDTVVYQCKIRKNTITRALKMYEDYYSHFK
SLYEQYHLNAAKKLI

>sp|Q65144|VF317_ASFB7 Uncharacterized protein F317L OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-043 PE=3 SV=1
MVETQMDKLGFLLNHIGKQVTTKVLSNAHITQTMKEIILENHGVDGGAAKNVSKGKSSPK
EKKHWTEFESWEQLSKSKRSFKEYWAERNEIVNTLLLNWDNVRGAIKKFLDDDREWCGRI
NMINGVPEIVEIIPSPYRAGENIYFGSEAMMPADIYSRVANKPAMFVFHTHPNLGSCCGG
MPSICDISTTLRYLLMGWTAGHLIISSNQVGMLTVDKRIIVDLWANENPRWLMAQKILDI
FMMLTSRRSLVNPWTLRDLKKILQDYGIEYIIFPSNDFFIYEDERLLMFSKKWTNFFTLH
ELLDDLETIETKASSTT
>sp|Q07384|VF421_ASFB7 Uncharacterized protein K421R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-052 PE=3 SV=1
MYTHVDVVGIAEASAALYVQKDRDRYLDVLTTIENFIYQHKCIITGESAHLLFLKKNIYL
YEFYSNNVAEHSKALATLLYKLDPEYLTRYTVLITKIPNHWYVINVDQREFVRLYAIPAV
KQHLPIPILPFYCTSALTHQELFCLGPELQLIQIYSKLCNPNFVEEWPTLLDYEKSMRML
FLEQFPRRLEIAGGKKEEEKHESIIKKIILEMVSTRQRIVVGGYIQKNLYNHVLKNRNRL
QLITSLDIYEEKDIIQQFCDSNGLKIKIRINNPLLPTNPELRRLTIYFNHNNDDDQSYLI
VDMYNTGSYELVPTNQINTLDGSFLIGTPFVQARFLLVEIWVLMLIAQQTKKDTKKIIQF
FINQYEMLMNSPWPSMEALFPSSSKRYLGNYVDPNALIKWAQLKLKRIPPFYPGKPDEES
C

>sp|Q65161|VF62_ASFB7 Uncharacterized protein C62L OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-070 PE=3 SV=1
MNWGSISSGTPGLFVESIRNTPSVVKINVIFLKVISNTAVSVFWRDRRIRFESDWLNSYF
QK

>sp|Q65207|VF63R_ASFB7 Uncharacterized protein DP63R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=BA71V-149 PE=3 SV=1
MPLKRQGRLLGKKKRVLQHILGLNIFKRELIPPCKDPDPYQIQILLKNYILKNVSTVFTY
YCQ
>sp|Q65171|VF125_ASFB7 Uncharacterized protein B125R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-082 PE=3 SV=1
MAVYAKDLDNNKELNQKLINDQLKIIDTLLLAEKKNFLVYELPAPFDFSSGDPLASQRDI
YYAIIKSLEERGFTVKICMKGDRALLFITWKKIQSIEINKKEEYLRMHFIQDEEKAFYCK
FLESR
>sp|Q65185|VF171_ASFB7 Uncharacterized protein H171R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-114 PE=3 SV=1
MVVYDLLVSLSKESIDVLRFVEANLAAFNQQYIFFNIQRKNSITTPLLITPQQEKISQIV
EFLMDEYNKNNRRPSGPPREQPMHPLLPYQQSSDEQPMMPYQQPPGNDDQPYEQIYHKKH
ASQQVNTELNDYYQHILALGDEDKGMDSMLKLPEKAKRGSDDEDDMFSIKN
>sp|Q65145|VF165_ASFB7 Uncharacterized protein F165R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-046 PE=3 SV=1
MANPNKRIMNKKSKQASISSILNFFFFYIMEYFVAVDNETPLGVFTSIEQCEETMKQYPG
LHYVVFKYTCPADAENTDVVYLIPSLTLHTPMFVDHCPNRTKQARHVLKKINLVFEEESI
ENWKVSVNTVFPHVHNRLSAPKFSIDEANEAVEKFLIQAGRLMSL
>sp|Q65175|VF263_ASFB7 Uncharacterized protein B263R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-086 PE=3 SV=1
MEDETELCFRSNKVTRLEMFVCTYGGKITSLACSHMELIKMLQIAEPVKALNCNFGHQCL
PGYESLIKTPKKTKNMLRRPRKTEGDGTCFNSAIEASILFKDKMYKLKCFPSTGEIQVPG
VIFPDFEDGKNIIQQWVDFLQHQPIEKKIQIIEFKTIMINFKFQINPVSPRVIIHLKKFA
ALLEHIPTPYPIREIKPPLEDSKVSAKFMVSPGKKVRINVFLKGKINILGCNTKESAEII
YTFLKDLISVHWQEILCVLPVPD
>sp|Q65168|VF354_ASFB7 Uncharacterized protein B354L OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-078 PE=3 SV=1
MALTTHSGKLIPELQFKAHHFIDKTTVLYGPSKTGKTVYVKHIMKILQPHIEQILVVAPS
EPSNRSYEGFVHPTLIHYRLWLADKQKKNDNKGAERFLEAIWQRQTMMSSIYSRVNNIDM
LKTLYHKLPIDIQQKENKNIAKVECLKAEQTDQKKEEKITSLYQQLLKKIIIQNIHMYKN
LCLTEDEKFTLNYINLNPRLLLILDDCAAELHPLFTKEIFKKFFYQNRHCFISMIICCQD
DTDLPANLRKNAFVSIFTNASICMSNFSRQSNRYSKQDKEYVEEISHIVFKGYRKLVYIR
EDENRQHFYHSTVPLPTAFSFGSKALLKLCKAVYSKEVVIDKSNPYWSKFRLNF
>sp|Q65167|VF475_ASFB7 Uncharacterized protein B475L OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-077 PE=3 SV=1
MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLY

