

# Identification and validation of putative therapeutic and diagnostic antimicrobial peptides against HIV: An *in silico* approach



UNIVERSITY *of the*  
WESTERN CAPE

**Marius Belmondo Tincho**

A thesis submitted in partial fulfillment of the requirements for the degree of

**Magister Scientiae**

In the Department of Biotechnology

University of the Western Cape

Supervisor: Dr A. Pretorius

Co-Supervisors: Dr M. N. Gabere

Dr M. Meyer

November 2013



UNIVERSITY *of the*  
WESTERN CAPE

UNIVERSITEIT VAN WES-KAAPLAND  
BIBLIOTEK

LIBRARY  
UNIVERSITY OF THE WESTERN CAPE

# Declaration

I declare that “**Identification and validation of putative therapeutic and diagnostic antimicrobial peptides against HIV: An *in silico* approach**” is my own work, that it has not been submitted for degree or examination at any other university, and that all the resources I have used or quoted, and all work which was the result of joint effort, have been indicated and acknowledged by complete references.

-----  
Marius Belmondo Tincho

Signed November, 2013



# Abstract

**Background:** Despite the effort of scientific research on HIV therapies and to reduce the rate of HIV infection, AIDS still remains one of the major causes of death in the world and mostly in Sub-Saharan Africa. To date, neither a cure, nor an HIV vaccine had been found and the disease can only be managed by using High Active Antiretroviral Therapy (HAART) if detected early. The need for an effective early diagnostic and non-toxic therapeutic treatment has brought about the necessity for the discovery of additional HIV diagnostic methods and treatment regimens to lower mortality rates. Antimicrobial Peptides (AMPs) are components of the first line of defence of prokaryotes and eukaryotes and have been proven to be promising therapeutic agents against HIV.

**Methods:** With the utility of computational biology, this work proposes the use of profile search methods combined with structural modelling to identify putative AMPs with diagnostic and anti-HIV activity. Firstly, experimentally validated anti-HIV AMPs were retrieved from various publicly available AMP databases, APD, CAMP, Bactibase and UniprotKB and classified according to super-families. Hidden Markov Model (HMMER) and Gap Local Alignment of Motifs (GLAM2) profiles were built for each super-family of anti-HIV AMPs. Putative anti-HIV AMPs were identified after scanning genome sequence databases using the trained models, retrieved AMPs and ranked based on their E-values. The 3-D structures of the 10 peptides that were ranked highest were predicted using I-TASSER. These peptides were docked against various HIV proteins using PatchDock and putative AMPs showing highest affinity and having the correct orientation to the HIV-1 proteins

gp120 and p24 were selected for future work so as to establish their function in HIV therapy and diagnosis.

**Results:** The results of the *in silico* analysis showed that the constructed models using the HMMER algorithm had better performances compare to that of the models built by the GLAM2 algorithm. Furthermore, the former tool has better statistical and probability explanation compared to the latter tool. Thus only the HMMER scanning results were considered for further study. Out of 1059 species scanned by the HMMER models, 30 putative anti-HIV AMPs were identified from genome scans with the family specific profile models after elimination of duplicate peptides. Docking analysis of putative AMPs against HIV proteins showed that from the 10 best performing anti-HIV AMPs with the highest E-scores, molecules 1, 3, 8 and 10 firmly binds the gp120 binding pocket at the V1/V2 domain and at the point of interaction between gp120 and T cells, with the 1<sup>st</sup> and 3<sup>rd</sup> highest scoring anti-HIV AMPs having the highest binding affinities. However, all 10 putative anti-HIV AMPs bind to the N-terminal domain of p24 with large surface interaction, rather than the C-terminal.

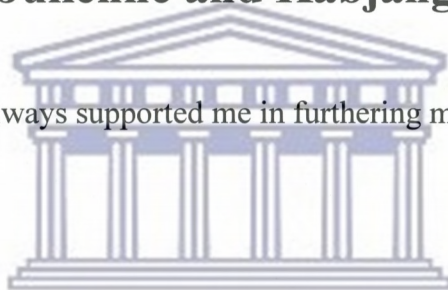
**Conclusion:** The *in silico* approach has made it possible to construct computational models having high performances, and which enabled the identification of putative anti-HIV peptides from genome sequence scans. The *in silico* validation of these putative peptides through docking studies has shown that some of these AMPs may be involved in HIV/AIDS therapeutics and diagnostics. The molecular validation of these findings will be the way forward for the development of an early diagnostic tool and as a consequence initiate early treatment. This will prevent the invasion of the immune system by blocking the V1/V2 domain and thus designing of a successful vaccine with broad neutralizing activity against this domain.

# Dedication

I dedicate this work to my mother and my father

**Tangué Julienne and Kabjang Joseph,**

who have always supported me in furthering my education



UNIVERSITY *of the*  
WESTERN CAPE

# Acknowledgements

First and foremost, I give appreciation to the **LORD JESUS CHRIST**, for His Infinite Blessings and Guidance in my life.

I would like to thank the Nanotechnology Innovation Centre (NIC) and the Medical Research Council (MRC) for funding during this project. I would like to express my sincere gratitude to my advisors Dr Ashley Pretorius, Dr Musa N. Gabere and Dr Mervin Meyer for their guidance, intellectual input, their enormous effort and time in the supervision of this thesis.

Special thanks go to Dr Mushal Ali and Dr Ruben Cleote for their bioinformatics support.

I cannot finish without mentioning the love, support and devotion of my family. They have always given me the strength, motivation and encouragement that made me complete my degree. Many thanks go to my mother, Julienne, my father, Joseph, all my brothers, Cyrille, Martial, Mark and Mike, my aunt Nyawa and, Mr and Mrs Saibu for their spiritual and academic support. A special thank goes to the lovely woman in my life, Hongue Jifedie Prudence for her long-time support and encouragement.

Lastly, I would like to extend my gratitude to the Bioinformatics Research Group (BRG) members especially to Habeeb, Namhla and Monray for their support.

## List of Conference Abstracts

Tincho, M. B., Gabere, M. N., and Pretorius, A. (2012) Identification and validation of novel therapeutic antimicrobial peptides against Human Immunodeficiency Virus: An *in silico* and molecular approach. 5<sup>th</sup> NIC-DST/Mintek annual conference at the MRC conference, Cape Town (Poster presentation)

Tincho, M. B., Gabere, M. N., Meyer, M. and Pretorius, A. (2013) Discovering of putative Antimicrobial peptides for early diagnostic and treatment of HIV-1 infection: a combination of computational and molecular biology approaches. 2<sup>nd</sup> Global Infectious Disease Research: a multidisciplinary approach, International Centre for Genetic Engineering and Biotechnology (ICGEB), Cape Town, South Africa (Poster presentation)

Tincho, M. B., Gabere, M. N., Meyer, M. and Pretorius, A. (2013) Identification and validation of putative therapeutic and diagnostic antimicrobial peptides against HIV: An *in silico* and molecular approach. 17<sup>th</sup> International Conference on AIDS and STIs in Africa (17<sup>th</sup> ICASA), Cape Town, South Africa (Oral presentation)



# List of Abbreviations

ACE	Atom Contact Energy
AIDS	Acquired Immunodeficiency Syndrome
AMP	Antimicrobial Peptide
ANN	Artificial Neural Networks
APD	Antimicrobial Peptides Database
BLAST	Basic Local Alignment Search Tool
CAMP	Collection of Anti-Microbial Peptides
CDC	Centers for Disease Control
CRFs	Circulating Recombinant Forms
DC	Dendritic Cells
DNA	Deoxyribonucleic Acid
EIA	Enzyme Immunoassay
FDA	Food and Drug Administration
FFT	Cartesian grid-based First Fourier Transformation
FN	False Negative



FP	False Positive
FRET	Fluorescence Resonance Energy Transfer
GLAM2	Gap Local Alignment of Motifs
GRID	Gay-Related Immune Deficiency
HAART	High Active Antiretroviral Therapy
HIV	Human Immunodeficiency Virus
HMMER	Hidden Markov Models
II	Integrase Inhibitors
I-TASSER	Iterative Threading ASSEMBLY Refinement
ITC	Isothermal Titration Calorimetry
KS	Kaposi's Sarcoma
LD	Linear Discriminant Analysis
NNRTIs	Non-Nucleoside Reverse Transcriptase Inhibitors
NRTIs	Nucleoside Analogue Reverse Transcriptase Inhibitors
PDB	Protein Data Bank
PI	Protease Inhibitors
QSAR	Quantitative Structure Activity Relationship
RF	Random Forest



RMSD	Root Mean Square Deviation
RNA	Ribonucleic Acid
RT	Reverse Transcriptase
RT-qPCR	Real Time Reverse Transcription quantitative Polymerase Chain Reaction
SDS	Sodium Dodecyl Sulfate
SIV	Simian Immunodeficiency Virus
SNPs	Single Nucleotide Polymorphisms
SPF	Spherical Polar Fourier
SPR	Surface Plasmon Resonance
SVM	Support Vector Machine
SW	Sliding Window
TN	True Negative
TP	True Positive
WHO	World Health Organization

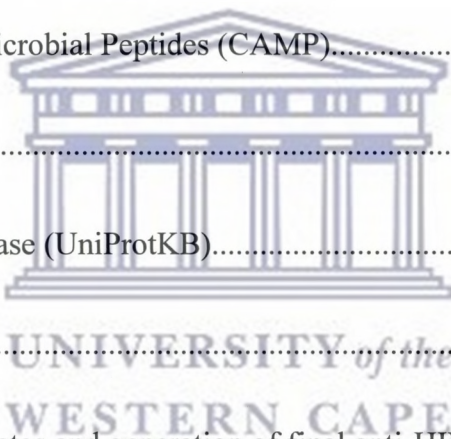


# Table of contents

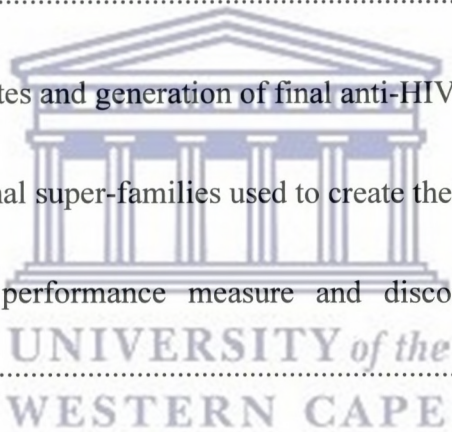
<b>1 Literature review</b> .....	1
1.1 Introduction.....	1
1.2 HIV/AIDS.....	5
1.2.1 History.....	5
1.2.1.1 Discovery.....	5
1.2.1.2 Origins.....	6
1.2.1.3 Classification of HIV.....	7
1.2.2 Transmission and life cycle of HIV.....	8
1.2.3 HIV infection .....	10
1.2.4 Immune invasion of HIV.....	11
1.2.5 Diagnosis of HIV/AIDS .....	12
1.2.6 Prevention of HIV/AIDS, management and treatment .....	14
1.2.7 Estimation of HIV statistics.....	16
1.3 Antimicrobial Peptides: the defence system of all living organisms.....	18
1.3.1 The adaptive immune system or acquired immunity.....	19
1.3.2 The innate immune system or natural immunity.....	19
1.3.3 Antimicrobial peptides as a new class of innate immunity.....	20
1.4 The therapeutic ability of antimicrobial peptides and clinical usage.....	21
1.4.1 The therapeutic ability of antimicrobial peptides.....	22
1.4.2 The role of antimicrobial peptides in clinics.....	24
1.5 Biosynthesis and diversity of antimicrobial peptides.....	28

1.5.1	Biosynthesis of antimicrobial peptides.....	28
1.5.2	Diversity of antimicrobial peptides.....	29
1.6	Biophysical properties of antimicrobial peptides.....	30
1.6.1	Charge ( <i>Q</i> ).....	31
1.6.2	Hydrophobicity ( <i>H</i> ).....	32
1.6.3	Amphipathicity ( <i>A</i> ).....	32
1.6.4	Structure and conformations of antimicrobial peptides.....	33
1.6.4.1	The $\alpha$ -helical peptides.....	33
1.6.4.2	The $\beta$ -sheet peptides and small proteins.....	34
1.6.4.3	Peptides with irregular amino acids composition or extended peptides.....	36
1.6.4.4	Peptides with loops or macrocyclic cystine knot peptides.....	37
1.7	Antimicrobial peptide mechanism of action.....	38
1.7.1	The Barrel-stave mechanism.....	39
1.7.2	The Toroidal pore or the Wormhole mechanism.....	40
1.7.3	The Carpet mechanism.....	40
1.8	<i>In silico</i> discovery of antimicrobial peptides.....	43
1.8.1	Classification of antimicrobial peptides: establishment of databases.....	43
1.8.2	Computational tools used in model construction and antimicrobial peptide identification.....	46
1.8.2.1	Computational tools used in antimicrobial peptide model construction.....	47
1.8.2.2	Computational tools used for antimicrobial peptide model construction and motifs finding.....	49
1.8.3	Prediction of antimicrobial peptides three-dimensional structures.....	50
1.8.4	Prediction of protein-protein interaction using docking tools.....	52
1.9	Experimental approaches to determine antimicrobial peptide activity.....	54

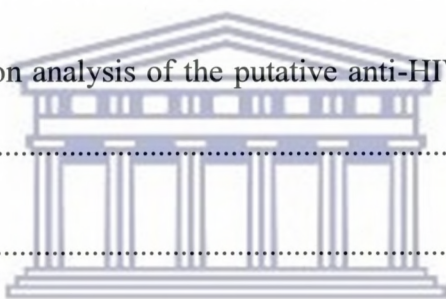
1.10	Rational of thesis.....	55
<b>2</b>	<b>Construction of HMMER and GLAM2 models for the identification of putative anti-HIV AMPs.....</b>	<b>59</b>
2.1	Introduction.....	59
2.2	Materials and Methods.....	64
2.2.1	Data mining: Extraction of anti-HIV antimicrobial peptides from the various databases.....	64
2.2.1.1	Antimicrobial peptide database (APD).....	64
2.2.1.2	Collection of Anti-Microbial Peptides (CAMP).....	65
2.2.1.3	Cybase.....	65
2.2.1.4	UniProt Knowledgebase (UniProtKB).....	66
2.2.2	Literature mining.....	66
2.2.3	Elimination of duplicates and generation of final anti-HIV AMP list.....	67
2.2.4	Construction of Hidden Markov Models (HMMER) and Gap Local Alignment of Motifs2 models (GLAM2).....	68
2.2.4.1	Hidden Markov Models profiles (HMMER).....	69
	a) Construction of Hidden Markov Models profiles.....	69
	b) Independent testing of Hidden Markov Models profiles.....	70
	c) Querying genome sequence databases using Hidden Markov Models profiles.....	71
2.2.4.2	Gap Local Alignment of Motifs 2 profiles (GLAM2).....	72



a) Construction of Gap Local Alignment of Motifs 2 profiles.....	72
b) Independent testing of Gap Local Alignment of Motifs 2 profiles.....	74
c) Querying genome sequence databases using Gap Local Alignment of Motifs 2 profiles.....	74
2.2.5 Performance measurement.....	75
2.3 Results.....	77
2.3.1 Data mining.....	77
2.3.2 Literature mining.....	77
2.3.3 Elimination of duplicates and generation of final anti-HIV list.....	78
2.3.4 Sequences from the final super-families used to create their respective models.....	80
2.3.5 Independent testing, performance measure and discovery of putative anti-HIV AMPs.....	87
2.3.5.1 Independent testing results.....	87
2.3.5.2 Performance measurement.....	89
2.3.5.3 Databases query and discovery of putative anti-HIV AMPs.....	90
a) Using Hidden Markov Models profiles.....	90
b) Using Gap Local Alignment of Motifs 2 profiles.....	94
2.4 Discussion.....	95
2.5 Conclusion.....	100



<b>3 3-D structure prediction and anti-HIV AMPs interaction study.....</b>	<b>103</b>
3.1 Introduction.....	103
3.2 Materials and methods.....	106
3.2.1 Selection of putative anti-HIV AMPs and HIV proteins.....	106
3.2.2 Physicochemical characterisation of the putative anti-HIV AMPs and HIV proteins.....	107
3.2.3 <i>De novo</i> structure predictions of putative anti-HIV antimicrobial peptides and HIV proteins.....	108
3.2.4 Docking and interaction analysis of the putative anti-HIV antimicrobial peptides and HIV proteins.....	110
3.3 Results.....	111
3.3.1 Physicochemical characterisation of the putative anti-HIV AMPs and HIV proteins.....	111
3.3.2 Prediction of the putative anti-HIV AMPs and HIV proteins 3-D structures.....	113
3.3.3 Superimposition of the predicted molecule 3-D structures with the known structures.....	115
3.3.4 Protein-protein interaction analysis of the anti-HIV AMPs and HIV proteins.....	116
3.4 Discussion.....	121
3.5 Conclusion.....	128
<b>4 General discussion and conclusion.....</b>	<b>130</b>



UNIVERSITY of the  
WESTERN CAPE



4.1 General discussion.....	130
4.2 Conclusion.....	136
4.3 Future work.....	138
<b>References.....</b>	<b>140</b>
<b>Appendix</b>	
<b>A Supplementary material for Chapter 2.....</b>	<b>179</b>
<b>B Supplementary material for Chapter 3.....</b>	<b>210</b>



# List of Figures

**Figure 1.1:** The different levels of HIV classification.

**Figure 1.2:** Global distribution map of HIV pandemic and their prevalence according to countries.

**Figure 1.3:** An overview of the major structural classes of host-defense peptides.

**Figure 1.4:** Different mechanisms of action used by Antimicrobial Peptides to enter their targets.

**Figure 1.5:** Alternative mode of action for intracellular Antimicrobial Peptide activity.

**Figure 2.1:** Architecture of the proposed method to build profiles using the profile HMMER algorithm.

**Figure 2.2:** Architecture of the proposed method to build profiles using the GLAM2 algorithm.

**Figure 3.1:** Cartoon representations of the 3-D structures predicted for 10 putative anti-HIV AMPs by the I-TASSER server.