EKMLDSFYKYFCNITQINSKTLNFSNFITTIVDSFIPKEYSQSISLEKKESILELLLCDY
ISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEP
ETELEETYARLSSYNRSLLHQIEELTSEKKSLLADLSTLRKKYEKRQSEYRRLVQLLYQQ
IQRSSTSKSSYPLTKFIETLPSEHFSNEEYQKETPADQKEVVEMELLRKQELLTSQELTS
KSPNNYPVPHSRTIVSKPPDNYPVPRSRTTTKLDFDNSLQNQELHTKNGFSEKDIVEFGQ
DKPEEENILAIDQDKPEEENILAIKQDIPEEENILAIDQDKPEFNQDTPEFKEAVLDTKE
NILEEENQDEPIVQNPFLENFWKPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
>sp|Q65203|VF111_ASFB7 Uncharacterized protein E111R OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-135
PE=3 SV=1
MSFSECPLVISACKKFLQKRITIENEALINALITALAQTSTLNDLCLLPIQTYLLSYKNA
FEWIHFVCIAITTILDNKYNWKNCTVDINYIFLHVTYIYNIKTKEYLDYCS
>sp|Q65204|VFE66_ASFB7 Uncharacterized protein E66L OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-136
PE=3 SV=1
MLTMVSLARFIINTASFNKSIYPLSQVYVKIHLSHKYNHILFTYNMRIYLIKRNHMLFTH
MLISCN
>sp|P27946|VF73R_ASFB7 Uncharacterized protein I73R OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=BA71V-141
PE=3 SV=1
METQKLISMVKEALEKYQYPLTAKNIKVVIQKEYNVVLPTGSINSILYSNSELFEKIDKT
NTIYPPLWIRKTN
>sp|Q65147|VF205_ASFB7 Uncharacterized protein K205R OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-048
PE=3 SV=1
MVEPREQFFQDLLSAVDQQMDTVKNDIKDIMKEKTSFMVSFENFIERYDTMEKNIQDLQN
KYEEMAANLMTVMTDTKIQLGAIIAQLEILMINGTPLPAKKTTIKEAMPLPSSNTNNDQT
SPPASGKTSETPKKNPTNAMFFTRSEWASSKTFREKFLTPEIQAILDEQFANKTGIERLH
AEGLYMWRTQFSDEQKKMVKEMMKK
>sp|Q65177|VF134_ASFB7 Uncharacterized protein G1340L OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-089
PE=3 SV=1
MDFQNDFLTNPLRVTLYNPAENEYTKTFIFLGSVPANVLQACRKDLQRTPKDKEILQNFY
GKDWEKKLSQYVVGGDSDDLDEFEKLFVEDSGEETNVMMPEIETMYSEYSIFPEDTFKDI
REKIYVATGIPPYRQHIFFFQNNALQVTYRLLLSGSGVALDIRDYKKEFQQVGGLNIDAS
MESQKDELYVEALDSFQLIKNIHHIFVADLNTLVAPMRRQISIAMEDNYQFDLLYYGLIM
KYWPLLSPDAFKLLVQSPLQMEKQYPALSPSLTSLKKRLLLEQKLINFTYARAQQVIAKY
EGNRLTRGTLAVTSAMIKISPLVNIQINVRNVFDLFPATPDIPQLVVFFYSKTGPTVVSK
HHITSTEPEKFSNKTFRVPTIILIRFINKKAFILTIQNNGHYFIESNWSENERHDFNSVV
STLNNFINPIIHTINDMGPAAFPRGGSLPLPSNEDIQISISSMSVSTFWPYTLSSKGFTE
LKSRWREYEQAGIISVRGLQQTGIYNFLFKKGIYSYDPHEIERMIIISSGPGRKMDINVA
LLQNTYAYLFDANVAARWETIYGGRNIRIYHRVTDIKIEMFNITQEEFNYLWVYLFVFLD
NLITGPDKILVNKLSQLHDKQQGKGASQLRALQEQDPDLYDLRKYDTQATVYSVLCQHPR
PPVIYSEAEVKSMPPAKRKELVKYWNFTEGVPAYYSCPHPDYPHLSLLEGRHPLNYCLPC
CQKTKALLGTKRFYINNTCLTKHTFVEQDLEDLNTQTSRHTLSYGKKIPVNRIAFLPHQI
ADELFLNTIKEPDIFCIVGVEQTMLGISNAGLFYSLARILDLAPKALAIEIAKAANTPQY
YILGNGAGNMFSSGAELANLILQTFVEQKNQLLQWDTTWQDIFLDLVAICYDLHCVFFKD
KQGDIEFEVSPSTIQKILSPSKKIAIIFDTDEGIYPMAITQQKRFLKNSEAQYIFTEDDP
VMEVVQSMSEFMCKDNWWDIHDVKNIPGYTVGKKLINRHNFCYALLIDSDTDRPIYFPIR
LSSYIHDDIPIDFDLRPTQIASFEETWKFITLFNKQYKQYEIVPSAVLQNIKKEFVGFLS

EGKTGLYFYYAPTQTLPATLEKLPIATLTIDPRDIDQAILYPLEEPYPQQNKANKAFYIN
HLYKFLLIEFFDVLYGLQSNSTRKHIENLFQKTDFQKITSVTEFYTKLSDFVDLNDIHTI
KHILETTDAEHALKVLQKNIFNFDYTLLSPLQSYTYDELCQHLKKLLTPRIEFYEDIETI
DRGLINIYTSCQYSTLNQPQCKKKRLRIPVNHFENYIHILAADILNPLKHSTLLLTGLGV
IDDLQFILRPQEIISVKNKF
>sp|Q65155|VF84L_ASFB7 Uncharacterized protein C84L OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-063
PE=3 SV=1
MLLFIWIRNNFSISYIPLIYSLLQKTTKKKENRFFIRSGKKIQRIQMMMFITVYDINQKQ
KKRYGLRGCNLNLKATVLLLHKRI
>sp|Q89675|VF79_ASFB7 Uncharacterized protein D79L OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-104
PE=3 SV=1
MNKTIEYQKEFLKENNQLLSIPVKKNILKEILQNDEQDTIITNCITKEVSINLDLIKNPK
VLYSIYIMVVEYLKSINIA
>sp|Q65189|VF233_ASFB7 Uncharacterized protein H233R OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-118
PE=3 SV=1
MILIASPFSLAHLEYLHTWHVTIKNIAQQHGLDIKVAIVVSTSHLNNFLPISGALNIECI
TFPSCGIKEIDLLWARIKLFQHYCAIGARLLWLVSADIRPPVSAWPAIADSLKKGADAVV
IPYPSRWNNLIPTVIKEIVVHQKKCLVAVDARHLDTDTQIVGAGMGCIVLTLKALMVRLS
IGKQPVKILWPDLHGTAEGIPLEGVEVGWFLNAYAHKLNIRCLGADHIAQHLT
>sp|Q65183|VF183_ASFB7 Uncharacterized protein S183L OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-110
PE=3 SV=1
MSVVVGGVEYSLNNWARYEIKRRAAELESVNYYPHCEYIMPEDIVVSILGSKPNCPFLEA
LKRFHDFLKKRRIIFKGEYLVIPWMGAQDVADMIHHVENRINLDHLEDLAHMLKLITYHK
SFDTCINQAFEHLYAFKFPDANIETHELKHIRQLEKKMYGYILRLEKLQTVLTFYIEFLL
KQV
>sp|Q65196|VF301_ASFB7 Uncharacterized protein E301R OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-128
PE=3 SV=1
MSEDIRRGPGRPPKKRVVPNFERKGILEKPVRPQSRLEFSYDNPLIFKNLFIYFKNLKSK
NILVRCTPTEITFFSRDQSQASFVIATIDGKNVNHYYASDVFWLGINRELVEKMFNSIDR
SFLKITIVHRYDKPETLFFIFTDFDIDKECTYQITVSEPELDMDLIEMEKSISEERLKNY
PLRWEFTSKQLKKTFSDLSNYTELVTIEKLGGDTPLHLYFQKFNSISYHEMYKSSNKINL
TSTIPKSQVFQINVKIAHIKSLASAMVTDKIRILCEENGNLIFQSEMDALMLNTITLNNT
I
>sp|Q65139|VF118_ASFB7 Uncharacterized protein A118R OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-035
PE=4 SV=1
MHSNAFFNLIACVLFPTPLIPSMVISIPRMINKWVKRVQFLTFLTNLFLYNIVQHYINRI
RCYSFIKYLLLYNLYRPIFGRSLQMAITKIKIISDATAAVLLKSCAAMYDVLIDKKFK
>sp|Q07385|VF145_ASFB7 Uncharacterized protein K145R OS=African swine
fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-051
PE=3 SV=1
MDHYLKKLQDIYTKLEGHPFLFSPSKTNEKEFITLLNQALASTQLYRSIQQLFLTMYKLD
PIGFINYIKTSKQEYLCLLINPKLVTKFLKITSFKIYINFRLKTFYISPNKYNNFYTAPS
EEKTNHLLKEEKTWAKIVEEGGEES

>sp|Q65172|VF117_ASFB7 Uncharacterized protein B117L OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-083 PE=3 SV=1
MGYTIQLDKDGDYCWDEDPTHHDPYMQANATSHVATSYATTSHAAVAAPHAAAHHTFHEP
FIKLNLTDKNIFNGLGFILIVIFIYLLLITLQQMLTRHIYNTVQHCVKAHLDSKNLQ
>sp|Q65206|VFD38_ASFB7 Uncharacterized protein DP238L OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=BA71V-147 PE=3 SV=1
MARGQNIRKRTFSDMDTPSDKNIGIHTNSLPKNNLYRRILFKGKISNYSISKDSLAKDHS
SKHSISKNGLIGKKRPAPLDISFQSMNSSISSSTQKKTRILDEEIKDQSLSNENDTDSPV
IVDITLKPSYMSKTSRITEIIHKMKELNMNRIEDGSSFNKKRSEHDDKNILLHTMEMEEE
DCEIEEDIAIDSPYLNTSLSEDDTDSIVGTDYSEKEKETISETESSSDDESYSLYDSF

>sp|Q65173|VF407_ASFB7 Uncharacterized protein B407L OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-084 PE=3 SV=1
MEDTTFLEGANLAGITTLMNNLHINEQANLEELEKQVMGKQQSFPTDHFDEELNGLAKSL
GINFNDPEFSLDSPHSVISKKPSGRGRDKVHGGIRRDSVCTDSICSDSVCSGSIRSGSIR
SGSIRNGSIRSGSVRDDSVRSGKTRRGLACNSSSRNDRGYSLSTHRKKYAESEASQKTAF
SKRDRKNHYAESEYSEKSIKPSTKQVDRLINHLRSNGDPNSFYKKDHDYERKTKLVKLEK
INMLLTYLGNEQISTDDIKIPTIDSSMQEIDDVIEMLTLRNVGIRYSSIAEEILIGLARG
LEIVFDGTREIPFLNYRPDYTGLHNTFMIKLFKMRYETSQVVGNLVQNMSPLSKICLELG
PSLLLYPALIRTKHKASEDLYNLLQKGPEDPFTAYNEIHETLKKNNK
>sp|Q65178|VF123_ASFB7 Uncharacterized protein CP123L OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-91 PE=3 SV=1
MPSTGTLVIIFAIVLILCIMLLFFYKTVEAEKPGVLPPPIPPPTPPPSKKKYDHNEYMEK
TDLEPEVKKNHRKWANEAEHLISSSVKGLENLDETAFLANHKGHGFRTFEHAKSLFKEFL
KKY
>sp|Q65154|VF129_ASFB7 Uncharacterized protein C129R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-062 PE=3 SV=1
MEHPSTNYTPEQQHEKLKYYVLIPKHLWSYIKYGTHVRYYTTQNVFRVGGFVLQNPYEAV
IKNEVKTAIRLQNSFNTKAKGHVTWAVPYDNISKLYAKPDAIMLTIQENVEKALHALNQN
VLTLASKIR
>sp|Q65186|VF124_ASFB7 Uncharacterized protein H124R OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-115 PE=3 SV=1
MNLEYVQVVQKFNQVLLELTKKVCTVVGGSKPTYWYHHIRRVCSECPSMPMSMIGPYLNV
YKAQILTRDKNFFMNFDPAHNEYTFIIQKLKEAARNMPEDELEQYWVKLLFLLKSYIKCK
PFIN
>sp|Q89445|VFD29_ASFB7 Uncharacterized protein D129L OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-103 PE=3 SV=1
MDINPLLYLQAFNNDATTFNTQGHILEQQSDSPYFDTFANAMQAYLDTKQGGNDEEGTII
LMDDEDFNDSESLEDFLQMLSEEELNDGFSSDDEPEEHVILTEDNQGEPSETPQATFDIT
EFIKIDDED
>sp|Q65152|VF249_ASFB7 Uncharacterized protein M1249L OS=African swine fever virus (strain Badajoz 1971 Vero-adapted) OX=10498 GN=Ba71V-060 PE=3 SV=1

MEEVITIAQIVHRGTDILSLNNEEIEALVDEIYSTLKGSNDIKNIRLIDFLFTLKDFVNH
VRAEQSKLPDLSMPMEAYIRQLLVDPDVVPIVSEKKKELRVRPSTRKEIFLINGTHLAVP
AEAPIEIYGLKLRLKSFSPQCFMRMAEIGSFSPETLGYVASGANLTNFIRVFMKCVDQET
WKKNGEGVVVTTKENIIQFTHQYIELYKFLRSGGHSWLINRLAEEMVHRKLDREDQGSHI
SNIVETEEIEPEENIKRVIFFLKELSTMYSVSPVFTSGYMPLLYDLYRAGYLEVLWNPVE
QKFLQHAEQREKEQMILQQVDMKLTEVITQARQYFKIMEEKIGRVQSDAIREILTMEGKV
DDPNSILQEVIKACGKQEAELITTEYLNIKKQWELQEKNACAHLKLVKQLRSGLQYAELL
KVLESIRVLYKEKNNTTNWNLCKACGFKLLCPHVDMLIQLQAAEASYDTMRTKLMKFSGI
NKEKENNQGLIYSYFCKICGEELAHFIQEDRTADVGVIGDLNSKLRIFIWQETMKACTFI
HFGKLVDVKQFANIAVNVCLPLVYSIENIKKEEDYDPLTQLYAVIYIYAYILNLIYSSQK
NKEFLTITIHGMKADSSLNAYVTFLLEKMMQQYSGIINQLSEITDQWIANNFREAFKKII
HQNGLQGLSVQDDTKVLLTEILLGPMYDYAATVARIDGSIPMHKPRTPKEAEYEFKTVIG
RTPAELLSQKEFYDKIYTSKYRPDFTQLARLNDIYFQEESLRVWWGGRDEEKTSTLIYLR
AYELFLKYLQNAPNFNSELAEFKTYENAYGEQKALLAQQGFYNIFDPNTGRADQRTRLFE
YKRLPISTLYDERGLPHKWTIYVYKAVDSSQKPAEIEVTRKDVIKKIDNHYALADLRCSV
CHVLQHEVGQLNIKKVQTALKASLEFNTFYAFYESRCPKGGLHDFQDKKCVKCGLFTYII
YDHLSQPELVHDYYNNYKDQYDKEKMSIRSIQIKKDMTTPSSETQPKPPQEPWTFDYGKI
IKTAKILDISPAVIEAIGAMEGRSYADIREGQGAPPPPTSMDDPRLMAVDSAVRIFLYNY
NCLRHVSTFNKPPMHVERLVKHLSYEEKEDLEKVLPNVVNEYHTTFKHLRVTDPASALLY
SIEFLCVSFLTLYEIKEPSWVVNIVREFALTELNTIIQSEKLLSKPGAFNFMIFGEDFVC
SGEDSSMDDISAYSSPGLFGEDIIDRLDDPFSIEDVDISLDVLDNLAPQ

>sp|Q89829|VF93L_ASFB7 Uncharacterized membrane protein KP93L/DP93R
OS=African swine fever virus (strain Badajoz 1971 Vero-adapted)
OX=10498 GN=BA71V-002 PE=3 SV=1
MFFFGIFRCNMDHWTTKRQVYIYLCFSLMTIALICYLIHICCHTKKNVVTNALPSNNMAL
IPYTPSNNMALIPYTPSNNTVPPPYTISGSCPQ