**Figure 3.2:** Cartoon representations of the 3-D structures of the four HIV proteins predicted by the I-TASSER server.

**Figure 3.3:** Superimposition of each putative anti-HIV AMP, the positive and negative controls with the closest solve 3-D structure found in the Protein Data Bank.

**Figure 3.4:** HIV protein gp120 3-D structure predicted by I-TASSER superimposed with the known solved 3-D structure having the PDB ID 2b4cG.

**Figure 3.5:** HIV protein gp41 3-D structure predicted by I-TASSER superimposed with the known solved 3-D structure having the PDB ID 2cmrA.

**Figure 3.6:** HIV protein p24 3-D structure predicted by I-TASSER superimposed with the known solved 3-D structure having the PDB ID 3gv2A.

**Figure 3.7:** HIV protein p17 3-D structure predicted by I-TASSER superimposed with the known solved 3-D structure having the PDB ID 1l6nA1.

**Figure 3.8a:** p24-Molecule1 complex formation during anti-HIV-p24 interaction.

**Figure 3.8b:** p24-Molecule2 complex formation during anti-HIV-p24 interaction.

**Figure 3.8c:** p24-Molecule3 complex formation during anti-HIV-p24 interaction.

**Figure 3.8d:** p24-Molecule5 complex formation during anti-HIV-p24 interaction.

**Figure 3.8e:** p24-Molecule6 complex formation during anti-HIV-p24 interaction.

**Figure 3.8f:** p24-Molecule8 complex formation during anti-HIV-p24 interaction.

**Figure 3.9a:** gp120-molecule1 complex formation during anti-HIV-gp120 interaction.

**Figure 3.9b:** gp120-molecule3 complex formation during anti-HIV-gp120 interaction.

**Figure 3.9c:** gp120-molecule7 complex formation during anti-HIV-gp120 interaction.

**Figure 3.9d:** gp120-molecule8 complex formation during anti-HIV-gp120 interaction.

**Figure 3.9e:** gp120-molecule10 complex formation during anti-HIV-gp120 interaction.

## Supplementary Figures for Chapter 2

**Figure A.1:** HMMER Amphibians profile query results using the Amphibians testing set.

**Figure A.2:** HMMER Microorganisms profile query results using the Microorganisms testing set.

**Figure A.3:** HMMER Vertebrates profile query results using the Vertebrates testing set.

**Figure A.4:** HMMER Human defensins profile query results using the Human Defensins testing set.

**Figure A.5:** HMMER Fish and crabs profile query results using the Fish and crabs testing set.

**Figure A.6:** HMMER Insects profile query results using the Insects testing set.

**Figure A.7:** HMMER Plants profile query results using the Plants testing set.

**Figure A.8:** GLAM2 Plants profile query results using the Plants testing set.

**Figure A.9:** GLAM2 Vertebrates profile query results using the Vertebrates testing set.

**Figure A.10:** GLAM2 Microorganisms profile query results using the Microorganisms testing set.

**Figure A.11:** GLAM2 Fish and crabs profile query results using the Fish and crabs testing set.

**Figure A.12:** GLAM2 Insects profile query results using the Insects testing set.

**Figure A.13:** GLAM2 Amphibians profile query results using the Amphibians testing set.

**Figure A.14:** GLAM2 Human defensins profile query results using the Human defensins testing set.

**Figure A.15:** Results of the *Heliconius melpomene* genome query with the Insects training model.

**Figure A.16:** Results of the *Danaus plexippus* genome query with the Insects training model.

**Figure A.17:** Results of the *Homo sapiens* genome query with the Humans defensins training model.

**Figure A.18:** Results of the *Zea mays* genome query with the Plants training model.

**Figure A.19:** Results of the *Setaria italica* genome query with the Plants training model.

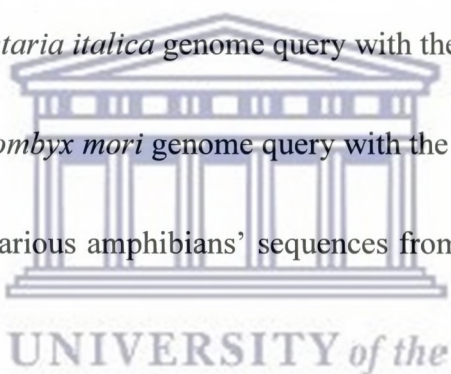
**Figure A.20:** Results of the *Bombyx mori* genome query with the Insects training model.

**Figure A.21:** Results of the various amphibians' sequences from UniprotKB query with the Amphibians training model.

**Figure A.22:** Results of the *Pelodiscus sinensis* genome query with the Vertebrates training model.

**Figure A.23:** Results of the *Taeniopygia guttata* genome query with the Vertebrates training model.

**Figure A.24:** Results of the *Ictidomys tridecemlineatus* genome query with the Vertebrates training model.



# List of Tables

**Table 1.1:** list of the frequently used and available FDA approved HAART in the pharmaceutical industry.

**Table 1.2:** HIV prevalence estimates and the number of people living with HIV in South Africa from 2001-2011.

**Table 1.3:** Number of HIV patients in need of HAART in South Africa from 2005-2011.

**Table 1.4:** AMPs drugs in development in clinical trials.

**Table 2.1:** The number of anti-HIV AMPs used for the construction and testing of each super-family model built by HMMER and GLAM2.

**Table 2.2:** Results of all the anti-HIV AMPs entries found into the query databases.

**Table 2.3:** Number of experimentally validated and predicted anti-HIV peptides retrieved from the databases.

**Table 2.4:** Final list of all experimentally validated and predicted anti-HIV peptides retrieved from the databases.

**Table 2.5:** Sequences of the experimentally validated anti-HIV AMPs used as the training set to create the family specific profiles.

**Table 2.6:** Sequences of the experimentally validated anti-HIV AMPs used as the testing set for the family specific profiles.

**Table 2.7:** Results obtained by querying the HMMER profile.

**Table 2.8:** Results obtained by querying the GLAM2 profile.

**Table 2.9:** Performance measurements generated for each super-family using the Model created by HMMER profile.

**Table 2.10:** Performance measurements generated for each super-family using the Model created by GLAM2 profile.

**Table 2.11:** Name and number of genome sequences predicted to have anti-HIV AMPs after scanning the genome sequence databases using the HMMER Model.

**Table 2.12:** Final list of all putative anti-HIV AMPs after removal of duplicate sequences and their classification according to the E-values.

**Table 2.13:** Number of putative anti-HIV AMPs obtained per model for both tools and the common putative anti-HIV AMPs found within the two tools.

**Table 3.1:** Physicochemical properties and parameters for the 10 putative anti-HIV AMPs and the positive control (Kn2-7) and the negative control (Mucroporin S1).

**Table 3.2:** Characterisation of the different physicochemical properties and parameters for the four HIV proteins.

**Table 3.3:** Quality assessment scores of the predicted 3-D structures of the putative anti-HIV AMPs and the positive and negative controls.

**Table 3.4:** Quality assessment scores of the predicted 3-D structures for the HIV proteins.

**Table 3.5:** Statistical explanation of the similarity between the predicted 3-D structure using the I-TASSER server and the 3-D structure of known proteins found in the PDB.

**Table 3.6:** Geometric scores of the binding affinity obtained for docking the anti-HIV AMPs to HIV proteins.

**Table 3.7:** The individual atomic interaction of the putative anti-HIV AMPs binding to the HIV protein p24.

**Table 3.8:** The individual atomic interaction of the putative anti-HIV AMPs binding to the HIV protein gp120.

### **Supplementary Tables for Chapter 2**

**Table A.1:** Results of all the experimentally validated anti-HIV peptides with their references.

**Table A.2:** List of all the single domains identified during the genome sequences scan of database by HMMER profiles and their classification according to their E-values.

### **Supplementary Tables for Chapter 3**

**Table B.1:** The 10 best putative anti-HIV AMPs with respect of their E-values, the positive control and negative control used in the 3-D structure prediction by the I-TASSER server.

**Table B.2:** Sequences of the HIV Proteins used for the 3-D structures prediction by the I-TASSER server.

**Table B.3:** Results of the docking parameters calculations used for the HIV proteins-anti-HIV AMP interaction studies.

**Table B.4:** Results of the docking parameters calculations used for the HIV proteins-anti-HIV AMP interaction studies.



## Chapter 1:

# Literature Review

### 1.1. Introduction

Since the discovery of the first case of an Human Immunodeficiency Virus-infected individual in 1981, the disease has been one of the major causes of death worldwide (Gallo, 2006). Acquired Immune deficiency Syndrome (AIDS) is a disease caused by the Human Immunodeficiency Virus (HIV) (Sepkowitz, 2001) which affects the immune system, causing the deterioration of the immune system in spite of the type of HIV virus (HIV-1 or HIV-2). Following the initial infection, an HIV infected person may have a prolonged asymptomatic period, followed by a variation of the CD4<sup>+</sup> T cells in the patient's blood, with an increase in CD4<sup>+</sup> T cell count at the initial stage of infection, followed by a decrease in CD4<sup>+</sup> T cell numbers and finally a breakdown of the patients immune system (Alimonti *et al.*, 2003). The vulnerability of the immune system encourages the emergence of opportunistic infections and various types of tumours in the human body which ultimately leads to the death of the HIV infection person.

With regards to the large distribution of the disease worldwide and the rate at which the HIV infection is spread, the UNAIDS has declared AIDS a pandemic (Kallings, 2008). The UNAIDS reported in 2009 that since the discovery of the disease, 60 million people have been infected with HIV, with 25 million deaths, and 14 million orphaned children in Southern

Africa alone. The World Health Organization (WHO) estimated in 2010 that there were 34 million people worldwide living with HIV/AIDS, with 2.7 million new HIV infections per year and 1.8 million annual deaths due to AIDS (UNAIDS, 2011). The high number of HIV-infected individuals is a result of unprotected sex with an infected person, blood transfusions and the use of hypodermic needles during drug use, the transmission from mother to child during pregnancy, delivery and breastfeeding (Markowitz, 2007). Currently there is no effective cure or vaccine for the disease but the HIV/AIDS infection is rather managed by a drug treatment regimen in addition to some preventive methods for infection and re-infection (Crosby and Bounse, 2012; Coutsooudis *et al.*, 2010).

The current treatment for AIDS consists of Highly Active Antiretroviral Therapy (HAART). The antiretroviral drugs however only slow the course of the disease and increase the life span of the individual infected with HIV (Dybul *et al.*, 2002; Anglemeyer *et al.*, 2011). Besides reducing mortality rates and disease progression, the antiretroviral drugs have some adverse side effects associated with the treatment. These include lipodystrophy syndrome, dyslipidaemia and diabetes mellitus (Mandell *et al.*, 2010), diarrhoea and an increased risk of cardiovascular disease (Montessori *et al.*, 2004; Burgoyne and Tan, 2008).

The lack of effective treatment has encouraged the development of additional therapeutic agents, with increased efficacy and to overcome the adverse side effects experienced by HAART treatment. Antimicrobial Peptides (AMPs) are excellent candidates as novel therapeutic agents as they have been reported to possess anti-HIV activities (Wang *et al.*, 2010).

AMPs are components of the first line of defence of the immune system of eukaryotes and prokaryotes (Brogden, 2005). They are small, positively charged, gene encoded peptides, which have selective toxicity towards gram-positive and gram-negative bacteria, protozoa,

fungi as well as viruses (Wang *et al.*, 2011; Brodgen, 2005; Andreu and Rivas, 1998). Their selective toxicity is due to the fact that the microbes' membrane bilayer is rich in lipopolysaccharides (LPS) and lipoteichoic acid (LPA) and is thus negatively charged in contrast to the positive charge of the AMPs (Ganz, 2003). A number of AMPs had been used as lead compounds for the development of several drugs which include Pexiganan which is used as a topical antibiotic, Omiganan which is used as a topical antiseptic and for the prevention of catheter infections, and Isegran which is used for oral mucositis in patients undergoing radiation therapy (Fjell *et al.*, 2012); whilst several AMPs have been shown to have anti-HIV activity. For example, LL-37, Circulin A, Circulin B, Cycloviolacin A, Cycloviolacin C, Cycloviolacin D, Cycloviolacin Y4, Cycloviolacin Y5, hBD-2 and hBD-3 (Ireland *et al.*, 2007; Wang *et al.*, 2008; Quiñones-Mateu *et al.*, 2003).

AMPs can be classified according to their various secondary structures,  $\alpha$ -helical,  $\beta$ -sheet, extended peptides and macrocyclic knot peptides, which they adopt upon interaction with the microorganism (Hancock and Lehrer, 1998). They can also be classified based on the different mechanism of action to penetrate the microbes. These mechanisms of action they use include the barrel-stave, the toroidal pore and the carpet mechanism (Brodgen, 2005).

Due to the high number of uncurated experimentally validated and non-experimentally proven Antimicrobial Peptides, repository databases have been created in which these molecules are classified based on their activities namely: anti-bacterial, anti-fungal, anti-viral or anti-HIV, anti-cancer and haemolytic (Wang and Wang, 2004; Mulvenna *et al.*, 2006; Wang *et al.*, 2008; Wang *et al.*, 2009; Thomas *et al.*, 2010). In addition, they are also classified according to their biological source, molecular properties and biosynthetic composition (Wang and Wang, 2004; Wang *et al.*, 2009). Furthermore, the data within these databases, because of their large volume, require the incorporation of computational tools. These tools embedded within these databases enable the proper classification and identification

of novel AMPs from sequence data. Thus with the flourishing of computational approaches, *in silico* as well as *de novo* building of molecules based on physical characteristics as well as chemical properties has made it possible to identify more AMPs (Torrent *et al.*, 2012).

Though a simple computational tool such as Basic Local Alignment Search Tool (BLAST) can help to identify similar AMP sequences, more sophisticated computational methods exist for designing and finding of putative and/or novel AMPs which include Support Vector Machine (SVM) (Thomas *et al.*, 2010), Quantitative Structure Activity Relationship (QSAR) (Fjell *et al.*, 2009) and the profile Hidden Markov Models (HMMER) (Brahmachary *et al.*, 2004; Fjell *et al.*, 2007). These machine learning tools take into consideration feature selection of the training dataset for the construction of an accurate and optimized stochastic model, which is then utilised for the retrieval and identification of novel/putative antimicrobial peptides from various genome sequence databases.

Besides the fact that the usage of computational biology has focused on the identification of putative Antimicrobial Peptides, more sophisticated tools have been designed to predict the 3-D structure of these peptides and their receptors (Schneidman-Duhovny *et al.*, 2005; Schwede *et al.*, 2008; Roy *et al.*, 2010).

The first section of this chapter will focus on the origin, classification, transmission, life cycle and infection of HIV, diagnosis, management and treatment of HIV, and HIV statistics. The second part will look at Antimicrobial Peptides, a class of peptides which acts as the first line of defence in many eukaryotic and prokaryotic organisms. The biosynthesis and diversity of AMPs will be highlighted, followed by their biophysical properties and mechanism of action. Lastly, the chapter will review their classification into repository databases, some computational tools which aid in the identification of AMPs and advanced tools which aid in their 3-D structure prediction and docking to other proteins.

## 1.2. HIV/AIDS

Human Immunodeficiency Virus (HIV) is a lentivirus, a member of the retrovirus family that causes Acquired Immune Deficiency Syndrome (AIDS), which results in the human immune system breakdown (Weiss, 1993). The Acquired Immune Deficiency Syndrome (AIDS) has become the fourth leading cause of death worldwide and the second leading cause of death in Africa after Malaria. The United Nations estimates that there are now 40 million people living with HIV/AIDS in the world (Quaranta *et al.*, 2012).