Appendix II: Perl script used in calculating posterior probabilities in CATH and SCOP

```
#!/usr/bin/perl
use Math::Trig;

############## set your paths for TMalign and stride ##########
$tmalign="/home/dickson/tasser/tmalign";
$stride="";
# examples #
$tmalign="/home/dickson/tasser/tmfold/TMalign";
$stride="/home/dickson/tasser/tmfold/stride";
# examples #


if(@ARGV<=1){usage();exit;}

for(my $i=0; $i<@ARGV; ++$i) {
   $ARGV[$i] eq $p1 && $i+1<@ARGV) { $pdbA =  $ARGV[++$i];}
   $ARGV[$i] eq $p2 && $i+1<@ARGV) { $pdbB =  $ARGV[++$i];}

}

sub usage {
   print STDERR "$0";
   print STDERR " -$p1 structure file1";
   print STDERR " -$p2 structure file2\n";
}

############## secodnary structure calculation ################
$pdbA_ss = &get_ss($pdbA);
$pdbB_ss = &get_ss($pdbB);

@tmpss=split(/\t+/,$pdbA_ss);
$pdb=$tmpss[0];
$pdb_nn_hash{$pdb}=$tmpss[1]; # number of residues in coil region at N terminal
$pdb_cc_hash{$pdb}=$tmpss[2]; # number of residues in coil region at C terminal

@tmpss=split(/\t+/,$pdbB_ss);
$pdb=$tmpss[0];
$pdb_nn_hash{$pdb}=$tmpss[1];
$pdb_cc_hash{$pdb}=$tmpss[2];

############## tmscore calculation by TMalign ################
$tmalign_rst = &get_alignment($pdbA,$pdbB);

@tmpts=split(/\t+/,$tmalign_rst);
$pdb1=$tmpts[0];
$pdb2=$tmpts[1];
```

```
$l1=$tmpts[3]; # length of protein1
$l2=$tmpts[4]; # length of protein2

# number of coil residues at N-,C-terminals
$chop1_n=$pdb_nn_hash{$pdb1};
$chop1_c=$pdb_cc_hash{$pdb1};
$chop2_n=$pdb_nn_hash{$pdb2};
$chop2_c=$pdb_cc_hash{$pdb2};

# number of unaligned residues at N-,C-terminals
$tms_N1=$tmpts[5];
$tms_N2=$tmpts[6];
$tms_C1=$tmpts[7];
$tms_C2=$tmpts[8];

# number of unaligned and coil residues at N-,C-terminals
$chop1n=($chop1_n<$tms_N1)?$chop1_n:$tms_N1;
$chop2n=($chop2_n<$tms_N2)?$chop2_n:$tms_N2;
$chop1c=($chop1_c<$tms_C1)?$chop1_c:$tms_C1;
$chop2c=($chop2_c<$tms_C2)?$chop2_c:$tms_C2;

# number of resiudes can be chopped
$chop1=($chop1n+$chop1c);
$chop2=($chop2n+$chop2c);

# new length used in normalizing TM-score
$l1_new=$l1-$chop1;
$l2_new=$l2-$chop2;
$short_new=($l1_new<$l2_new)?$l1_new:$l2_new;
$short_old=($l1<$l2)?$l1:$l2;

# new TM-score
$tms_old=$tmpts[2];
$tms_new=$tmpts[2]*$short_old/$short_new;
if($tms_new>=1){
   $tms_new=1;
}
$tms_new=sprintf ("%.4f",$tms_new);
$tms_old=sprintf ("%.4f",$tms_old);

################## statistical analysis of tmscore ###############
# set parameters of posterior probability and EVD
$a1_scop=-0.0071735;
$a2_scop=0.99803;
$x0_scop=0.57917;
$dx_scop=0.048934;

$a1_cath=-0.004103;
$a2_cath=0.99217;
$x0_cath=0.52695;
```

```
$dx_cath=0.04688;

$u=0.1512;
$sigma=0.0242;

# pvalue and posterior probability calculation
$pvalue=1-exp(0-exp(($u-$tms_new)/$sigma));
$prob_scop=$a2_scop+($a1_scop-$a2_scop)/(1+exp(($tms_new-$x0_scop)/$dx_scop));
$prob_cath=$a2_cath+($a1_cath-$a2_cath)/(1+exp(($tms_new-$x0_cath)/$dx_cath));
if($prob_scop<0){$prob_scop=0;}
if($prob_scop>1){$prob_scop=1;}
if($prob_cath<0){$prob_cath=0;}
if($prob_cath>1){$prob_cath=1;}

if($pvalue<0){$pvalue=0;}
if($pvalue>1){$pvalue=1;}
$prob_scop=sprintf ("%.15f",$prob_scop);
$prob_cath=sprintf ("%.15f",$prob_cath);
$pvalue=sprintf ("%.15f",$pvalue);

############################## output result ###################################
do {
print "Proteins in Comparison:\n";
print "Structure 1: $pdb1\n";
print "Structure 2: $pdb2\n\n";



print "Similarity Scores:\n";
print "TM-score after  Normalization:                    $tms_new\n";
print "TM-score before Normalization:                     $tms_old\n\n";

print "Statistical Scores:\n";
print "Probability of sharing same fold (SCOP):             $prob_scop\n";
print "Probability of sharing same topolgy (CATH):            $prob_cath\n";
print "P-value (Significance of the alignment):           $pvalue\n\n";
print "$alignment";
print "//\n";}
while
exit();

sub get_alignment{
   my ($pdbA, $pdbB)=@_;
   my $k=0;

   my @result=`$tmalign $pdbA $pdbB -b`;
   chomp(@result);
   $result[8]=~ m/Chain 1:[\d\D\w\W\s]{5,}Size=([0-9\s]{2,5})\s*/;
   my $length1=$1;
   $result[9]=~ m/Chain 2:[\d\D\w\W\s]{5,}Size=([0-9\s]{2,5})\s*/;
```

```
    my $length2=$1;

    $result[11]=~ m/Aligned length=([0-9\.\s]{2,5}),/;
    my $aligned=$1;
    $result[11]=~ m/Aligned length=[0-9\.\s]{2,5}[\d\D\w\W\s]{5,}TM-score=([0-9\.\s]{2,}),/;
    my $tmscore=$1;

    my $ch1=$result[20];
    my $dot=$result[21];
    my $ch2=$result[22];
    $ch1 =~ s/\s+//g;
    $ch2 =~ s/\s+//g;
    my @tmp1=split(//,$ch1);
    my @tmp2=split(//,$ch2);
    my @tmpD=split(//,$dot);
    my $N1=0;
    my $N2=0;
    my $C1=0;
    my $C2=0;
    for($k=0;$k<@tmp1;$k++){
            if($tmpD[$k]=~/\.|:/){
               last;
            }
            if($tmp1[$k]=~/[A-Z]/){
               $N1++;
            }
            if($tmp2[$k]=~/[A-Z]/){
               $N2++;
            }
    }

    for($k=@tmp1;$k>=0;$k--){
            if($tmpD[$k]=~/\.|:/){
               last;
            }
            if($tmp1[$k]=~/[A-Z]/){
               $C1++;
            }
            if($tmp2[$k]=~/[A-Z]/){
               $C2++;
            }
    }

undef $alignment;
    for($k=13;$k<@result;$k++){
            $alignment.= $result[$k]."\n";
    }
    return "$pdbA\t$pdbB\t$tmscore\t$length1\t$length2\t$N1\t$N2\t$C1\t$C2";
}
```

```
sub get_ss{
   my ($pdb) = @_;
   my @sec;
   my $k;
   my $q;
   my $j;
   my @lines=`$stride $pdb 2> /dev/null`;
   chomp(@lines);
   for($j=0;$j<@lines;$j++){
         my $line=$lines[$j];
         if($line=~/^ASG/){
            $k++;
            $sec_tmp=substr($line,24,1);
            $sec[$k]=1 if($sec_tmp eq "H");
            $sec[$k]=2 if($sec_tmp eq "E");
            $sec[$k]=3 if($sec_tmp ne "E" && $sec_tmp ne "H");
         }
   }
   my $chl=@sec;

   my $N1=0;
   my $C1=0;
   for($q=1;$q<$chl;$q++){
         if($sec[$q]!=3){
            last;
         }
         $N1++;
   }
   for($q=($chl-1);$q>=1;$q--){
         if($sec[$q]!=3){
            last;
         }
         $C1++;
   }
   my $tot=$N1+$C1;
   return "$pdb\t$N1\t$C1\t$tot";
}
```

Appendix III: Sequences used in phylogenetic classification

	Description	Accession	Organism
pB263R Isolate			
pB263R|Ba71v	Uncharacterized protein B263R Ba71v isolate	Q65175	*African swine fever virus*
NCLDVs			
Apm|TBP	TATA-box-binding protein-like	NC_014649.1	*Acanthamoeba polyphaga mimivirus*
Hav| Putative TBP	Putative TATA-box-binding protein	KX008963.1	*Heterosigmaakashiwo virus*
Mv|TBP	TATA-box-binding protein-like protein	AEX61565	*Megavirus courdo7*
YlpvA|Putative TBP	putative TATA box binding protein	YP_009174782	*Yellowstone lake phycodnavirus 1*
Pbcv |TFIID/TBP1	transcription factor TFIID (or TATA-binding protein TBP)	AGE48549	*Paramecium bursaria Chlorella virus AN69C*
Pbcv |TFIID/TBP2	transcription factor TFIID (or TATA-binding protein TBP)	AGE48777	*Paramecium bursaria Chlorella virus AP110A*
Atcv |TFIID/TBP	Transcription factor TFIID (Or TATA-binding protein TBP)	AGE56691	*Acanthocystisturfacea Chlorella virus*
Archaea, Eukaryotes TBP			
Sc|1ytb_A|TBP	TATA-binding protein	AEP68421	*Saccharomyces cerevisiae*
Sc|4b0a_A|TIF TFIID1	Yeast TATA-binding protein-yeast TAF1	4B0A_A	*Saccharomyces cerevisiae*
Sc|1rm1_A|TFIIA TBP	TATA-binding protein	NP_011075.1	*Saccharomyces cerevisiae*
Sc|TBP	TATA-binding protein	NP_011075.0	*Saccharomyces cerevisiae S288c*
At|TBP2	TATA binding protein 2	NP_175948	*Arabidopsis thaliana*
me| cd04516|TBP	TATA binding protein	OAQ30812	*Mortierellaelongata AG-77*
At|TBP1	TATA binding protein1	NP_187953	*Arabidopsis thaliana*
At|TBP CTD	TBP Recognizing The Minor Groove Of A Tata Element.	1VTL_E	*Arabidopsis thaliana*
Ec|TBP	Mot1 N-Terminal Domain In Complex With TBP	3OC3_C	*Encephalitozoon cuniculi*
In|TLF-Iso X1	TATA-box-binding protein-like isoform X1	XP_019186975	*Ipomoea nil*
Ce| TBP-1	TATA-box-binding protein	NP_498635	*Caenorhabditis elegans*
Mmus|TRF3	TATA box-binding protein-like protein 2 isoform 1	NP_951014	*Mus musculus*

Dm|TBP	TATA binding protein	NP_523805	*Drosophila melanogaster*
Dm|TRF	TBP-related factor	NP_476939	*Drosophila melanogaster*
Hs|TBP2	TATA box-binding protein	2119243A	*Homo sapiens*
Hs|TL	TATA box-binding protein-like protein 1	NP_004856	*Homo sapiens*
Hs| 5iyb P|TBP	TATA-box-binding protein isoform 2	NP_001165556	*Homo sapiens*
Hs|TBP	TATA-box-binding protein isoform 1	NP_003185	*Homo sapiens*
mv|Cd00652|TBP TLF	TATA-box-binding protein	KFH70047	*Mortierellaverticillata NRRL 6337*
Ce|TLF-1	TBP-Like Factor	NP_492356	*Caenorhabditis elegans*
ag|Cd04517 TLF TBPL	TATA box-binding protein-like protein 1	XP_018577452	*Anoplophoraglabripennis*
Hs|TLP2	TATA box-binding protein-like protein 2	NP_950248	*Homo sapiens*
Dm|TRF-IsoJ	TATA box binding protein-related factor 2, isoform J	NP_511084	*Drosophila melanogaster*
Dm|TRF-IsoH	TATA box binding protein-related factor 2, isoform H	NP_996377	*Drosophila melanogaster*
Sa|1mp9A|TBP TF	TATA box-binding protein	Q9UWN7	*Sulfolobusacidocaldarius*
Sa|TBP CTD	TATA-box-binding protein	WP_011278173	*Sulfolobusacidocaldarius*
SsP2|TFIID	TATA-box-binding protein	AAK41225	*Sulfolobussolfataricus P2*
mk|TF	transcription factor	WP_011019207	*Methanopyruskandleri AV19*
sp|TBP	TATA-binding protein (TBP)	NP_594566	*Schizosaccharomyces pombe*
Ap|TFIID TBP	transcription factor IID TBP	NP_148219	*Aeropyrumpernix K1*
Tp|TBP	TATA box-binding protein	WP_011751968	*Thermofilum pendens*
Ckc|TFP	Archaeal TATA-box-binding family protein	WP_012310109	*CandidatusKorarchaeum cryptofilum*
Nm|TFP	TATA-box binding family protein	ABX13415	*Nitrosopumilusmaritimus SCM1*
Pa|TF	transcription factor TATA-boxbinding protein	WP_011008464	*Pyrobaculumaerophilum*
pw|1ais A|TBP	Chain A, Tata-Binding Protein TRANSCRIPTION FACTOR	1AIS_A	*Pyrococcuswoesei*
pw|d1aisa2|TBP CTD	TATA-box-binding protein	WP_014734019	*Pyrococcuswoesei*
ph|TFP	TATA box binding protein (TBP) Transcription Factor	WP_010885095	*Pyrococcushorikoshii*
pw|d1aisa1| TBP	Chain A, Tata-Binding Protein	1D3U_A	*Pyrococcuswoesei*

Appendix IV: Untruncated DNA parameters used in simulations

Base pairing								
[Pair ID	pairnum	pair	chain1	resnum1	base1	base2	resnum2	chain2
1	G-C	A	2	G	C	19	B	3
2	T-A	A	3	T	A	18	B	2
3	A-T	A	4	A	T	17	B	2
4	T-A	A	5	T	A	16	B	2
5	A-T	A	6	A	T	15	B	2
6	T-A	A	7	T	A	14	B	2
7	A-T	A	8	A	T	13	B	2
8	C-G	A	9	C	G	12	B	3

Local base-pair parameters							
[Pair ID	pair	shear	stretch	stagger	buckle	propeller	opening
1	G-C	-0.16	0	-0.14	-19.92	-3.34	1.47
2	T-A	-0.25	-0.05	-0.3	-4	-5.14	2.73
3	A-T	0.06	-0.18	0.08	-4.3	-6.94	-1.42
4	T-A	0.03	-0.04	-0.1	8.08	-4.38	0.88
5	A-T	-0.08	-0.22	0.13	-6.16	-13.52	1.17
6	T-A	-0.05	-0.07	0.1	-0.99	-15.36	0.25
7	A-T	-0.02	-0.16	-0.01	0.43	-1.76	4.23
8	C-G	0.16	-0.15	-0.25	15.15	4.46	-0.77