### 1.2.1. History

#### 1.2.1.1. Discovery

Since the first clinical case was reported in 1981, initially there have been many speculations about the exact causes of AIDS. The disease was first reported in drug injection users and gay men with no known weakened immune systems but presenting with symptoms of *Pneumocystis carinii* pneumonia (PCP). After these observations, many gay men developed a rare skin cancer namely Kaposi's sarcoma (KS) (Friedman-Kien, 1981; Hymes *et al.*, 1981; Gottlieb, 2006). Due to the fact that the disease was mostly reported in gay men, it led to the appellation of gay-related immune deficiency (GRID) by the press (Altman, 1982). However, after noticing that it was also associated with drug users and haemophiliacs and not limited to the gay community, the name GRID was changed to AIDS (Centers for Disease Control (CDC), 1982). The proper name of the cause of this disease will only come in 1983, with two research laboratories independently identifying a new virus from patients presenting with the AIDS disease and publishing their findings in the same journal Science (Gallo *et al.*, 1983; Barre-Sinoussi *et al.*, 1983). Both research groups led by Robert Gallo and Luc Montagnier gave different names to their findings. While Gallo gave the name human T-lymphotropic

viruses III (HTLVs), Montagnier named his virus lymphadenopathy-associated virus (LAV) but with both finally agreeing on the name HIV.

#### 1.2.1.2. HIV origins

The origin of this virus has been one of the most controversial issues during these last decades. HIV-1 is closely related genetically to the Simian Immunodeficiency Virus (SIVcpz) which infects the wild chimpanzee, a subspecies *Pan troglodytes troglodytes*. It is hypothesized that HIV-1 is derived from the SIVcpz through evolution (Gao *et al.*, 1999; Keele *et al.*, 2006). HIV-2 is closely related to SIVsmm, a virus of the Sooty Mangabey (*Cercocebus atys atys*) (Reeves and Doms, 2002). The controversy lies within the question: How did humans come into contact with the parent virus SIV and were they infected with this virus through the process of zoonosis? Many hypotheses were put forward and are enumerated as follows: (i) the hunter theory is the common theory put forward since scientist believe that humans could have been infected with the SIV by eating infected primates (Wolfe *et al.*, 2004); (ii) another theory focused on the fact that the vast program of Polio vaccine development was done using African primates and could have been transferred to humans during the administration of the oral polio vaccine (OPV) (Hooper, 1999). However, this theory was later rejected after analysis of some of the initial vaccine samples. The company behind the vaccine campaign however does not recognized that this vaccine might have contributed to the transmission of the HIV-1, so this theory remains questionable (Blancou *et al.*, 2001). OPV accounts only for the M-subtype of the HIV-1; (iii) the colonialism or “Heart of Darkness” theory also accounts for the understanding of how the cross-contamination from the SIV to the HIV took place. Even though the theory appears to be recent, it is also taken into consideration, to explain the SIV evolution to HIV. This theory

is based on the basic “hunter” hypothesis. It is believed that between the late 19<sup>th</sup> and the early 20<sup>th</sup> century, most African countries ruled by colonial regimes were forced into labour camps where sanitation was poor and food was scarce. These workers might have eaten primates to survive. But due to their poor health, the SIV in the primates could have easily infiltrated their weakened immune systems to become HIV. Nonetheless, the link between this theory and the origin of HIV has never been established (Chitnis *et. al.*, 2000); (iv) though the contaminated needle theory is put forward as a contributor of the crossover of species, it is hypothesized that cross-contamination with needles have occurred between humans during the treatment of patients in poor hygienic conditions with disposable plastic syringes in the 1950s. Without being more controversial, the following questions can be posed: Why was the first case of HIV not reported in these African hospitals at the time? And why was HIV only observed in 1981 in the USA and not in these reported African countries? (v) The last theory is the conspiracy theory, which believes that the virus has been man-made. However, this theory is only based on speculation and supposition (Fears, 2005). In addition to the main theories stated above, other theories such as travelling, blood transfusions and drug users have emerged. It is difficult to address the issue regarding the evolution of SIV to the HIV and the answer should be sought in a scientific perspective rather than speculation.

### **1.2.1.3. Classification of HIV**

Although it is known that there are two types of HIV, HIV-1 and HIV-2; the difference between the two types lies in the fact that the HIV-2 is predominantly found in West Africa and rarely elsewhere. It is not easily transmitted and takes a longer period between the infection and illness manifestation. However, HIV-1 is found worldwide and it is generally used by people to refer to HIV without considering a specific type of the virus. The HIV-1 subtype is subdivided into four groups: the "major" group M, the "outlier" group O and two

new groups, N and P. The M group is the so-called major group because more than 90 percent of HIV-1 infections belong to this group and it comprises of at least nine known genetically distinct subtypes of HIV-1. These subtypes include A, B, C, D, F, G, H, J and K of HIV-1 (WHO, 2011). The Circulating Recombinant Forms (CRFs) are regarded as a sub-group because it may comprise a mixture of A and B subtypes or A and F subtypes. This confusion is based on the fact that it is a fusion of two viral sub-groups through the process of “viral sex”, thus the CRFs should be classified as a sub-group on his own (Bailes *et al.*, 2003). Whilst the group O is known to be restricted only to West Africa, the groups N and P were recently discovered as new groups belonging to the HIV-1 virus (Plantier *et al.*, 2009). Figure 1.1 shows the classification of the various HIV types.



**Figure 1.1:** The different levels of HIV classification.

### 1.2.2. Transmission and life cycle of HIV

The major mode of HIV transmission is through unprotected sexual relations with an infected partner of the same or opposite sex. Other ways of HIV transmission include: body fluids mostly during blood transfusions, sharing a needle during drug injection and mother-to-child transmission during pregnancy, delivery or during breastfeeding after delivery (Markowitz, 2007, Coutsooudis *et al.*, 2010).

Once the HIV virus has been transmitted to a healthy individual through body fluids or via mucosal surfaces, HIV initiates its entry into the macrophages, monocytes and the CD4+ T



lymphocytes by attaching via unspecific binding to the lectins, glycosaminoglycans or via receptor-ligand interaction of the viral surface proteins (Stolp and Fackler, 2011). This entry is made possible by the interaction of the viral envelope glycoprotein 120 (gp120) to the cell surface of the CD4+ molecules and a chemokine receptor (either CXCR4 or CCR5) on the viral cell surface (Chinen and Shearer, 2002). Upon binding to the cell surface of the CD4+ T cells, the viral gp120 molecule undergoes a conformational change, which will allow the exposure of the chemokine receptor binding domains of gp120. When exposed, the gp120 molecule will then bind and form a more stable complex with these chemokines (Stolp and Fackler, 2011).

The stable complex formed by the conformational change will allow the N-terminal fusion peptide glycoprotein 41 (gp41) to penetrate the cell's membrane then engage, causing the formation of a heparin sulphate compound on the host plasma membrane. The loop structure encourages the membrane and the virus to come closer, leading to the fusion of membranes and consequently the release of the viral capsid into the cytoplasm of T cells (Chinen and Shearer, 2002; Chen *et al.*, 2009). Once fused, the viral RNA, reverse transcriptase, integrase, ribonuclease and protease are infused into the cell. The single stranded HIV RNA is reverse transcribed to double-stranded viral DNA, which is carried across the host nucleus and where the viral DNA integrates into the human DNA, resulting in the formation of a new viral RNA which is subsequently used to produce genomic RNA and other HIV proteins. The immature HIV matures by proteases, releasing individual HIV proteins into the circulating system and other proteins such as p24, p17 and p9 needed for virion assembly (Chinen and Shearer, 2002).

### 1.2.3. HIV infection

Following transmission of the HIV, the infection is characterised by an acute phase and a chronic phase which might lead to AIDS. During HIV infection, both the innate as well as the adaptive immune responses are raised. These responses however are insufficient or too late to eliminate the virus. Once the host body has been infected with this virus, there is a drop in the CD4<sup>+</sup> T cell count, which corresponds to an increase in the viral load with exponential growth kinetics (Cadogan and Dalgleish, 2008; Quaranta *et al.*, 2012). The peak of this growth coincides with the onset of a strong host immune response, resulting in a decrease of the viral load and an increase in the number of circulating virus-specific CD4<sup>+</sup> T cells. The acute phase of an HIV infection is accompanied by a selective and dramatic depletion of CD4<sup>+</sup>CCR5<sup>+</sup> memory T cells principally from the mucosal surfaces. This step is an irreversible process which ultimately causes the failure of the host immune system and the inefficiency of clearing the HIV infection (Warrilow *et al.*, 2007).

The inefficiency of the immune system allows the HIV to institute a life-long latency and chronic infection. However, the viral load remains stable during this phase while the CD4<sup>+</sup> T cells levels decline, and without medical intervention, the infection progresses to the symptomatic phase which is characterised by a rapid decrease of both CD4<sup>+</sup> and CD8<sup>+</sup> T cells and an increased viral load in the patient. Decreased number of the CD4<sup>+</sup> and CD8<sup>+</sup> T cells levels, favours the emergence of opportunistic infections in the patient and consequently leads to AIDS. Additionally, the very same cells and responses aimed to eliminate HIV seems to play deleterious roles by driving chronic immune activation, which plays a central role in the immunopathogenesis and progression to AIDS (Deeks and Walker, 2004; Quaranta *et al.*, 2012).

#### 1.2.4. Immune evasion of HIV

Since the first step of HIV infection is the attachment to the surface of the T cells with the help of the envelope glycoprotein gp120, the primary humoral response to this infection is to target the envelope glycoprotein gp120. Nevertheless, the HIV has developed various mechanisms to avoid the effect of the neutralizing epitopes within the antibodies. The reasons for this failure may be explained by these neutralizing epitopes being hidden within the protein structure of the molecule and only being exposed momentarily during the binding onto the HIV gp120. As such, a high binding affinity is required for the neutralizing antibodies to bind to the target molecules and compete with these natural ligands. Thus, the high affinity triggers effective binding of the neutralizing antibodies to HIV. Also, the major neutralizing epitopes are protected by protein glycans by formation of a shield to provide steric hindrance against anti-gp120 interaction (Back *et al.*, 1994). Another reason for the ineffectiveness of the neutralizing antibodies might also be the high mutation rate of the glycoprotein. These mutations confer resistance to antibodies by point mutation on the V2/V3 loop and N-linked carbohydrate glycans on gp120 and hinder the ability of the neutralizing antibodies (Quaranta *et al.*, 2012).

The HIV mutations, at targeted epitopes, are to avoid MHC restricted recognition by decreasing the binding affinity of the epitopes to the presenting protein MHC-1, because of the down-regulated expression of MHC-1 molecules on the surface of these cells (Goulder *et al.*, 2001). Evidence of the HIV evading the host immune system (the innate and adaptive immune response) is indicated by the functional impairment of Dendritic Cells (DCs) during HIV-1 infection. This impairment contributes to a lack of efficient priming of adaptive immunity, which is caused by either direct viral interaction with DCs or as a result of indirect mechanisms. Such indirect mechanisms can include either the production of interleukin-10

(IL-10) by monocytes during the HIV infection or the activation of plasmacytoid dendritic cell precursors (pDCs) by HIV-1, which produce type I Interferon (IF) which besides killing the virus also contribute to CD4+ T cell death (Quaranta *et al.*, 2012).

### **1.2.5. Diagnosis of HIV/AIDS**

There are currently two approaches for the diagnosis of HIV. It is either by monitoring the viral load within the plasma which indicates viral replication and virus particles in the infected individuals or by looking for the antibodies which are produced to resist the viral infection. While the diagnostic system is able to detect these molecules in the blood of an infected individual, the main goal is to detect with confirmation these molecules at the onset of infection. As a consequence, early medications can be prescribed to prevent the replication of the virus and consequently slow the progression of the virus.

The diagnostic approach to detect antibodies comprises of the enzyme immunoassay (EIA) followed by western blotting (or immune blot) and the rapid test. EIA is the gold standard for HIV detection. The first generation EIAs, detects immunoglobulin G (IgG) antibodies where as the third generation EIAs detects immunoglobulin M (IgM) antibodies produced in response to HIV infection. The antibody binds to various HIV antigens and indicates a positive test (Hauck *et al.*, 2010). Western blotting serves as a confirmatory test because the same antibodies are utilised to detect the HIV antigens during this assay. Unlike the EIAs, the result of this complex formation is run on a gel matrix which is separated by an electric current. The proteins on the gel are transferred onto a nitrocellulose membrane, which is then coated with radio-labelled antibodies specific to these proteins and the result is visualised by autoradiography.

The rapid tests with traditional enzyme immunoassays have been developed for HIV antibody detection and have comparable sensitivities (98%-100%) and specificities (86%-100%)

(Hauck *et al.*, 2010). However, the detection of IgG and IgM antibodies is only possible 3-6 weeks after infection whereas the viral RNA detection is possible within 9 days after HIV infection (Hauck *et al.*, 2010). Thus the sensitivity and specificity of the techniques ought to be improved for better and earlier detection.

The diagnostic approach to detect viral replication or virus particles include the Ultrasensitive p24 assay, the ExaVir<sup>TM</sup> RT viral load and Real Time Reverse Transcription quantitative Polymerase Chain Reaction (RT-qPCR). The Ultrasensitive p24 assay makes use of a specific antibody which binds to the p24 viral core protein in a serological sample via a EIA technique to diagnose an individual but also requires a window period of a few days, an average of 7 days, and the result of the test is received within an hour. However, the results of the test are insensitive in comparison with the nucleic acid based assay (Buttò *et al.*, 2010).

The ExaVir<sup>TM</sup> RT viral load and RT-qPCR techniques amplify the HIV RNA extracted from the patients' blood. The method operates by converting the HIV RNA into DNA via an enzyme reverse transcriptase which makes it easier for quantification of the amount of viral particles present (Wang *et al.*, 2010).

Whist these techniques are important for the diagnosis of HIV, they require a window period after infection has occurred, to be able to detect either the antibodies or the viral particles within a serological sample. Most of these assays such as the ExaVir<sup>TM</sup> RT viral load, RT-qPCR, the EIAs and western blots are expensive and require specialized trained staff, regular and expensive maintenance of equipment, are laborious, and suitable for centralized laboratories, but not for district clinics in resource-limited settings (Wang *et al.*, 2010). Inexpensive techniques such as the Ultrasensitive p24 assay and rapid tests though insensitive fulfil the point-of-care purpose for a proper global diagnostics system. However, a more sensitive and specific technique ought to be put in place so as to reduce the window period, a common drawback of all the above-mentioned assays.

### 1.2.6. Prevention of HIV/AIDS, Management and treatment

Even though the disease prevalence is still high around the globe and mostly in Africa, many groups working on the awareness of HIV/AIDS propose many preventative methods to avoid the spread and transmission of HIV.

Some of these preventative methods include the use of condoms for males and females to avoid sexual contact with an infected individual (Crosby and Bounse, 2012). While it is advised to circumcise male children, it is said that this practice reduces the risk of HIV infection in Sub-Saharan Africa heterosexual partners (Siegfried *et al.*, 2009). Other preventive methods include the mother-to-child programs which reduce the transmission of the virus during pregnancy (Coutsoudis *et al.*, 2010) and bottle feeding after birth rather than breastfeeding (Siegfried *et al.*, 2009; Coutsooudis *et al.*, 2010). Considerable effort is also dedicated to vaccine development as a preventative method against HIV/AIDS (Reynell and Trkola, 2012). To date, no effective vaccine exists as a preventative tool for the HIV pandemic (UNAIDS, 2012). However, medications can be administered to pre-exposed and post-exposed individuals to the HI Virus (UNAIDS, 2012; No authors listed, 2012).

Despite the efforts of many researchers, there is neither a cure nor a vaccine to date. The infected patient has to rely on high active antiretroviral therapy (HAART), a cocktail of medication, which reduces the progression of the virus but does not eradicate it. This looks promising because the HAART prolongs the lifespan of the infected patient, thus it is a dramatic advancement in pharmacological intervention of this infectious virus (Powderly, 2000). The HAART is designed to target the various molecules in the life cycle of HIV and is composed of approved drugs targeting each of the viral enzymes: Reverse Transcriptase (RT), Protease Inhibitors (PI) and Integrase Inhibitors (IN). The Reverse Transcriptase (RT) is sub-divided into non-nucleoside reverse transcriptase inhibitors (NNRTIs) and nucleoside analogue reverse transcriptase inhibitors (NRTIs) (WHO, 2010). The initial treatment

classically makes use of RTs, with a combination of RTs and PIs given to the patients when the RT therapy becomes ineffective. The most prominent NRTIs includes: Zidovudine (AZT) or Tenofovir (TDF) and Lamivudine (3TC) or Emtricitabine (FTC) (WHO, 2010). Table 1.1 represents the frequently used HAART combinations for HIV/AIDS management. Even though these drugs can extend the patient's survival and look promising, the patient has to rely on HAART lifelong as it does not eradicate the virus but only slows its progression (Marsden and Zack, 2009). However, it comes with adverse side effects and some of the side effects related to the intake of these drugs include: microalbuminuria dyslipidaemia, insulin resistance, impaired glucose tolerance and increase risk of cardiovascular disease (Dimock *et al.*, 2011; Barbaro and Barbarini, 2011).

**Table 1.1:** list of the frequently used and available FDA approved HAART in the pharmaceutical industry

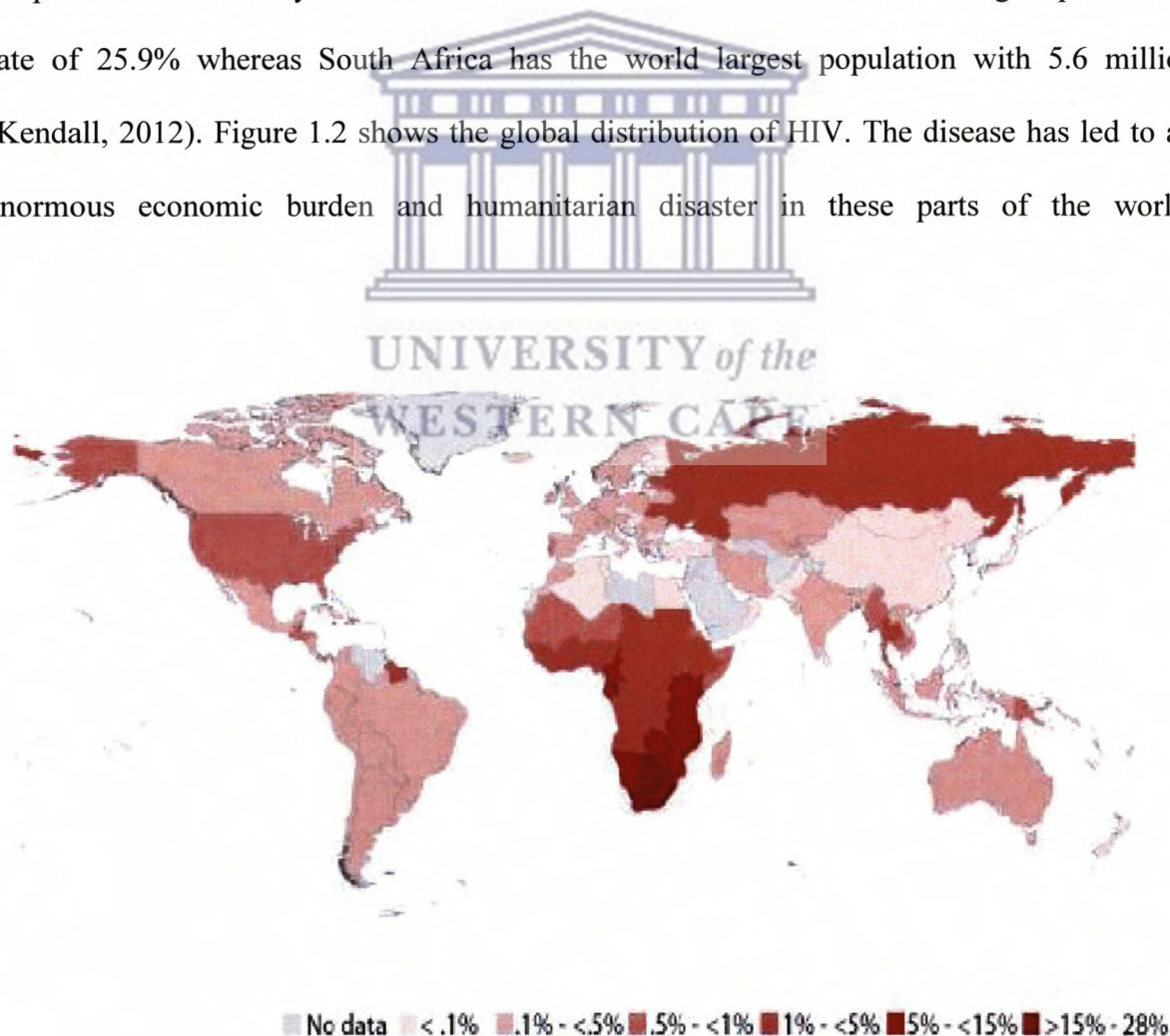
HAART classes	Abbreviation	Generic name	Brand name	Food restrictions and notes	Date of FDA approval
NRTIs	FTC	Emtricitabine	Emtriva	Take with or without food	02-Jul-03
	TDF	Tenofovir	Viread	Take with or without food	26-Oct-01
	3TC	Lamivudine	Epivir	Take with or without food	17-Nov-95
	AZT or ZDV	Zidovudine	Retrovir	Take with or without food	19-Mar-87
NNRTIs	ETR	Etravirine	Intelence	Take following a meal	18-Jan-08
	DLV	Delavirdine	Rescriptor	Take with or without food	04-Apr-97
	NVP	Nevirapine	Viramune	Take with or without food	21-Jun-96
PIs	APV	Amprenavir	Agenerase	Take with or without food; avoid high-fat meals	15-Apr-99
	ATV	Atazanavir	Reyataz	Take with food	20-Jun-03
	DRV	Darunavir	Prezista	Take with food	23-Jun-06
INs	RAL	Raltegravir	Isentress	Take with or without food	12-Oct-07

NRTIs: Nucleoside/Nucleotide Reverse Transcriptase Inhibitors, PIs: Protease Inhibitors

NNRTIs: Non-Nucleoside Reverse Transcriptase Inhibitors, INs: Integrase Inhibitors

### 1.2.7. Estimation of HIV statistics

The window period, the inefficacy of the diagnostic methods and the therapeutic prescriptions may have contributed to the increased statistics of infected individuals around the globe. Consequently, the disease has been considered a pandemic by the United Nations and they estimated that since the discovery of the disease in 1981, 34 million people are living with HIV/AIDS and 30 million HIV-related deaths as in 2009 (Wang *et al.*, 2010). This figure has now risen to 40 million people living with HIV/AIDS (Quaranta *et al.*, 2012). A total of 94% of infected people worldwide are from Sub-Saharan Africa, South and Southeast Asia, Latin America, Eastern Europe and Central Asia with 60% of the infected people accounted for by Sub-Saharan Africa. Swaziland has the world's largest prevalence rate of 25.9% whereas South Africa has the world largest population with 5.6 million (Kendall, 2012). Figure 1.2 shows the global distribution of HIV. The disease has led to an enormous economic burden and humanitarian disaster in these parts of the world.



**Figure 1.2:** Global distribution map of HIV pandemic and their prevalence according to country.



According to the current statistics, the number of HIV patients has increased as from 2001 in South Africa and, the country has one of the highest levels of prevalence in the world besides having the world's largest population living with HIV. Table 1.2 shows the statistics of HIV prevalence estimates as from 2001 to 2011 in South Africa whilst Table 1.3 shows the number of HIV infected patients receiving HAART treatment in South Africa (Statistics South Africa, 2011).

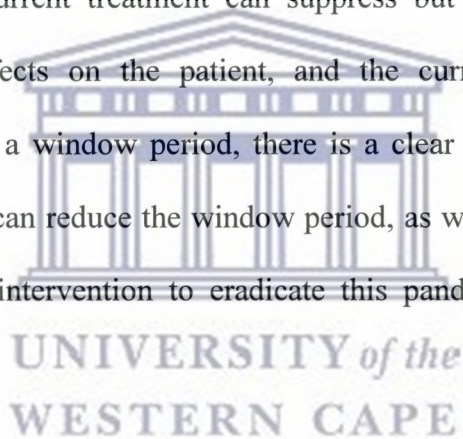
**Table 1.2:** HIV prevalence estimates and the number of people living with HIV in South Africa from 2001-2011 (Statistics South Africa, 2011)

HIV prevalence estimates and the number of people living with HIV, 2001–2011 Prevalence				Incidence Adult 15-49	HIV population (millions)
Years	Women 15-49	Adult 15-49	Total population %		
2001	17,4	16,0	9,4	1,72	4,21
2002	17,7	16,2	9,6	1,59	4,37
2003	18,0	16,2	9,7	1,58	4,49
2004	18,1	16,2	9,8	1,63	4,59
2005	18,3	16,2	9,9	1,73	4,69
2006	18,9	16,6	10,2	2,11	4,87
2007	18,9	16,5	10,2	1,54	4,95
2008	18,9	16,4	10,3	1,43	5,02
2009	19,1	16,4	10,4	1,45	5,13
2010	19,3	16,5	10,5	1,43	5,26
2011	19,4	16,6	10,6	1,38	5,38

**Table 1.3:** Number of HIV patients in need of HAART in South Africa from 2005-2011. (Statistics South Africa, 2011)

Number of persons in need of ART, 2005–2011 Year	Adults (15+ years)	Children (0–14)
2005	54 104	199 636
2006	163 017	215 042
2007	306 598	260 519
2008	504 809	270 024
2009	732 809	282 646
2010	966 266	368 357
2011	1 115 284	377 097

Due to the fact that the current treatment can suppress but not eradicate the HIV, has numerous adverse side effects on the patient, and the current diagnostic systems are insensitive since it requires a window period, there is a clear need for new and improved early detection tools which can reduce the window period, as well as a need for a non-toxic, more effective therapeutic intervention to eradicate this pandemic virus as soon as it is detected.

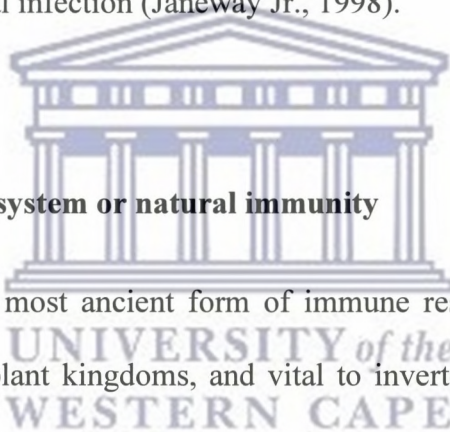


### **1.3. Antimicrobial Peptides: The defence system of all living organisms**

All living organisms are exposed daily to various potential pathogenic organisms and their survival is the result of the immune response defence mechanisms that they possess. These defence mechanisms are the adaptive immune response and the innate immune response. The study of the immune system and its responses to invading pathogens is called Immunology and the Immune system is the collection of cells, tissues and molecules that mediate resistance to infections (Abbas *et al.*, 2006).

### **1.3.1. The adaptive immune system or acquired immunity**

The adaptive immune response or acquired immunity is slower and exerts specific responses to microbes. This immune system involves majorly the use of antibodies created by B cells thus is known as the humoral immunity. However, when the cells involved in the defence of the organism are produced by T cells, the adaptive immunity is said to be a cell-mediated immune response. The primary response to pathogens invasion requires a lag time between exposure and the maximal response to the pathogen, hence the slow response of this defence system. The exposure leads to immunological memory and the response to pathogens is gene rearrangement since this response is imperfect. The body has to rely on the innate immune response to prevent additional infection (Janeway Jr., 1998).



### **1.3.2. The innate immune system or natural immunity**

The innate immunity is the most ancient form of immune response protection, conserved throughout the animal and plant kingdoms, and vital to invertebrate host defences because they lack potent antibodies to combat invasive pathogens and is a cell-mediated immune response (Hancock and Diamond, 2000). The innate immune system is the first line of defence of the immune system which is present in many organisms. This defence system is prepared to defend the body against any infectious attacker at any time. The innate immune system is native, rapid and an antigen-independent response, thus innate immunity is immediate, non-specific and diverse (Boman, 1991; Gallo and Nizet 2003). Though the adaptive immune responses share to some extent the same characteristics with the innate immune response, both immune systems are present in vertebrates.

The inducible effectors of the innate immune response differ from those of the adaptive immune response in that they are relatively non-specific; and have conserved the molecular patterns of recognition of stimulatory molecules, which include bacterial lipopolysaccharide (LPS) and lipoteichoic acid (LPA). Hence the inducible effectors of the innate immune system are rapidly induced within minutes to hours (Ganz, 2003). The effectors of the innate immune response include the phagocytic cells such as the neutrophils and macrophages, the leukocytic cells and the lymphocytic cells. These effectors have been deemed incomplete for the innate response and that a group of Antimicrobial Peptides, mostly the cationic Antimicrobial Peptides are one of the major players contributing to the innate immune response (Zasloff, 1992; Ganz and Lehrer, 1995; Ganz, 2003).



### **1.3.3. Antimicrobial peptides a new class of innate immunity**

The Antimicrobial Peptides (AMPs) were recently discovered to be part of the innate immune response in multi-cellular organisms and many have demonstrated direct antimicrobial activity against gram-positive and gram-negative bacteria, fungi, eukaryotic parasites as well as viruses (Shai, 2002). Thus they constitute the first line of the body's defence immune system. Their presence have been reported in many animal and plant species (Zheng *et al.*, 2002; Wong and Ng, 2003; Wang and Ng, 2005).

In spite of their structural diversity, most of the Antimicrobial Peptides are cationic (positively charged). However, several anionic Antimicrobial Peptides are present in plants and animals. A few examples of the anionic AMPs include the polyAsp-containing fragments from the ovine pulmonary surfactant and propieces of ovine trypsinogen and PYLa frog activation peptide. A common feature of the cationic Antimicrobial Peptides is that they all have an amphipathic structure which allows them to attract and bind to the negative charge

on the microbial membrane interface. Most Antimicrobial Peptides interact with the microbes' membranes, and may be cytotoxic as a result of disturbance of the bacterial inner and outer membranes (Erand and Vogel, 1999).

Antimicrobial Peptides are relatively small molecules of approximately 6 to 100 amino acid residues, with an overall positive charge (generally +2 to +9) as a result of the presence of multiple Lysine, Histidine and Arginine amino acid residues and, a substantial portion ( $\geq 30\%$  or more) of hydrophobic amino acid residues (Giuliani *et al.*, 2007). The hydrophobic nature of Antimicrobial Peptides, facilitates the folding of the peptides into an amphipathic structure in three dimensions, allowing their permeability into the microorganism upon contact, and avoids resistance by the microbes (Andreu and Rivas, 1999).

The cationic Antimicrobial Peptides are diverse and it arises from the different antimicrobial functions as well as the different pathogenic microbes. As such, several Antimicrobial Peptides involved in immunity have showed similar rapid evolution to the host-defense peptides and a single Antimicrobial Peptide might have therapeutic abilities across a vast range of pathogens (Hancock and Sahl, 2006). Antimicrobial Peptides can be a new source for novel therapeutic drugs, due to their low toxicity and the capability of these peptides not to induce resistance by the microorganism, unlike conventional antibiotics drugs which the microorganism may develop resistance to with time. This is due to the molecular composition of these peptides in comparison to their target membranes (Fjell *et al.*, 2012).

#### **1.4. The therapeutic ability of antimicrobial peptides and clinical usage**

Several studies had documented the importance of Antimicrobial Peptides in invertebrate and vertebrate innate immune systems, acting as therapeutic molecules to prevent the invasion of the organism by diverse pathogens (Zasloff, 2002).

#### 1.4.1. The therapeutic ability of antimicrobial peptides

These peptides have played an important role in the innate immune system as well as in the adaptive immune response of some organisms as they are produced immediately following a microbial challenge, to rapidly counteract a wide range of microorganisms (Chertov *et al.*, 2000). Antimicrobial Peptides are produced in different parts of an organism (Zasloff, 2002).

Examples of the activity of these peptides can be demonstrated against various microorganisms which infect multicellular organisms. Histatin, a Histidine-rich natural defensin protein predominantly found in the saliva of humans has shown strong activity against *Candida* microbes (Xu *et al.*, 1991). Histatin also reduces *Candida* microbes' adhesion to dentures (Edgerton *et al.*, 1995) and has also been shown to inhibit the activity of proteinase K from *Bacteroides gingivalis* (Nishikata *et al.*, 1991). Increased circulating levels of cathelicidin (PR-39) Antimicrobial Peptide have been noticed in pigs as a result of infection with *Salmonella*, showing up to a threefold increase of this peptide (Zhang *et al.*, 1997). On the other hand, an increase expression of  $\beta$ -defensin levels was registered in the intestines of cows that tested positive for *Mycobacterium paratuberculosis* infection (Stolzenberg *et al.*, 1997). High expression level of human  $\beta$ -defensin-2 was recorded in humans with an inflamed intestinal epithelium, relative to the normal colon (Scott *et al.*, 2002). The same was also observed for this Antimicrobial Peptide in inflamed gingival epithelium (Fukumoto *et al.*, 2005). Many Antimicrobial Peptides as well as Polyphemusin, a potent Antimicrobial Peptide from the horseshoe crab have been shown to cross the lipid bilayer of microorganism and destroy anionic intracellular targets of the microbes (Zhang *et al.*, 2000; Kragol *et al.*, 2001; Patrzykat *et al.*, 2002).