Simple base-pair parameters									
Step ID	step	base1	base2	shift	slide	rise	tilt	roll	twist
1	C/G	A:1:C	A:2:G	10.02	1.74	11.64	-94.16	-80.28	103.35
2	G/T	A:2:G	A:3:T	-0.7	-0.95	2.88	11.44	1.17	33.1
3	T/A	A:3:T	A:4:A	-1.02	0.2	3.64	-3.91	1.41	35.29
4	A/T	A:4:A	A:5:T	0.59	-0.04	2.9	9.85	1.55	27.89
5	T/A	A:5:T	A:6:A	0.32	1.39	3.66	-4.95	-10.8	51.95
6	A/T	A:6:A	A:7:T	0.23	-0.46	3.18	5.12	-1.84	26.68
7	T/A	A:7:T	A:8:A	1.07	1.17	3.23	8.8	7.1	45.51
8	A/C	A:8:A	A:9:C	0.37	0.16	2.87	9.78	11.07	20.85
9	C/G	A:9:C	A:10:G	17.67	2.11	-7.36	132.25	-2.41	140.92
10	G/C	A:10:G	B:11:C	-5.19	15.4	-26.74	98.42	41.82	-43.67
11	C/G	B:11:C	B:12:G	6.46	2.73	13.17	-77.74	-69.85	88.31
12	G/T	B:12:G	B:13:T	-0.29	-0.24	3.1	4.26	1.86	26.53
13	T/A	B:13:T	B:14:A	-0.45	1.48	3.39	-1.04	-5.11	42.13
14	A/T	B:14:A	B:15:T	0.61	-0.4	3.18	6.72	1.63	27.41
15	T/A	B:15:T	B:16:A	0.41	1.4	3.49	-0.27	-3.43	50.85
16	A/T	B:16:A	B:17:T	-0.36	-0.28	3.11	4.82	-1.83	25.76
17	T/A	B:17:T	B:18:A	1.73	0.59	3.09	7.5	5.69	38.64
18	A/C	B:18:A	B:19:C	0.7	-0.23	3.11	6.39	9.61	30.83
19	C/G	B:19:C	B:20:G	7.77	6.27	4.02	11.79	58.07	116.79

Main chain and chi torsion angles										
Serial	Chain	Residue ID	Base	alpha	beta	gamma	delta	epsilon	zeta	chi
1	A	1	DC	---	-174.9	48	144.5	175	-111	-106.1
2	A	2	DG	76.9	134.3	31.5	102.2	-158.5	-73.3	-170.7
3	A	3	DT	-70	-163.2	55.7	144.7	-159.1	-123.7	-113.4
4	A	4	DA	-69.6	-173.2	43.2	141.5	-175.2	-94.8	-101.3
5	A	5	DT	-62.3	172.2	47.7	138.9	-99.6	168.1	-98.4
6	A	6	DA	-75.5	146.3	49.6	137.9	-164.5	-92.9	-113
7	A	7	DT	-54.1	165.2	39.5	104.9	-177.1	-94.2	-120.9
8	A	8	DA	-60.5	-171.4	48.6	142.2	-164.8	-88	-91.6
9	A	9	DC	-70.2	171.4	35.7	94.3	-152.3	82.2	-104.9
10	A	10	DG	77	-164.7	49.4	115.4	---	---	-122.2
11	B	11	DC	---	-174.8	48	144.6	175.1	-111	-106.1
12	B	12	DG	84.7	157.2	40.8	144.3	-172.3	-96	-105.5
13	B	13	DT	-70.3	178.3	42	132.5	-116.5	-175.7	-99.3
14	B	14	DA	-65.3	150.7	41.2	133.9	174.4	-70.7	-104.1
15	B	15	DT	-81.3	-176.9	55.6	136.6	-111.8	-176.9	-105.2
16	B	16	DA	-49.3	152.2	26.1	137.7	-177.7	-92.5	-101.2
17	B	17	DT	-60.8	170.9	37.8	108.9	-173	-92.6	-112.5
18	B	18	DA	-61.5	162.8	55.7	112.3	-167.7	-96	-117.6
19	B	19	DC	-56.7	167.6	43.3	133.6	-136.5	-70.1	-107.9
20	B	20	DG	-65.2	177.1	54.7	95.5	---	---	-69.2

Pseudo eta(η)/theta(θ) torsion angles									
Serial	Chain	Residue ID	Base	eta	theta	eta'	theta'	eta''	theta''
1	A	1	DC	---	-102.7	---	-130.4	---	-132.4
2	A	2	DG	-103.7	-128.8	-76.3	-128.9	-50.6	-116.5
3	A	3	DT	173.2	-143.8	-148.3	-156.4	-126.2	-121.5
4	A	4	DA	165.3	-126.5	-171.9	-144.8	-133.3	-108.5
5	A	5	DT	-170.7	167.6	-147.1	167.5	-112.3	-121.1
6	A	6	DA	171.5	-113	-162.3	-144.6	-97.1	-113.4
7	A	7	DT	166.7	-169.7	-169.8	-166.3	-131.8	-109
8	A	8	DA	178.4	-116.3	-155	-141.4	-101.1	-114.1
9	A	9	DC	154.7	37.6	170.5	-6.2	-156.5	-42.1
10	A	10	DG	---	---	---	---	---	---
11	B	11	DC	---	-117.7	---	-160.9	---	-139.6
12	B	12	DG	-84.4	-114.5	-42.6	-144.7	-48.6	-114.6
13	B	13	DT	177.3	174.3	-161	175	-130.7	-129.2
14	B	14	DA	163.8	-116.6	-168.9	-146.3	-116.9	-113.6
15	B	15	DT	-172.5	171.7	-150.9	176.5	-118.1	-117.4
16	B	16	DA	174.4	-127.9	-160.2	-156.5	-100.3	-126.9
17	B	17	DT	170.9	-152.1	-171.4	-156.2	-138.9	-100.7
18	B	18	DA	171.9	-150.3	-161.5	-158.4	-104.2	-120.6
19	B	19	DC	176.9	-90.4	-159.7	-92.5	-120.7	-25.9
20	B	20	DG	---	---	---	---	---	---

Appendix V: Truncated DNA parameters used in simulations

[Pair ID	pair number	pair	chain1	resnum1	base1	base2	resnum2	chain2
1	G-C	A	2	G	C	19	B	3
2	T-A	A	3	T	A	18	B	2
3	A-T	A	4	A	T	17	B	2
4	T-A	A	5	T	A	16	B	2
5	A-T	A	6	A	T	15	B	2
6	T-A	A	7	T	A	14	B	2
7	A-T	A	8	A	T	13	B	2
8	C-G	A	9	C	G	12	B	3

long basepair parameters							
[Pair ID	pair	shear	stretch	stagger	buckle	propeller	opening
1	G-C	-0.16	0	-0.14	-19.92	-3.34	1.47
2	T-A	-0.25	-0.05	-0.3	-4	-5.14	2.73
3	A-T	0.06	-0.18	0.08	-4.3	-6.94	-1.42
4	T-A	0.03	-0.04	-0.1	8.08	-4.38	0.88
5	A-T	-0.08	-0.22	0.13	-6.16	-13.52	1.17
6	T-A	-0.05	-0.07	0.1	-0.99	-15.36	0.25
7	A-T	-0.02	-0.16	-0.01	0.43	-1.76	4.23
8	C-G	0.16	-0.15	-0.25	15.15	4.46	-0.77

Simple Base-pair Parameters									
Step ID	step	pairnum1	pairnum2	shift	slide	rise	tilt	roll	twist
1	GT/AC	1	2	-0.7	-0.58	3.02	2.71	5.51	32.01
2	TA/TA	2	3	-1.38	0.41	3.38	-5.68	3.6	35.61
3	AT/AT	3	4	0.48	-0.15	3.01	2.52	-0.14	27.04
4	TA/TA	4	5	-0.05	1.39	3.59	-2.35	-7.14	52.02
5	AT/AT	5	6	-0.19	-0.42	3.21	-0.79	-0.06	26.95
6	TA/TA	6	7	0.77	1.33	3.32	4.81	1.05	44.1
7	AC/GT	7	8	0.32	-0.03	2.99	2.65	6.6	22.42