Antimicrobial Peptides, Cecropins and their analogues as well as Magainin 2 and their analogues have been well described for their ability to react positively toward transformed cells (Chen *et al.*, 1990; Baker *et al.*, 1993; Moore *et al.*, 1994). Examples of well studied

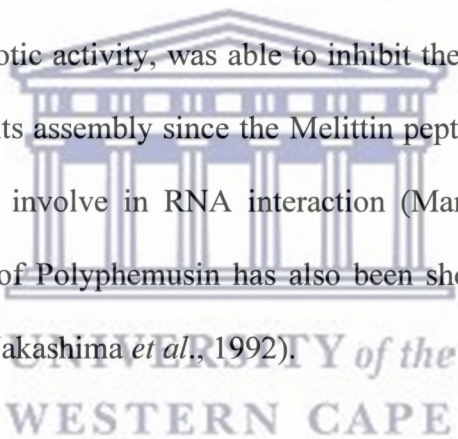
Antimicrobial Peptide efficacy on murine tumors include the all-D MSI-511 Magainin analogues, which completely alleviated an induced melanoma in athymic nude mice (Soballe *et al.*, 1995). In addition, a Magainin analogue, MSI-238 increased the lifespan by 100% of mice with induced ascites and spontaneous ovarian tumours, subjected to an intraperitoneal treatment of the Antimicrobial Peptide (Baker *et al.*, 1993).

Certain Antimicrobial Peptides have also been implicated in ocular infections, and have shown some promising activity. The rabbit (defensin) NP-1, Magainins, Cecropin-derived and Cecropin D5C, have been shown to play a role as the preserving media for cornea storage, contact lens disinfectants or ocular antiseptics (Schuster *et al.*, 1992; Sousa *et al.*, 1996; Schwab *et al.*, 1992). The Cecropin A-melittin hybrids have proven to be effective in rabbits infected with *Pseudomonas aeruginosa* in an *in-vivo* experiment, and more effective as gentamycin in the clearance of the infection (Nos-Barberá *et al.*, 1997).

Some Antimicrobial Peptides have inhibitory activity on spermatozooids such as Magainins and their analogues (De Waal *et al.*, 1991), and can be used as contraceptive agents. Other Antimicrobial Peptides, Protegrins have therapeutic activity against several sexually transmitted diseases, including the HI Virus; and as such could be considered as a possible combination of an antibiotic as well as contraceptive (Qu *et al.*, 1996). A group of HIV patients with lower salivary levels of Histatin peptides showed higher incidence of oral candidiasis and fungal infection (Andreu and Rivas, 1998).

An important number of Antimicrobial Peptides have also been mentioned for their inhibitory ability to neutralize viruses, mostly the human immunodeficiency virus (Wang *et al.*, 2010). The actions of anti-HIV AMPs have been categorized into three levels of activity. The first level of anti-HIV AMP activity is binding to the viral membrane or particles, through electrostatic interaction. This is due to the positive charge of the peptide and the negative

charge of the viral envelope (Andreu and Rivas, 1998). Herpes virus has been targeted using this approach, by an  $\alpha$ -defensin Antimicrobial Peptide, the amphipathic model peptide Modelin-1 and a Melittin analogue Hecate (Daher *et al.*, 1986; Aboudy *et al.*, 1994; Baghian *et al.*, 1997). Also the inhibition of the HIV by Polyphemusins and their analogues (Tamamura *et al.*, 1994) and the stomatitis virus by Tachyplesin I (Murakami *et al.*, 1991) have been described. The second level of anti-viral inhibiting activity has been proposed to be the inhibition of virion production. Examples are Melittin and Cecropin A peptides that have been shown to inhibit the HI Virus through this mechanism (Wachinger *et al.*, 1992; Wachinger *et al.*, 1998). The third level of anti-viral activity exhibited by AMPs is by mimicking the viral infective processes. Using this approach, Melittin and its subK71 analogue, which lacks antibiotic activity, was able to inhibit the infective ability of Tobacco Mosaic Virus by perturbing its assembly since the Melittin peptide is similar to the Tobacco Mosaic Virus capsid region involve in RNA interaction (Marcos *et al.*, 1995). The T22 [Tyr.<sup>5,12</sup> Lys<sup>7</sup>], an analogue of Polyphemusin has also been shown to inhibit the HI Virus, using the mimicry method (Nakashima *et al.*, 1992).



#### **1.4.2. The role of antimicrobial peptides in clinics**

Regarding the growing resistance of current antibiotics to treat the existing diseases and health conditions, and the widespread knowledge about antimicrobial development in the past decade, Antimicrobial Peptides might be the root in designing novel therapeutic agents against these diseases. The uses of such therapeutical applications have been largely envisaged in the treatment of bacterial infections (Chopra, 1993); viral infections (Aboudy *et al.*, 1995) and cancer (Soballe *et al.*, 1995). Studies both in the laboratory and in the clinic, confirmed that the emergence of resistance against Antimicrobial Peptides is less probable than that observed for conventional antibiotics, and provided both naturally occurring and



laboratory conceived therapeutically (synthetic) useful agents (Zasloff, 2002). Additionally, their development might not be impeded as systematic therapies as their activity showed an increased ability to inhibit microorganisms. In addition, their low toxicity makes them attractive candidates as therapeutic agents (Hancock and Diamond, 2000; Hiemstra *et al.*, 2004). However, some Antimicrobial Peptides have effective concentrations in animal models of infection which are often close to the toxic concentration of these peptides (Darveau *et al.*, 1991), thus, making them unsafe for peptide based drug design.

Knowledge of functional active sites located on different regions of the peptides has enabled the development of synthetic peptides with potent and specific antimicrobial functions. This is done by mutational alteration of these amino acid residues at these active sites (Powers and Hancock, 2003; Braff *et al.*, 2005; Chen *et al.*, 2012). Indeed, several Antimicrobial Peptides have shown some promising therapeutic indices, suitable for drug design; and many companies have pursued the launch of Antimicrobial Peptides with the aim of pharmaceutical development (Giuliani *et al.*, 2007). Several naturally occurring AMPs or their synthetic counterparts have entered clinical trials and at least 15 peptides or mimetics are in (or have completed) clinical trials as antimicrobial or immune-modulatory agents for a broad range of diseases (Fjell *et al.*, 2012). In spite of over two decades of pharmaceutical design of peptide-based drugs, few Antimicrobial Peptides have shown promise in clinical trials, and most of the peptides were removed or stopped at the phase I or II of the clinical trial (Zhang and Falla, 2006). This may be due to the fact that the cost of manufacturing these Antimicrobial Peptides has elevated and the peptides did not produce the expected therapeutic activity (Marr *et al.*, 2006).

Conversely, some Antimicrobial Peptides have advanced to the pharmaceutical trial level and to date four cationic peptides have reached phase III of clinical-efficacy trials (Table 1.4). Pexiganan, the frog Magainin derivative MSI-78, which has been indicated to cure or prevent

impetigo and diabetic foot ulcers; Isegran, a pig protegrin derivative IB-367 which cures oral mucositis; Neuprex, the human bactericidal permeability protein derivative rBPI<sub>23</sub> for sepsis; and Omiganan variant, the cattle Indolicidin variant CP-226 or MX-226/MBI-226 and CLS001 or MX-594AN, which cures catheter-associated infections and dermatological related infections respectively. Only two peptides have demonstrated efficacy at their phase III clinical trials (Pexiganan and Omiganan), but have not yet been approved by the Food and Drug Administration (FDA) (Hancock and Sahl, 2006; Fjell *et al.*, 2012). Even though most Antimicrobial Peptides are still at phase I or II clinical trials, it will just be a matter of time to properly assess the results of the trials before going to the next step. There is hope that these peptides will help challenge the microbes' resistance to current drugs, and the Antimicrobial Peptides structure and composition can be optimized, for improved therapeutic indices, drug design and good pharmaceutical application (Hancock and Sahl, 2006; Silva *et al.*, 2011; Fjell *et al.*, 2012).



**Table 1.4:** AMPs drugs in development in clinical trials.

Name	Sequence	Company	Description	Application	Trial phase	Comments	Clinical trial identifiers and further information
Pexiganan acetate (MSI 78)	GIGKFLKK AKKFGKAF VKILKK	MacroChem	Synthetic analogue of magainin 2 derived from frog skin	Topical antibiotic	III	No advantage demonstrated over existing therapies	NCT00563433 and NCT00563394
Omiganan (MX-226/ MBI-226)	ILRWPW WPWRRK	Migenix/ BioWest therapeutics	Synthetic cationic peptide derived from indolicidin	Topical antiseptic, prevention of catheter infections	III	Missed primary end point (infections) but achieved secondary end points of microbiologically confirmed infections and catheter colonization	NCT00027248 and NCT00231153
Omiganan (CLS001)	ILRWPW WPWRRK	Cutanea Life Sciences/ Migenix	Synthetic cationic peptide derived from indolicidin	Severe acne and rosacea; anti-inflammatory	II/III	Significant efficacy in Phase II trials for both indications; in Phase III trials	NCT00608959
Iseganan (IB-367)	RGGLCY CRGRFC VCVGR	Ardea Biosciences	Synthetic 17-mer peptide derived from protegrin 1	Oral mucositis in patients undergoing radiation therapy	III	No advantage demonstrated over existing therapies	NCT00022373
XOMA 629	KLFR-(d-naphtho -Ala)- QAK- (d-naphtho -Ala)	Xoma	Derivative of bactericidal permeability-increasing protein	Impetigo	IIa	No Phase IIa results available (trial started in July 2008)	<a href="#">XOMA website</a>
PAC-113	AKRHHG YKRKFH	Pacgen Biopharmaceuticals	Synthetic 12-mer peptide derived from histatin 3 and histatin 5	Oral candidiasis	IIb	Phase IIb results (announced June 2008): 34% increase in primary end point efficacy level; Phase III trial not initiated	NCT00659971
IMX942	KSRIVPA IPVSLL	Inimex	Synthetic cationic peptide derived from IDR1 and batenecin	Nosocomial infection, febrile neutropenia	Ia	Phase Ia trial completed in 2009; no Phase II trial announced yet	<a href="#">Inimex Pharmaceuticals website</a>
OP-145	IGKEFK RIVERIK RFLREL VRPLR	OctoPlus; Leiden University, The Netherlands	Synthetic 24-mer peptide derived from LL-37 for binding to lipopolysaccharides or lipoteichoic acid	Chronic bacterial middle ear infection	II (completed)	Clinical proof-of-efficacy in Phase II trials; no Phase III trials proposed yet	ISRCTN84220089
PMX-30063	Structure not disclosed	PolyMedix	Arylamide oligomer mimetic of a defensin	Acute bacterial skin infections caused by <i>Staphylococcus</i> spp.	II	Mimetic rather than peptide; currently in Phase II trials	NCT01211470; PolyMedix website

Table 1.4 was taken from Fjell *et al.*, 2012

## 1.5. Biosynthesis and diversity of antimicrobial peptides

### 1.5.1. Biosynthesis of antimicrobial peptides

Living creatures on earth have always been exposed and survived environmental challenges (climatic change, drought), famine, and mostly pathogenic infections. It is surprising to know how the egg of an animal can develop without proper defence from white blood cells and manage to overcome the situation of being surrounded by microbes. The reason for surviving can be hypothesized that these organisms have defence systems to fight these pathogens, and that these defences system have contributed to their natural selection in response to the challenges and diseases they encountered. These methods of surviving can be better comprehended if we consider the case of animals such as frogs, monkeys, pig, insects and even plants which external bodies or skin are constantly exposed to the microbial world. As such, various species under the plant and animal kingdoms have produced substances as defence tools which protected them against pathogenic attacks throughout their evolutionary period (Andreu and Rivas, 1999). Hence, plants and animals have an inborn defence system called innate immune system of defence.

Besides having the innate immune system, they also have an acquired immune system to fight against all type of diseases. Yet, the innate immune system is largely responsible to defend the organism in response to microbe invasion. The majority of Antimicrobial Peptides synthesized by these multi-cellular organisms are gene encoded, via a regular process of transcription and ribosomal translation followed sometimes by further proteolytic cleavage of the gene product. However, some of the Antimicrobial Peptides are not ribosomally synthesized and are synthesized from non-protein amino acids such as  $\beta$ -Alanine,  $\gamma$ -aminobutyric acid, orthonine, 2,3-Diamonosuccinic acid or they are modified after translation (Epanand and Vogel, 1999). The diversity of Antimicrobial Peptides is as a result of post-

translational modification of the proteins synthesized from protein amino acids or they are synthesized directly from the non-protein amino acids to fulfil their biological activities and combat the various microorganisms to which they are exposed in their environment (Andreu and Rivas, 1999; Zasloff, 2002).

### **1.5.2. Diversity of antimicrobial peptides**

Despite the widespread distribution of Antimicrobial Peptides throughout the animal and plant kingdoms, some sub-classes of peptides have ancient lineage precursors in common, regardless of their diversity at their active sites. As such, it is somehow seen as if they are not different and diverse except in their amino acid composition and secondary structures (Boman, 1995; Gennaro and Zanetti, 2000; Vizioli and Salzet, 2002). The categorization of Antimicrobial Peptides based on their amino acid composition and structure have enabled their grouping as follows: linear peptides which form amphipathic and hydrophobic helices, cyclic peptides and small proteins which form  $\beta$ -sheet structures, cationic peptides with unique amino acid compositions also called extended peptides, macrocyclic knotted peptides or peptides with loops (Powers and Hancock, 2003). However, their amino acid composition can also enhance the variety of antimicrobial peptide sizes, charges, hydrophobicities and amphipathicities.

Besides deriving their diversity from their fundamental amino acid composition and secondary structure, Antimicrobial Peptide variations mostly arise from the different genes that encode them after which proteolytic cleavage follow, or after post-translational modification of the original peptide sequences. The post-translational modification include proteolytic processing, and in some cases glycosylation (Bulet *et al.*, 1993; Andreu and

Rivas, 1999); carboxyl-terminal amidation and amino acid isomerization (Simmaco *et al.*, 1998) and halogenation (Shinnar *et al.*, 1996).

The diversity can also occur as a result of single mutation on the original peptide sequence thus, changing the biological activity of the peptide. Mutations on the peptide probably reflects the species adaptation to the unique microbial environment that is characterised by the niche they occupy, including the microbes associated with their acceptable food sources (Simmaco *et al.*, 1998; Boman, 2000). Producing peptides in response to the presence of various microbes had encouraged the diversity of Antimicrobial Peptide sequences in such a way that two similar sequences are rarely produced by two different species, regardless of their kingdom, animal or plant, and despite even being very closely related. Exceptions include peptides cleaved from highly conserved proteins such as Buforin II (Zasloff, 2002).

However, similarity and conservation of amino acid residues in the pre-pro-region of the precursor molecules can be found within Antimicrobial Peptides from a single species, and even between certain classes of different peptides from diverse species (Simmaco *et al.*, 1998). Even though the Antimicrobial Peptides are diverse, their structures always adopt a hydrophobic and amphipathic spatial design, to target the negative charged lipopolysaccharide on the microbial membrane.

## **1.6. Biophysical properties of antimicrobial peptides**

Despite the varieties of Antimicrobial Peptides, they exhibit common physicochemical properties, which enable their classification and characterisation. Hence, their biological activities as natural antibiotic molecules and their selective toxicity towards the microbial

target relative to the host. Some of these properties include their positive charge (cationicity), hydrophobicity, amphipathicity and structural conformations.

### 1.6.1. Charge ( $Q$ )

Antimicrobial Peptides are naturally occurring antibiotic molecules, responsible to thwart off potential pathogens from the organism. Many Antimicrobial Peptides are small molecules of variable lengths ranging from 6 to 100 amino acids and are positively charged with +2 to +9 charges. The positive charge is as the result of an excess of basic amino acids in their composition, mostly made up of Lysine, Arginine and Histidine residues (Hancock and Chapple, 1999).

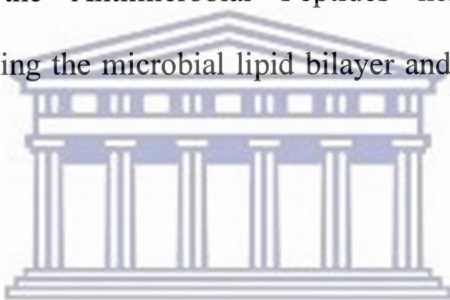
The positive charge allows the electrostatic attraction of microbes, seeing that the microbes' bilayer membrane is made up of negatively charged phospholipid molecules such as lipopolysaccharides, phosphatidylglycerols and cardiolipins (Matsuzaki *et al.*, 1995; Matsuzaki, 2009). Therefore, the microbe becomes the preferential target of the Antimicrobial Peptides due to their selective toxicity property. However, in a specific example such as Magainin, it was shown that a charge of +5 was found to increase the microbial activity of the peptide but an increase of the charge up to +7 altered the microbial activity. The increase in positive charges rather increased the haemolytic activity and led to a decrease of the membrane permeabilizing ability, thus resulting in a likely decrease in hydrophobicity paralleled to the increase in charge (Powers and Hancock, 2003).