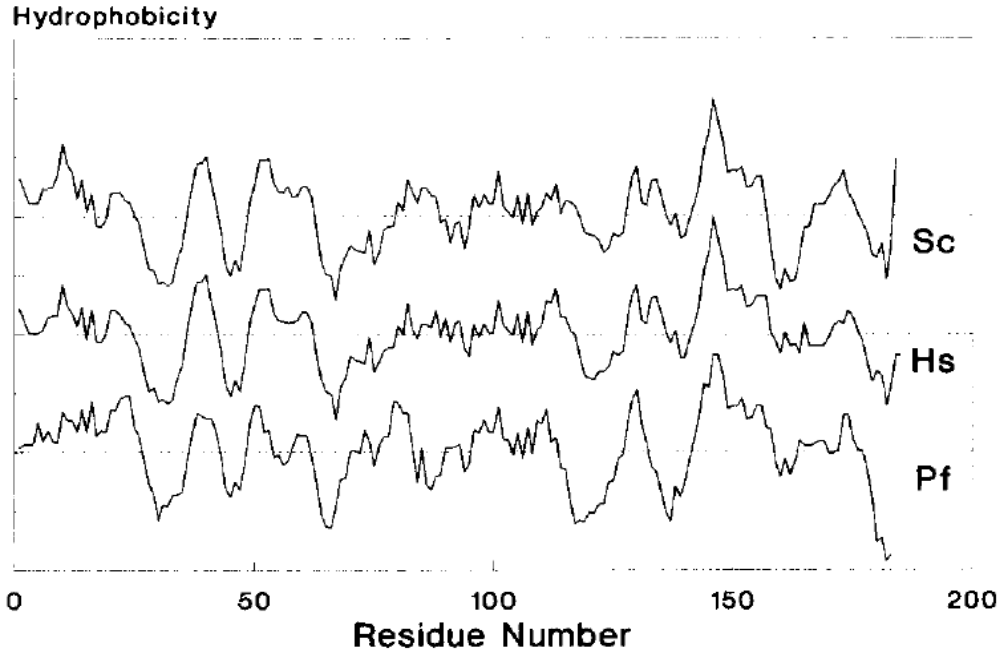
simple base-pair parameters									
Step ID	step	base1	base2	shift	slide	rise	tilt	roll	twist
1	G/T	A:2:G	A:3:T	-0.7	-0.95	2.88	11.44	1.17	33.1
2	T/A	A:3:T	A:4:A	-1.02	0.2	3.64	-3.91	1.41	35.29
3	A/T	A:4:A	A:5:T	0.59	-0.04	2.9	9.85	1.55	27.89
4	T/A	A:5:T	A:6:A	0.32	1.39	3.66	-4.95	-10.8	51.95
5	A/T	A:6:A	A:7:T	0.23	-0.46	3.18	5.12	-1.84	26.68
6	T/A	A:7:T	A:8:A	1.07	1.17	3.23	8.8	7.1	45.51
7	A/C	A:8:A	A:9:C	0.37	0.16	2.87	9.78	11.07	20.85
8	C/G	A:9:C	B:12:G	-0.11	0.25	-0.18	164.21	1.11	32.79
9	G/T	B:12:G	B:13:T	-0.29	-0.24	3.1	4.26	1.86	26.53
10	T/A	B:13:T	B:14:A	-0.45	1.48	3.39	-1.04	-5.11	42.13
11	A/T	B:14:A	B:15:T	0.61	-0.4	3.18	6.72	1.63	27.41
12	T/A	B:15:T	B:16:A	0.41	1.4	3.49	-0.27	-3.43	50.85
13	A/T	B:16:A	B:17:T	-0.36	-0.28	3.11	4.82	-1.83	25.76
14	T/A	B:17:T	B:18:A	1.73	0.59	3.09	7.5	5.69	38.64
15	A/C	B:18:A	B:19:C	0.7	-0.23	3.11	6.39	9.61	30.83

Main chain and chi torsion angles										
Serial	Chain	Residue ID	Base	alpha	beta	gamma	delta	epsilon	zeta	chi
1	A	2	DG	---	134.3	31.5	102.2	-158.5	-73.3	-170.7
2	A	3	DT	-70	-163.2	55.7	144.7	-159.1	-123.7	-113.4
3	A	4	DA	-69.6	-173.2	43.2	141.5	-175.2	-94.8	-101.3
4	A	5	DT	-62.3	172.2	47.7	138.9	-99.6	168.1	-98.4
5	A	6	DA	-75.5	146.3	49.6	137.9	-164.5	-92.9	-113
6	A	7	DT	-54.1	165.2	39.5	104.9	-177.1	-94.2	-120.9
7	A	8	DA	-60.5	-171.4	48.6	142.2	-164.8	-88	-91.6
8	A	9	DC	-70.2	171.4	35.7	94.3	---	---	-104.9
9	B	12	DG	---	157.2	40.8	144.3	-172.3	-96	-105.5
10	B	13	DT	-70.3	178.3	42	132.5	-116.5	-175.7	-99.3
11	B	14	DA	-65.3	150.7	41.2	133.9	174.4	-70.7	-104.1
12	B	15	DT	-81.3	-176.9	55.6	136.6	-111.8	-176.9	-105.2
13	B	16	DA	-49.3	152.2	26.1	137.7	-177.7	-92.5	-101.2
14	B	17	DT	-60.8	170.9	37.8	108.9	-173	-92.6	-112.5
15	B	18	DA	-61.5	162.8	55.7	112.3	-167.7	-96	-117.6
16	B	19	DC	-56.7	167.6	43.3	133.6	---	---	-107.9

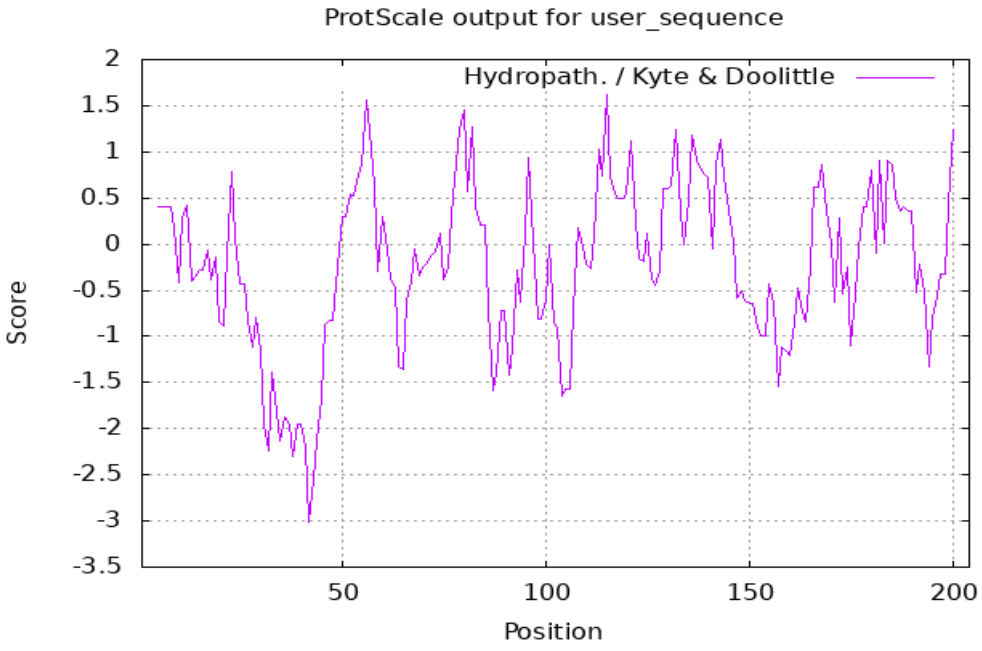
Pseudo eta(η)/theta(θ) torsion angles									
Serial	Chain	Residue ID	Base	eta	theta	eta'	theta'	eta"	theta"
1	A	2	DG	---	-128.8	---	-128.9	---	-116.5
2	A	3	DT	173.2	-143.8	-148.3	-156.4	-126.2	-121.5
3	A	4	DA	165.3	-126.5	-171.9	-144.8	-133.3	-108.5
4	A	5	DT	-170.7	167.6	-147.1	167.5	-112.3	-121.1
5	A	6	DA	171.5	-113	-162.3	-144.6	-97.1	-113.4
6	A	7	DT	166.7	-169.7	-169.8	-166.3	-131.8	-109
7	A	8	DA	178.4	-116.3	-155	-141.4	-101.1	-114.1
8	A	9	DC	---	---	---	---	---	---
9	B	12	DG	---	-114.5	---	-144.7	---	-114.6
10	B	13	DT	177.3	174.3	-161	175	-130.7	-129.2
11	B	14	DA	163.8	-116.6	-168.9	-146.3	-116.9	-113.6
12	B	15	DT	-172.5	171.7	-150.9	176.5	-118.1	-117.4
13	B	16	DA	174.4	-127.9	-160.2	-156.5	-100.3	-126.9
14	B	17	DT	170.9	-152.1	-171.4	-156.2	-138.9	-100.7
15	B	18	DA	171.9	-150.3	-161.5	-158.4	-104.2	-120.6
16	B	19	DC	---	---	---	---	---	---