Antimicrobial Peptides can also possess a negative charge. For example, Maximins and Lactoferrin enriched by Aspartic and Glutamic amino acids in their structure. The antimicrobial activities of these molecules are enhanced by the incorporation of zinc as a

cofactor. These also have broad activity on both gram-negative and gram-positive bacteria (Brogden *et al.*, 1997).

### 1.6.2. Hydrophobicity (*H*)

A large proportion of hydrophobic residues are found on the peptide molecule, generally up to 30% and more (Hancock and Sahl, 2006; Giuliani *et al.*, 2007). Feature selection algorithms indicate that composition and distribution of charged and hydrophobic residues of peptides are the major determinants of the antimicrobial activity (Thomas *et al.*, 2010). The hydrophobic domains in the Antimicrobial Peptides help in effective membrane permeabilization by partitioning the microbial lipid bilayer and consequently results in lysis (Brogden, 2005).



### 1.6.3. Amphipathicity (*A*)

Amphipathicity is the amount of hydrophobic and hydrophilic domains in a protein molecule. The amphipathicity structures of these peptides have a variety of antimicrobial activities ranging from membrane permeabilization to actions on a range of cytoplasmic targets. The amphipathicity of protein secondary elements can be quantified by the concept of hydrophobic moment ( $M_H$ ).

The hydrophobic moment is the sum-vector of the hydrophobicities of the side chains of the amino-acyl residues of the helix or strands. Other peptides often show spatial separation of polar and hydrophobic residues that is less easy to quantify (Eisenberg *et al.*, 1982). A more technical approach to assess the amphipathic nature of a peptide is to measure the retention time of the peptide in reversed phase high performance liquid chromatography. Using this



technique, findings show that peptides with high amphipathicity exhibited high haemolytic activity but had low antimicrobial activity. However, Antimicrobial Peptides with low amphipathicity exhibited higher antimicrobial activity (Kondejewski *et al.*, 1999). Further studies showed that optimizing peptide incorporation into the bacterial membrane may not always produce the most effective antimicrobial activity of the peptide, but the primary and most essential element for antimicrobial activity is the interaction of the positive charges on the peptide and the negative charges on the bacterial membrane (Oren *et al.*, 1999).

#### **1.6.4. Structure and Conformations of antimicrobial peptides**

The positive charged, amino acids composition of the peptide sequences coupled with their high hydrophobic nature have attributed to various secondary structures of the Antimicrobial Peptides. They fold into a variety of secondary structures (often after they have inserted into the membrane bilayer) through the use of charged, polar and hydrophobic residues forming patches on the surface of the microbe, despite their small size (Hancock and Rozek, 2001). The most common 3-D structures fit into four known classes (Hancock and Lehrer, 1998), the most common classes are the  $\alpha$ -helices and the  $\beta$ -sheets. The less common ones are the extended peptides and loop peptides.

##### **1.6.4.1. The $\alpha$ -helical peptides**

The  $\alpha$ -helical class of Antimicrobial Peptides are characterized by their helical conformation and lack of cysteine amino acid residues, for example human Cathelicidin LL-37, (Figure 1.3A). Cationic amphipathic and hydrophobic  $\alpha$ -helix conformations are favoured by the presence of Alanine, Lysine, Leucine, Phenylalanine, Tyrosine, Tryptophan, Cysteine, Methionine, Asparagine and Valine amino acids. On the other hand the  $\alpha$ -helices are

destabilised by amino acid residues such as Serine, Isoleucine, Threonine, Glutamic acid, Aspartic acid, Glycine, Proline and Hydroproline.

The most studied  $\alpha$ -helical Antimicrobial Peptide is Magainin, mainly Magainin 1 and 2, which were isolated from the skin of the African clawed frog, *Xenopus laevis* (Zhang *et al.*, 1999). Other well studied  $\alpha$ -helical Antimicrobial Peptides include but are not restricted to Cecropins A, Andropin, Moricin, Ceratotoxin and Melittin from insects, Cecropin P1 from *Ascaris nematodes* (Andersson *et al.*, 2003); CAP18 from rabbits; LL-37 from humans; PMAP from cattle, sheep and pigs (Brogden, 2005). Most of these peptides are known to be disordered when in aqueous solutions, but they are immediately converted to an  $\alpha$ -helix in the presence of Trifluoroethanol, Sodium Dodecyl Sulfate (SDS) micelles, phospholipid vesicles and liposome or in the presence of lipid A of the microbes membrane (Brogden, 2005).

#### 1.6.4.2. $\beta$ -sheet peptides and small proteins

This class of peptide conformation is characterized by the presence of an anti-parallel  $\beta$ -sheet or a cyclization structure (Figure 1.3B). The  $\beta$ -sheets peptides are held in place by 2-4 disulfide bridges such as in the case of the Tachyplesins, Protegrins, human  $\alpha$ -defensins, and Lactoferricin peptides. Nevertheless,  $\beta$ -sheeted peptides may occasionally contain a short  $\alpha$ -helical stretch for long peptide sequences, such as in human  $\beta$ -defensin 1 (Figure 1.3C). Cyclization of the peptide backbone may appear in peptides such as Gramicidin or Polymycin B (Epanand and Vogel, 1999).

The nature of the  $\beta$ -sheets structure formation has been reported to be a matter of monomer aggregation, or a bend has to be formed to allow an intramolecular anti-parallel structure. However, the latter conformation is not easily achieved for small peptides because there will be a loss in the entropy, which would not be compensated for by the favourable bonding

interaction that might result. Also, the monomer  $\beta$ -structure with only two interacting peptides segments would still have half the hydrogen-bonding potential of the amide groups unfulfilled (Epan and Vogel, 1999). Thus, Antimicrobial Peptides which form  $\beta$ -structures are able to do so because they are cyclical peptides and hence, there is less entropy loss on the formation of the  $\beta$ -structure.

The importance of the Cysteine amino acids as well as their exact role in  $\beta$ -structured Antimicrobial Peptide action and the mechanism of causing damage to the microbes are not well established. However, a study on the orientation of Protegrin in membranes using circular dichroism, demonstrated that there are two different states of insertion of the peptide into a membrane that depends (a) on the Antimicrobial Peptide concentration, (b) the nature of the lipid bilayer membrane and (c) the extend of hydration (Heller *et al.*, 1998). The cyclic anti-parallel  $\beta$ -sheet structure peptide Tachyplesin, held together by two disulfide bonds, passes across lipid bilayers of bacterial and artificial lipid membranes, coupled with transient pore formation (Matsuzaki *et al.*, 1997b). On the other hand, the importance of the Cysteine amino acids on the  $\beta$ -sheet peptide is required for the formation of disulfide bonds, which helps to maintain the structural integrity of the Antimicrobial Peptide. Studies involving linear Tachyplesin, chemically protected with acetamidomethyl groups, (T-Acm) demonstrated reduced antimicrobial and antiviral activity of the compound (Tamamura *et al.*, 1993) as well as a reduction in calcein release from model membranes (Matsuzaki *et al.*, 1993). Though being less effective at permeabilization of model membranes, T-Acm however possessed greater disrupting ability as assayed by measuring lipid chain orientation (Matsuzaki *et al.*, 1993).

Additionally, the linear analogue completely lacks the ability of the parent peptide to translocate across lipid membranes. These studies showed that the stabilization force of disulfide bonds of Tachyplesin is not absolutely required for antimicrobial activity, but it is

rather necessary for the folding and structural rigidity of the peptide, and permits membrane translocation in model systems (Matsuzaki *et al.*, 1997b).

#### **1.6.4.3. Peptides with irregular amino acid composition or extended peptides.**

The class of extended peptides is characterized by the presence of unusual amino acid composition in the Antimicrobial Peptide sequence, which is rich in one or more amino acids of the same residue. Hence, the inability of the extended structure of these peptides to possess a classical secondary structure (Figure 1.3E), which is generally due to the presence of high Histidine, Tryptophan, Proline and/or Glycine residues.

In fact, these peptides form their final secondary structure not through inter-residue hydrogen bonds, but by hydrogen bonds and Van der Waals interactions with the membrane's lipids (Powers and Hancock, 2003). Well known characterized Antimicrobial Peptides of extended structure include: Histatin, Indolicidin, PR-09, Prophenin and Tritrpticin. Histatin is an Antimicrobial Peptide produced by saliva, and is extremely rich in histidine residues (Gallo *et al.*, 1994; Futaki *et al.*, 2001; Gao *et al.*, 2001). PR-09 and Prophenin, other Antimicrobial Peptides of this class are produced by porcine neutrophils, which are very rich in Proline and Arginine or Proline and Phenylalanine whereas Indolicidin and Tritrpticin are richer in Tryptophan amino acid residues. Nevertheless, Indolicidin produced from the cytoplasmic granules of bovine neutrophils, is a 13 residue C-terminal peptide and has 5 Tryptophan amino acids, in fact, making Indolicidin the peptide with the highest known number of Tryptophan residues (Selsted *et al.*, 1992).

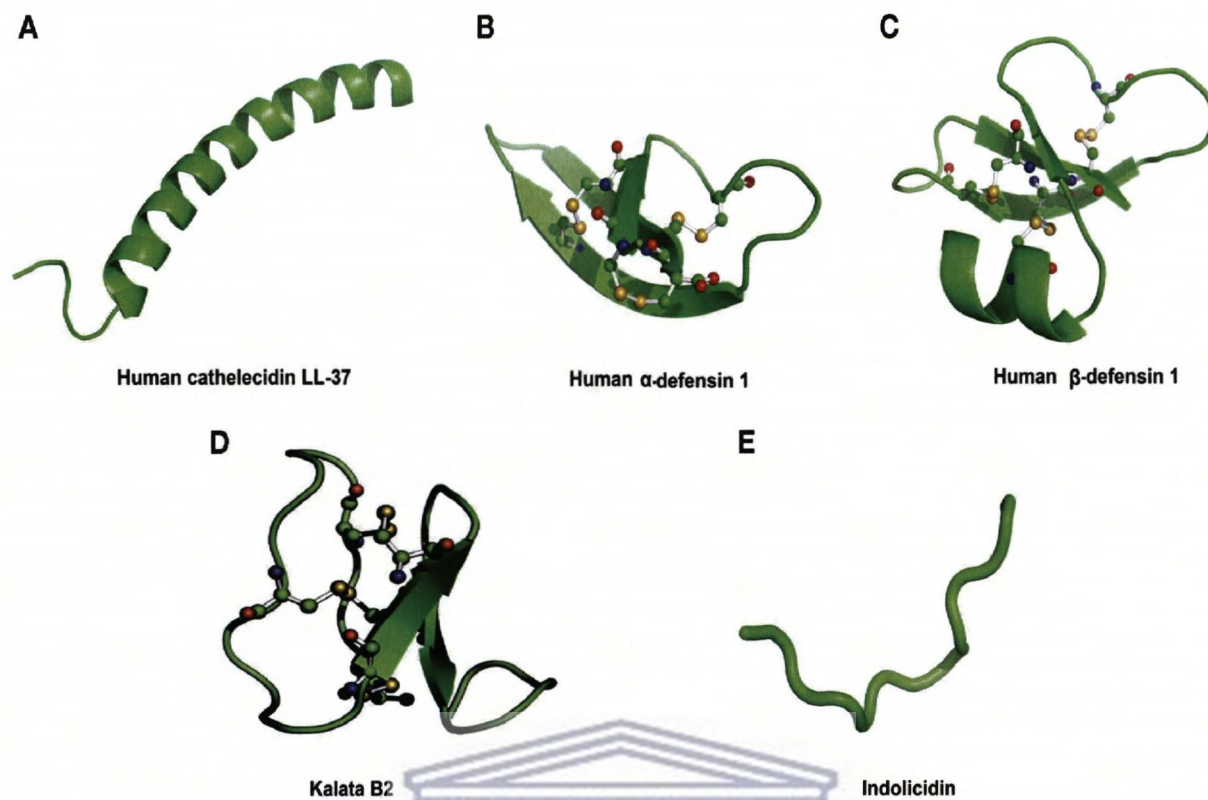
Although the secondary structure has an unusual conformation, the conformation of the Antimicrobial Peptide is a major parameter for Antimicrobial Peptides to translocate across the lipid membrane (Rozek *et al.*, 2000; Friedrich *et al.*, 2001; Zhang *et al.*, 2001).

#### 1.6.4.4. Peptides with loops or macrocyclic cystine knot peptides

Many Antimicrobial Peptides have been reported to have a cyclic peptide structure and this cyclic feature involves not less than 15 amino acids. The loop structure, which forms the main characteristic of this class of secondary structures, is imparted by the presence of a single bond, either disulfide or amide and iso-peptide bonds (Figure 1.3D) (Cornut *et al.*, 1994; Hancock and Diamond, 2000; Powers and Hancock, 2003). The single disulfide bond is the most important feature of all, because it enhances the formation of a cystine knotted motif in the peptide and the cyclic backbone; and confers a high rigidity to the structure.

Recently, another well characterized AMP having this structure are the four macrocyclic end-to-end 30 amino acid residues cyclic peptides from plants of the *Rubiaceae* family: Kalata, Circulin A and B, Cyclopsychotride (Tam *et al.*, 1999). Circulin A and B have been reported to have antiviral activity (Gustafson *et al.*, 1994), with particular activity on human immunodeficiency virus (HIV) inhibition, and can thus play a pivotal role in the design of novel anti-HIV drugs (Gustafson *et al.*, 1994). Other well characterized Antimicrobial Peptides having this structure is Thanatin, a 21-residue peptide, isolated from the spined soldier bug, *Podisus maculiventris* (Fehlbaum *et al.*, 1996).

Besides the four 3-D structures cited above, the Antimicrobial Peptides can bear other 3-D structures such as peptaibols, characterized by a high proportion of  $\alpha$ -amino-isobutyric acid (Aib) residues, isovaleric acid (Iva) and the amino acid Hydroxyproline (Hyp) (Monaco *et al.*, 1998; Wiest *et al.*, 2002). Other classes include peptides with thio-ether rings, characterized by their possession of small ring structures enclosed by a thio-ether bond. This group is referred to as Lantibiotics (Stahl, 1994; Montville and Chen, 1998). An example of Lantibiotics is Nisin, an antimicrobial agent used for food preservation (de Kruijff *et al.*, 1999).



**Figure 1.3:** An overview of the major structural classes of host-defense peptides including: (A)  $\alpha$ -helices, (B)  $\beta$ -sheets, (C) a mixture of  $\alpha$ -helices/ $\beta$ -sheets structures, (D) cyclic, and (E) extended structures. Disulfide bonds are represented in ball and stick (Silva *et al.*, 2011).

### 1.7. Antimicrobial peptide mechanism of action

Despite their vast variety, the only defining characteristic of Antimicrobial Peptide targets is their possession of a membrane. Therefore, most Antimicrobial Peptides work directly against microbes through a mechanism involving membrane disruption and pore formation, allowing efflux of essential ions and channel formation.

Although the exact mechanisms by which Antimicrobial Peptides exert their killing actions is not completely understood, a common hallmark seems to be their interaction with the negative phospholipid component of the cytoplasmic membrane of the microbe, leading to membrane permeabilization, cell lysis and death (Rinaldi, 2002; Zasloff, 2002). Thus, the first step in the mechanism of action is the electrostatic interaction between the cationic peptide and negatively charged or anionic lipopolysaccharides on the outer membrane of the

pathogen hence, their selective toxicity for the microbial target relative to the host. After the Antimicrobial Peptide binding to the microorganism, the mechanism of action can be divided into two types: (I) the membrane disruptive and (II) the non-membrane disruptive mechanisms. The Barrel-stave, Toroidal pore or Wormhole and the Carpet mechanism are the proposed mechanisms for Antimicrobial Peptides acting via the membrane disruptive mechanism or permeabilization.

### 1.7.1. The Barrel-stave mechanism

This mechanism is proposed for a selected group of peptides. The Barrel-stave mechanism involves the perpendicular insertion, and the aggregation of a relatively small number of individual peptides, also referred to as *staves* in a *barrel-like* ring inside the membrane leading to a membrane pore or channel with a cylindrical structure (Giuliani *et al.*, 2007). The hydrophobic surfaces of  $\alpha$ -helical or  $\beta$ -sheet peptides face outward, toward the acyl chains of the membrane, whereas the hydrophilic surfaces form the pore lining (Breukin and Kruijff, 1999). When bound peptide reaches a threshold concentration, peptide monomers self-aggregate and insert deeper into the hydrophobic membrane core. As a result of aggregation and recruitment of peptide monomers, a pore structure will be formed on the microbes' membrane and will increase as peptide monomers are added (Figure 1.4A). A minimal length of about 22 amino acids is required to transverse the lipid bilayer with  $\alpha$ -helical peptides, whereas about 8 amino acids are required for that of a  $\beta$ -sheeted structure (Lehrer and Ganz, 2002). An example of an Antimicrobial Peptide that inserts and aggregates by using the Barrel-stave mechanism is Alamethicin (Alm) (Yang *et al.*, 2001).

### 1.7.2. The Toroidal pore or the Wormhole mechanism

Also known as the Wormhole mechanism, it is a well characterised mode of antimicrobial action in peptide-membrane interaction. In addition to being similar to the Barrel-stave mechanism of antimicrobial action, the Toroidal pore mechanism shows some differences in the process of penetrating the outer membrane of the microbe. In this mode of action, the Antimicrobial Peptide helices penetrate into the membrane and induce the lipid monolayer to bend continuously through the pore. Consequently, the membrane curves inward to form a hole, with the hydrophobic head groups facing the centre whilst the hydrophilic surfaces of the peptide lines the hole (Figure 1.4B). Examples of Antimicrobial Peptides that employ this mechanism of action are Magainins, Melittin and Protegrins (Yang *et al.*, 2001; Brogden, 2005).

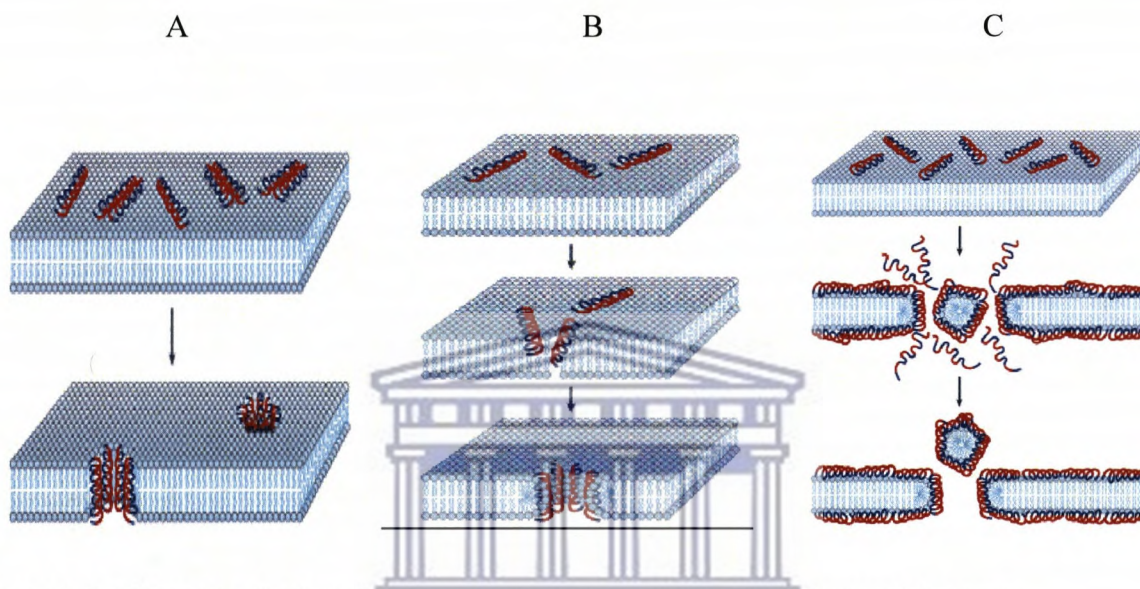


### 1.7.3. The Carpet mechanism

The Carpet mechanism acts against microorganisms through a relative diffuse manner. As in other models, peptides do not insert into the membrane but rather bind initially to the membrane mainly via electrostatic interactions, from which peptides accumulate on the surface carpeting the phospholipid bilayer. When a threshold peptide density is reached, the microbial cell membrane is subjected to a significant curvature strain, leading to disruption (Figure 1.4C). From this perspective, membrane dissolution occurs in a dispersion-like manner that does not involve channel formation, and the peptides do not necessarily insert into the hydrophobic membrane core (Giuliani *et al.*, 2007). An example of Antimicrobial Peptides that use this model is Dermaseptin, Cecropin and Melittin (Brogden, 2005).



Following the interaction of the cationic peptide and the anionic membrane of the microbe and the insertion of the peptide into the microbial membrane, using the various mechanisms cited above, insertion is subsequently followed by the disruption of the physical integrity of the membrane or inhibition of functional biomolecules found within the cytoplasm of the microbe.



**Figure 1.4:** Different mechanisms of action used by Antimicrobial Peptides to enter their targets.

(A) **The Barrel-stave model of Antimicrobial Peptide induced killing.** In this model, the attached peptides aggregate and insert into the membrane bilayer so that the hydrophobic peptide regions align with the lipid core region and the hydrophilic peptide regions form the interior region of the pore. Hydrophilic regions of the peptide are shown coloured in red, hydrophobic regions of the peptide are shown coloured in blue (Brogden, 2005).

(B) **The Toroidal model of Antimicrobial Peptide induced killing.** In this model the attached peptides aggregate and induce the lipid monolayer to bend continuously through the pore so that the water core is lined by both the inserted peptides and the lipid head groups. Hydrophilic regions of the peptide are shown coloured in red, hydrophobic regions of the peptide are shown coloured in blue (Brogden, 2005).

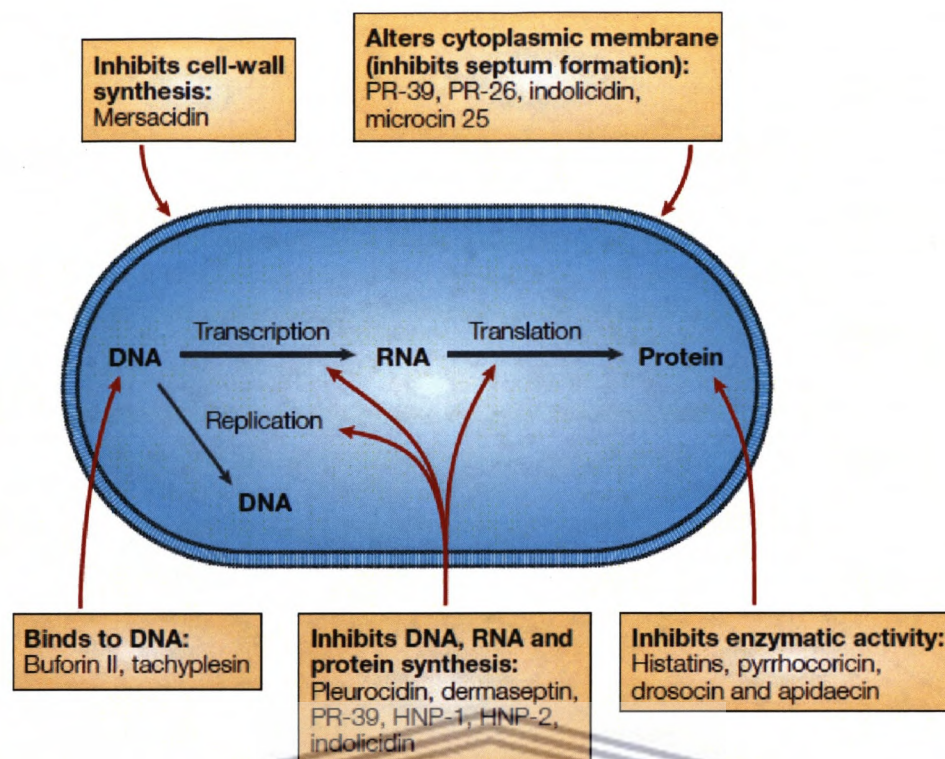
(C) **The Carpet model of Antimicrobial Peptide induced killing.** In this model, the peptides disrupt the membrane by orienting parallel to the surface of the lipid bilayer and forming an extensive layer or carpet. Hydrophilic regions of the peptide are shown coloured in red, hydrophobic regions of the peptide are shown coloured in blue (Brogden, 2005).

There has been evidence that Antimicrobial Peptides use other modes of action besides the Barrel-stave model, the Toroidal model and the Carpet model in the disruptive mechanism of their action (Gennaro and Zanetti, 2000). The alternative mechanism of Antimicrobial Peptide activity is the non-disruptive membrane mechanism. Here, the peptide does not act on the

pathogen's membrane but rather involves a direct diffusion and permeabilization of the peptide molecule into the microorganism, to target the cytoplasmic components, inhibiting nucleic-acid synthesis, protein synthesis or enzymatic reactions taking place within the cytoplasm of the microbe (Brogden, 2005). This mode of action is mostly encouraged by arginine-rich and proline-rich peptides. Example: PR-39, which is capable of crossing the cellular and nuclear membrane of microbes (Futaki *et al.*, 2001). Upon entering the microbe cytoplasm, the immediate effect of the cationic Antimicrobial Peptide PR-39 will be the inhibition of cell wall biosynthesis and DNA replication (Boman *et al.*, 1993). Furthermore, PR-39 and its N-terminal 1-26 fragment, PR-26 have been implicated to induce filamentation of *Salmonella enteric* serovar Typhimurium (*S. typhimurium*) (Shi *et al.*, 1996).

HNP-1 and HNP-2 inhibits the synthesis of DNA, RNA and enzymatic activities such as cellular protein synthesis in *E. coli* (Lehrer *et al.*, 1989); while Indolicin will reduce the replication of *E. coli* DNA and RNA, it might also affect protein synthesis at a higher dose of peptide concentration (Subbalakshmi and Sitaram, 1998), and similarly the action of Buforin II on *E. coli* (Park *et al.*, 1998). Indilocin and Defensins prevent the uptake of (<sup>3</sup>H)thymidine, (<sup>3</sup>H)uridine, and (<sup>3</sup>H)leucine by *E. coli*, hence inhibiting the synthesis of DNA, RNA and proteins (Lehrer *et al.*, 1989; Subbalakshmi and Sitaram, 1998).

Specific enzymatic activities of microbes are also targeted by Antimicrobial Peptides. Examples of Antimicrobial Peptides which inhibit enzymatic activities of microbes include the proline-rich Pyrrhocoricin, Drosocin and Apidaecin. These Antimicrobial Peptides have been shown to bind to the heat shock protein DnaK of the chaperone-assisted protein folding class, hence interfering and inhibiting the folding and the activity of this protein (Otvos *et al.*, 2000; Kragol *et al.*, 2001). Additionally, Lantibiotics and Mersacidin Antimicrobial Peptides from *Bacillus* have been shown to bind to lipid II, leading to the inhibition of peptidoglycan biosynthesis (Brotz *et al.*, 1998).



**Figure 1.5:** Alternative modes of action for intracellular Antimicrobial Peptide activity. In this figure, *Escherichia coli* are shown as the target microorganism. Figure was taken from Brogden, 2005 and modified.

## 1.8. *In Silico* discovery of antimicrobial peptides

The basic technique for Antimicrobial Peptides discovery and identification has always been through the use of molecular approaches. However, with the creation of a new area such as Bioinformatics and computational biology, *in silico* discovery and identification of novel/putative Antimicrobial Peptides have been made easier. This is due to the fact that the method is less time consuming, cost effective and less labour intense.

### 1.8.1. Classification of antimicrobial peptides: establishment of databases

The growing number of new Antimicrobial Peptides and the increase in the number of synthetic Antimicrobial Peptides has encouraged the creation of repository databases to store, classify and organise these peptides. The creations of computational tools have also been

incorporated into these databases, to better and faster analyse and classify the activity and properties of the Antimicrobial Peptides.

Since the discovery of the first Antimicrobial Peptide 30 years ago on the skin of the European frog *Bombina variegata* (Rinaldi, 2002); thousands of Antimicrobial Peptides have been found and experimentally validated. Examples include but not restricted to Attacin, Drosocin, Andropin, Cecropin, Maximins, Lactoferin, Hemocymmin, Ascalin, Dermaseptin, Ginkbilobin, Cycoviolin, Cycloviolacin, Kalata, Circulin, Palicourein, Tricyclon, Vh11, Aurein, Tachyplesin, Melittin, Magainin, Siamycin I and II, Indilicidin, Ranatuerin, Sesquin, Gymnin, Cicadin etc.

The multitude of Antimicrobial Peptides is due to different sub-classes of Antimicrobial Peptides within a species. As a result, Antimicrobial Peptides will have many multi-isoforms of peptide sequence in the species, hence their diversity. Nevertheless, similarity is often found only within defined groups of defence peptides from closely related species and they have the same Antimicrobial Peptides on a defined microorganism (White *et al.*, 1995; Wachinger *et al.*, 1998; Daly *et al.*, 2006). However, Antimicrobial Peptides originating from the same species might have different antimicrobial activities (Hill *et al.*, 1991; Lai *et al.*, 2001; Navon-Venezia *et al.*, 2002).

The need to classify these Antimicrobial Peptides is due to the fact that many of these Antimicrobial Peptides were uncurated and displaced. Hence, repository databases were created to store these peptides, which are experimentally validated, i.e. their antimicrobial activity proven with the appropriate assay. These databases include: AMSdb (Tossi and Sandri, 2002), AMPer (Fjell *et al.*, 2007), Antimic (Brahmachary *et al.*, 2006; Brahmachary *et al.*, 2004), APD (Wang and Wang, 2004; Wang *et al.*, 2009), Bagel (de Jong *et al.*, 2006; de Jong *et al.*, 2010), CAMP (Thomas *et al.*, 2009), Cybase (Mulvenna *et al.*, 2006; Wang *et*

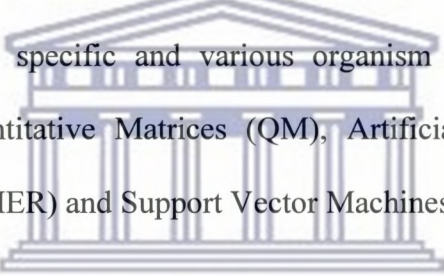
*al.*, 2008), Defensins knowledgebase (Seebah *et al.*, 2006), PenBase (Gueguen *et al.*, 2006), Peptaibols (Whitmore and Wallace, 2004; Chugh and Wallace, 2001), PhytAMP (Hammami, 2009), RAPD (Li and Chen, 2008), SAPD (Wade and Englund, 2002), Bactibase (Hammami *et al.*, 2007) and DAMP (Sundararajan *et al.*, 2011).

Each database has its own particularity and specificity, advantages and limitations; but most of these databases have analytic tools incorporated into their servers, to aid in Antimicrobial Peptide prediction. In that regard, AMSdb (Tossi and Sandri, 2002), PhytAMP (Hammami, 2009), Bactibase (Hammami *et al.*, 2007) and PenBase (Gueguen *et al.*, 2006) are databases of Antimicrobial Peptides from eukaryotes, plants, bacteria and shrimps, respectively and are based on specific classes of Antimicrobial Peptides. Other databases such as RAPD (Li and Chen, 2008) deals with recombinant Antimicrobial Peptides, SAPD (Wade and Englund, 2002) includes information on synthetic Antimicrobial Peptides. Defensin knowledge base deals with defensins (Seebah *et al.*, 2006), Peptaibol Database is specific to peptaibols (Chugh and Wallace, 2001; Whitmore and Wallace, 2004).

While these databases are only dealing with specific classes of Antimicrobial Peptide, other databases have tools for Antimicrobial Peptide analysis and further detailed information on various Antimicrobial Peptide classes. Examples of such databases include APD2 (Wang and Wang, 2004; Wang *et al.*, 2009), AMPer (Fjell *et al.*, 2007) and CAMP (Thomas *et al.*, 2009). APD and CAMP have information on the taxonomy and the activity of the peptide, experimentally validated peptides, patents and predicted Antimicrobial Peptides. Both databases also allow for search and query of the sequences found within the datasets, using keywords such as antiviral peptides, antifungal peptides, antibacterial peptides, anti-HIV-1 peptides etc. However, predicted peptides are of little use since their activity on the target organism is not established.

The prediction of putative Antimicrobial Peptides is based on their sequence similarity with known Antimicrobial Peptides, as well as activities based on properties of known peptides. Prediction servers such as APD (Wang and Wang, 2004; Wang *et al.*, 2009), Bactibase (Hammami *et al.*, 2007) and PhytAMP (Hammami, 2009) are capable of sequence alignment of input Antimicrobial Peptide sequences to identify similar peptides in these databases.

APD database have tools which are able to predict if a peptide is an AMP or a non-AMP. This prediction is based on the physicochemical properties of the said peptide. Bactibase and PhytAMP possess tools such as HMMER-based models for the prediction of the Antimicrobial Peptide families of the query sequence/s; conversely, the mathematical algorithm tools have the ability to cluster and classify, train, test and predict the activity of Antimicrobial Peptides from specific and various organism sources. Examples of such sophisticated tools are Quantitative Matrices (QM), Artificial Neural Network (ANN), Hidden Markov Model (HMMER) and Support Vector Machines (SVM).