Appendix VI: Alignment of 5'flanking sequences of late ASFV genes

Appendix VII: Hydropathy plot profiles



***Saccharomyces cerevisiae* (Sc), *Homo sapiens* (Hs) and *Plasmodium falciparum* (Hs) hydropathy profiles**



ASFV pB263R hydropathy profile

Appendix VIII: Publication 1

PLOS ONE

RESEARCH ARTICLE

Comparative *in silico* study of congocidine congeners as potential inhibitors of African swine fever virus

Dickson Kinyanyi[1]*, Peris Amwayi[1], Mark Wamalwa[2], George Obiero[3]

1 Department of Biochemistry and Biotechnology, Technical University of Kenya, Nairobi, Kenya, **2** Department of Biochemistry and Biotechnology, Kenyatta University, Nairobi, Kenya, **3** Center for Biotechnology and Bioinformatics, University of Nairobi, Nairobi, Kenya

* dkinyanyi@gmail.com

OPEN ACCESS

Citation: Kinyanyi D, Amwayi P, Wamalwa M, Obiero G (2019) Comparative *in silico* study of congocidine congeners as potential inhibitors of African swine fever virus. PLoS ONE 14(8): e0221175. https://doi.org/10.1371/journal.pone.0221175

Editor: Dennis Salahub, University of Calgary, CANADA

Received: March 7, 2019

Accepted: July 31, 2019

Published: August 28, 2019

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

African swine fever virus (ASFV) infection is fatal in domesticated pigs, with a mortality rate approaching 100%. This may result in economic losses and threats to food security. Currently, there are no approved vaccines or antiviral therapies for ASFV. Therefore, in this study, we evaluated congocidine congeners and a tris-benzimidazole as potential inhibitors of ASFV transcription using an *in silico* approach. We applied redocking of congocidine and docking of its congeners and a tris-benzimidazole to a receptor containing B-DNA with AT-motifs as a target to mimic conserved ASFV late gene promoters. Subsequently, the binding scores of DNA-ligand docked complexes were evaluated and their binding affinity was estimated. Molecular dynamics (MD) simulation was then used to assess ligand behavior within the minor groove. From our results, it is evident the less toxic congocidine congeners and tris-benzimidazole could dock to AT-rich regions significantly. Additionally, the predicted binding affinities had suitable values comparable to other experimentally determined minor groove binders, MD simulation of the docked DNA-ligand complexes and subsequent molecular trajectory visualization further showed that the ligands remained embedded in the minor groove during the time course of simulation, indicating that these ligands may have potential applications in abrogating ASFV transcription.

Introduction

DNA is a major target for various types of drugs [1]. Results from the analysis of several high-resolution structures suggest that the minor groove of DNA may function as a receptor for proteins and small molecules [2]. Moreover, drugs that bind to the minor groove may be exploited when pursuing a subset of viruses that replicate in the cytoplasm, such as the African swine fever (ASF) virus (ASFV). ASFV causes ASF, a fatal disease that affects domestic pigs. ASFV infection can affect the food supply, as pork is one of the most commonly consumed kinds of meat worldwide [3].

Currently, there are no vaccines or antiviral drugs approved for use against ASFV [4], and to date, reversible minor groove binders have not been applied or studied in mitigating ASFV

Appendix IX: Publication 2

PeerJ

In silico structural and functional prediction of African swine fever virus protein-B263R reveals features of a TATA-binding protein

Dickson Kinyanyi[1,*], George Obiero[2,*], George F.O. Obiero[1,*], Peris Amwayi[1,*], Stephen Mwaniki[1,*] and Mark Wamalwa[3,*]

[1] Department of Biochemistry and Biotechnology, Technical University of Kenya, Nairobi, Kenya
[2] Center for Biotechnology and Bioinformatics, University Of Nairobi, Nairobi, Kenya
[3] Department of Biochemistry and Biotechnology, Kenyatta University, Ruiru, Kenya
[*] These authors contributed equally to this work.

ABSTRACT

African swine fever virus (ASFV) is the etiological agent of ASF, a fatal hemorrhagic fever that affects domestic pigs. There is currently no vaccine against ASFV, making it a significant threat to the pork industry. The ASFV genome sequence has been published; however, about half of ASFV open reading frames have not been characterized in terms of their structure and function despite being essential for our understanding of ASFV pathogenicity. The present study reports the three-dimensional structure and function of uncharacterized protein, pB263R (NP_042780.1), an open reading frame found in all ASFV strains. Sequence-based profiling and hidden Markov model search methods were used to identify remote pB263R homologs. Iterative Threading ASSEmbly Refinement (I-TASSER) was used to model the three-dimensional structure of pB263R. The posterior probability of fold family assignment was calculated using TM-fold, and biological function was assigned using TM-site, RaptorXBinding, Gene Ontology, and TM-align. Our results suggests that pB263R has the features of a TATA-binding protein and is thus likely to be involved in viral gene transcription.

Submitted 4 October 2017
Accepted 30 January 2018
Published 22 February 2018

Corresponding author
Dickson Kinyanyi,
prodicky7777@yahoo.com,
dkinyanyi@gmail.com

Academic editor
Rob Russell

Additional Information and Declarations can be found on page 14

DOI 10.7717/peerj.4396

Subjects Biochemistry, Bioinformatics, Biotechnology, Computational Biology
Keywords African swine fever virus, TATA-binding protein, BA71V pB263R, *In silico* characterization, I-TASSER, Sequence homology search, African swine fever transcription factor

INTRODUCTION

African swine fever virus (ASFV) is a large enveloped double-stranded DNA virus and the sole member of the family Asfarviridae (*Dixon et al., 2004*), which belongs to the nucleo-cytoplasmic large DNA virus (NCLDV) superfamily, an apparently monophyletic class of viruses with eukaryotic hosts. NCLDVs partially replicate in the cytoplasm, with varying degrees of dependence on the host nucleus (*Dixon et al., 2013*); members of the superfamily have been classified to belong to the proposed new order named Megavirales, some include, Poxviridae, Iridoviridae, Mimiviridae, Phycodnaviridae, Marseilleviridae, and Ascoviridae (*Colson et al., 2012*; *Colson et al., 2013*). The natural hosts of ASFVs include bush pigs, warthogs, and argasid ticks of the genus *Ornithodoros* (*Carrillo et al., 1994*; *Rodríguez & Salas, 2013*) in these hosts, infection by the virus is not fatal. ASF in

How to cite this article Kinyanyi et al. (2018), *In silico* structural and functional prediction of African swine fever virus protein-B263R reveals features of a TATA-binding protein. PeerJ 6:e4396; DOI 10.7717/peerj.4396