**1.8.2. Computational tools used in model construction and antimicrobial peptide identification**

Besides being a promising approach because it is less time consuming, cost effective and less labour intense, the major breakthrough of computational biology and bioinformatics in the event of Antimicrobial Peptide discovery, has been its capability to cluster Antimicrobial Peptides into curated databases, construction of new tools for stochastic optimization of Antimicrobial Peptides and to search for putative Antimicrobial Peptides for therapeutic purposes (Wang *et al.*, 2009; Sundararajan *et al.*, 2011).

### 1.8.2.1. Computational tools used for antimicrobial peptide model construction

Various computational tools are often used in the *in silico* and *de novo* discovery of putative Antimicrobial Peptides, using experimentally validated peptides as a training set. Though these tools are included into some databases, they can also be constructed as a separate entity for antimicrobial discovery. The most frequent tools used for model construction for the training, testing and prediction of peptides with antimicrobial activity, are but not limited to, Hidden Markov Models (HMMER) (Brahmachary *et al.*, 2004; Fjell *et al.*, 2007), Support Vector Machine (SVM) (Thomas *et al.*, 2010), Quantitative Structure Activity Relationship (QSAR) (Fjell *et al.*, 2009), Linear Discriminant Analysis (LD) (Thomas *et al.*, 2010), Random Forest (RF) (Thomas *et al.*, 2010), Gap Local Alignment of Motifs 2 (GLAM2) (Frith *et al.*, 2008) and Sliding Window (SW) (Torrent *et al.*, 2012).

Despite the fact that most of these algorithms make use of data mining and high-throughput screening techniques and apply vector-like analysis to scan protein and peptide sequences, they also take advantage of the available bioinformatic machine learning techniques such as SVM, DA, RF and ANN (Torrent *et al.*, 2009). These computer-based design algorithms enable easy and consistent assessment of the complex Antimicrobial Peptide data. Whilst the main reason for using computational methods and associated tools is to enable the analysis of the large scale of experimental data, these methods make use of three common approaches for the discovery of Antimicrobial Peptides:

- (i) Modify a known Antimicrobial Peptide sequence (known as template) with limited computational input
- (ii) Rigorous biophysical modelling to understand peptide activity and
- (iii) Virtual screening.

The first approach does not necessitate high-throughput computational demand but instead improves the activity of a template peptide by a single mutation modification at a particular

point of the peptide backbone (Robinson, 2011). The modification is done on the basis of the importance of charge and amphiphilicity, i.e. physicochemical properties for activity alone (Pag *et al.*, 2009). The choice of only these three properties might explain the limitation of the approach, however the approach has shown the importance of certain amino acids and their positions on the peptide structure and how it affects the peptide's activity.

While the second approach makes use of computational modelling such as Support Vector Machine (SVM) (Thomas *et al.*, 2010), profile Hidden Markov Models (HMMER) (Brahmachary *et al.*, 2004; Fjell *et al.*, 2007), Gap Local Alignment of Motifs 2 (GLAM2) (Frith *et al.*, 2008), Quantitative Structure Activity Relationship (QSAR) (Fjell *et al.*, 2009), Linear Discriminant Analysis (LD) (Thomas *et al.*, 2010), Random Forest (RF) (Thomas *et al.*, 2010) and Sliding Window (SW) (Torrent *et al.*, 2012) to search for putative Antimicrobial Peptides by understanding AMP activity, the third approach however does not require model creation but rather an impute peptide structure of the primary sequences or a model created by one of the second approaches. Nonetheless, both approaches require the usage of physicochemical properties to achieve the end goal of design. The properties take into account the primary structure of the peptides. Some of the biophysical properties include thermodynamic calculations of the interactions of peptides with the microbial membranes, molecular modelling based on free energy perturbation, as well as molecular dynamics simulation (Mátyus *et al.*, 2007). Other properties might include hydrophobicity, charge of the side chains, amphiphilicity, hydropathic index, and several others. The properties, so-called descriptors, are incorporated into computational design methods by features selection tools and ought to be quantified to attain a high level of optimization.



### **1.8.2.2. Computational tools used for antimicrobial peptide model construction and motifs finding**

With the development of computational biology, mathematical algorithms and stochastic tools, some added tools such as those mentioned in **1.8.2.1** were created and have been incorporated into databases as predictions tools. However, models can also be constructed independently, in conjunction with the incorporation of properties related to the structure and the activity of existing peptides using the same tools.

SVM, QSAR, LD, SW and RF require structure-activity relationship information in order to enhance their strength and performances. Furthermore, the calculation of structure-activity relationship suffers from one drawback. That is, it uses the index of physicochemical properties, which sums each property of the sequence (Torrent *et al.*, 2012). The HMMER and GLAM2 algorithms only require the usage of naturally occurring, experimentally validated peptide sequences to construct the models, which will displace their features based on the motifs of the sequences.

The HMMER and the GLAM2 algorithms work on the principle of motif profile description generated by sequences of the same family for their profile construction, called a training data set. The main novelty of these tools is that they allow insertions and deletions in motifs. The motifs found within the protein/peptide sequences, forms the patterns or features that will be exhibited by the profiles.

To achieve good model construction for both algorithms, proteins/peptides sequences having the same function are divided into a training set and a testing set. The training set accounts for three quarters of the total initial sequences while the testing sets accounts for one quarter of the total sequence number. The steps for the models are alignment, model building and querying of the model. However, an additional calibration step is added into the HMMER algorithm, making this tool more sensitive for database searching. Also, the two tools differ

in their statistical scoring and probability prediction annotation. The models or profiles generated by the HMMER and the GLAM2 algorithm can serve as templates to search for sequences of the same signature or the same sequences of selected families in databases.

### **1.8.3. Prediction of antimicrobial peptides three-dimensional structures**

Molecular interaction between the Antimicrobial Peptide and the microbial proteins are required to take place for the biological activity of the Antimicrobial Peptide to be effective and prevent the continuation of the disease pathology. For these protein molecules to interact, they ought to be in their right conformation with the correct three-dimensional (3-D) structure and within range of the target.

While most traditional methods used to solve the 3-D structure of proteins have been the usage of tools such as X-ray crystallography, high-resolution electron microscopy (EM) and nuclear magnetic resonance (NMR), they have proven to have some limitations such as: being time consuming, expensive and not constantly appropriate. However, these methods have managed to validate around 50000 experimentally derived protein structures and these proteins have been released into the Protein Data Bank (PDB). Additional 3500 proteins have been deposited but are waiting for release into PDB following curation (Berman *et al.*, 2000). Even after solving these high number of protein structures, more than 3.3 million sequences still remain for the elucidation of their 3-D structures. As a result, it has been noted that there are more protein sequences without experimentally proven structures, and to date, the experimental methods have only managed to solve the structures of 84757 protein molecules (<http://www.rcsb.org/pdb/statistics/holdings.do>).

Despite the efforts to solve these 3-D structures by traditional experimental methods, limitations encountered, and the initiative to reduce this gap between the available protein sequences and their solved 3-D structures, computational biology has been a method of

choice for predicting the 3-D structure of most proteins. Although novel, computational biology enjoys high interest and application in many research fields (Peitsch, 1997; Peitsch *et al.*, 2000; Baker and Sali, 2001).

Besides the fact that predicting the 3-D structures of most proteins from their primary amino acid sequences, which can be considered one of the most difficult and challenging aspect in computational biology and biochemistry, four types of approaches are commonly used. They include: (1) the “fold recognition and threading” methods, (2) the “integrative” or “hybrid” methods, (3) the “comparative” or “homology” modelling approach and (4) the “*de novo*” or “*ab initio*” methods (Schwede *et al.*, 2008).

The first method is employed to model proteins with low or statistically insignificant sequence similarity to a known protein structure. The “integrative” or “hybrid” method applies the combination of information from experimental and different computational sources to predict a protein structure, including the three abovementioned methods (1), (2) and (3). The third method uses experimentally solved structures of the related protein ancestor member as a template protein to derive the structure of the protein of interest. This can only be feasible once a detectable template of known structure is available. The fourth method on the other hand predicts the protein structure by making use of the primary amino acid sequence of that protein, by applying the principles of physics that govern protein folding and information derived from a known structure but do not employ a known ancestor protein to recognize folds (Schwede *et al.*, 2008). Although the *de novo* and the comparative methods are different; both methods make use of experimentally validated structures to assist them in the prediction of the unknown protein structure. Furthermore, they also have similarity steps such as identification of modelling templates and sequence alignment, generating all-atom models, model refinement using the Monte Carlo search strategy and a

model evaluation step, using scoring function in respect to the closest native protein structure molecule (Das *et al.*, 2007).

Since the comparative method for protein structure modelling can only produce a highly accurate model of an unknown protein sequence if known template information is available on the homologous proteins, this method will be imperfect if used to predict the structure of proteins and/or peptides molecules which cannot be aligned with the template sequences due to the presence of long variable loop regions or a novel fold which has not been studied before (Dill *et al.*, 2004). As such, the *de novo* method becomes the method of choice to predict with high accuracy the 3-D structure of most proteins/peptides and has made tremendous progress over the years. Examples of this method include the most frequently used Rosetta method (Rohl *et al.*, 2004) and the I-TASSER method (Wu *et al.*, 2007; Zhang, 2008; Roy *et al.*, 2010).

#### **1.8.4. Prediction of protein-protein interaction using docking tools**

Only a few complexes of protein-protein interaction have been solved in structural biology despite the fact that enough progress has been made in the recent years to build the 3-D structures of many proteins and/or peptides (Berman *et al.*, 2000). While it is imperative to solve the 3-D structure of proteins, the same is also true for the construction of the complexes that are formed between the two proteins interacting. Since most molecules ought to be studied in order to know their function (s) and/or activities, it is essential to find a more reliable method to resolve this problem. As such, computational methodologies have become a vital component in predicting the interaction of proteins in the docking process. These approaches are most important as both the proteins conformation and its orientation, are useful in designing experiments for site-directed mutagenesis, predicting ligand binding sites during docking of small molecules in structure-based drug discovery programmes, studying

the effect of mutations and Single Nucleotide Polymorphisms (SNPs) as well as protein engineering and design (Hillisch *et al.*, 2004; Poole and Ranganathan, 2006; Feyfant *et al.*, 2007).

Since computational methodologies have been the pioneer of docking small molecules to protein binding sites in the early 1980s, many docking software and servers have seen the light of day. Some of these servers include but not limited to PatchDock and SymmDock (Schneidman-Duhovny *et al.*, 2005), GRAMM-X (Tovchigrechko and Vakser, 2006), RosettaDock (Lyskov and Gray, 2008), PepSite (Petsalaki *et al.*, 2009), HexServer (Macindoe *et al.*, 2010), Haddock (Dominguez *et al.*, 2003), ClusPro (Comeau *et al.*, 2004) and ZDOCK (Chen *et al.*, 2003).

Although these docking tools have their applications in the abovementioned fields, the protein's interaction will only make sense if additional information is provided on how good the binding affinity is during the virtual screening of the macromolecule targets and their complementarity to the binding sites in drug discovery and other applications. This is regardless of the robustness and accuracy of the available algorithms. As such, a scoring system ought to be added to these tools to make the prediction valuable. However, the scoring continues to be the main challenge of protein-protein interaction prediction (Kitchen *et al.*, 2004). From the above mentioned only PatchDock and SymmDock (Schneidman-Duhovny *et al.*, 2005) and PepSite (Petsalaki *et al.*, 2009) have a well integrated scoring system that are able to rank the best predicted protein bindings.

Whilst many algorithms suppose that the protein molecules are rigid despite the fact that they are flexible, they use the geometric hashing (Bachar *et al.*, 1993) or the Fast Fourier Transformation (FFT) correlation techniques (Katchalski-Katzir *et al.*, 1992) to predict the binding affinities of small molecules bound to their targets. PatchDock, RosettaDock and Haddock servers all use geometric hashing (method used for protein's recognition in 2-D and

3-D based on structural alignment). Whilst PatchDock (Schneidman-Duhovny *et al.*, 2005) is rapid, RosettaDock and Haddock are more computational demanding and incorporate some flexibility (Dominguez *et al.*, 2003; Lyskov and Gray, 2008). Several FFT-based docking programs include web servers such as GRAMM-X (Tovchigrechko and Vakser, 2006), ClusPro (Comeau *et al.*, 2004), ZDOCK (Chen *et al.*, 2003) and HexServer (Macindoe *et al.*, 2010). Similar to the geometric hashing, the FFT-based approaches suppose that the molecules are rigid during the docking. However, the proteins and/or peptides ought to be compact for the molecules to accomplish all possible rigid-body orientations in the 6-D search space.

Because the FFT-based approach makes use of the 3-D Cartesian grid representations of the proteins and can only compute translational correlations, the programs have to repeat several rotational cycles in order to search for the 6-D space; consequently the Cartesian grid-based FFT dock programs becomes more computationally demanding. In order to fill this gap, Hexserver (Macindoe *et al.*, 2010) has been put in place and makes use of the Spherical Polar Fourier (SPF) approach which uses rotational correlations, with less time spend on searching during the docking. However, this robust way of searching the 6-D transformation space may reduce the efficiency of the prediction and the identification of the complementary binding sites during protein-protein or ligand interactions (Schneidman-Duhovny *et al.*, 2005).

### **1.9. Experimental approaches to determine antimicrobial peptide activity**

Antimicrobial Peptides are abundant and a diverse group of biomolecules produced by a variety of organisms; invertebrates, plants and animals species. They possess therapeutic potential against a wide range of microbes, including gram-negative and gram-positive bacteria, fungi, viruses and protozoa. A variety of techniques have been used to assess the

activities of Antimicrobial Peptides. These activities can be assayed *in vitro* and *in vivo* using experimental approaches or predicted with *in silico* approaches.

The experimental approaches include microscopy, fluorescent dyes, ion channel formation, circular dichroism, dual polarization interferometry, solid-state NMR spectroscopy and neutron diffraction. Each technique provides a slightly different view of peptide activity. Microscopy is used to visualize the effects of Antimicrobial Peptides on microbial cells and helps to identify general target sites. Fluorescent dyes enable the measurement of Antimicrobial Peptides to permeabilize membrane vesicles by the release of internal fluorescent-labelled dextran, immunoglobulin, calcein or other probes. Ion channel formation is another useful technique for assessing the formation and stability of an antimicrobial-induced pore. The orientation and secondary structure of an Antimicrobial Peptide bound to a lipid bilayer can be measured by circular dichroism in a controlled humid environment with light incident normal to the sample surface (Wu *et al.*, 1990). Dual polarization interferometry measures the different mechanisms of Antimicrobial Peptides. Solid-state NMR spectroscopy measures the secondary structure, orientation and penetration of Antimicrobial Peptides into the lipid bilayer in the biologically relevant liquid crystalline state. The neutron diffraction technique measures the diffraction patterns of peptide-induced pores within membranes in orientated multilayers or liquids (Brogden, 2005).

### **1.10. Rationale of the study**

The rationale of the thesis arises as result of the following gaps that were found in the literature, which is,

- Acquired Immunodeficiency Syndrome (AIDS) is a disease that affects the human immune system and is caused by the Human Immunodeficiency Virus (HIV), and it

accounts for many deaths worldwide each year, for which a cure is not yet available (United Nations Programme on HIV/AIDS, 2010). Current treatment for HIV infection consists of high active antiretroviral therapy (HAART) (Department of Health and Human Services, 2005). HAART treatment although being promising for the management of the disease, several side effects and medication intolerance is associated with the treatment regime. They include lipodystrophy syndrome, dyslipidaemia and diabetes mellitus; diarrhoea and an increased risk of cardiovascular disease (Montessori *et al.*, 2004; Burgoyne and Tan, 2008; Mandell *et al.*, 2010). The failure of conventional antibiotics to treat a HIV patient and its eradication from the human body have encouraged the evolution and spread of the disease. Prevention and treatment of the disease and transmission between individuals is a big challenge, thus new tools ought to be put in place which could eradicate the proliferation and the invasion of the body's immune system by HIV/AIDS. The development of new tools such as the Antimicrobial Peptides (AMPs) will be of great importance in the design of novel therapeutic drugs against HIV/AIDS, and overcome the adverse side effects experienced by the current treatment regimes.

- AMPs have been reported to play a role in the first line of defence in many organisms and have a wide range of activity against all types of microbes. They possess anti-viral properties, inhibiting viral fusion and egress; thus preventing infection and viral spread via direct interaction with the membranous viral envelope and the host cell surface molecules (Wang *et al.*, 2010). Moreover, unlike conventional antibiotics, which microbes readily circumvent, AMPs do not appear to induce antibiotic resistance, most likely due to profound changes in membrane structure warranted to confer the microbial cell with resistance (Andreu and Rivas, 1999; Fjell *et al.*, 2012). Presently, AMPs represent one of the most promising future strategies for combating

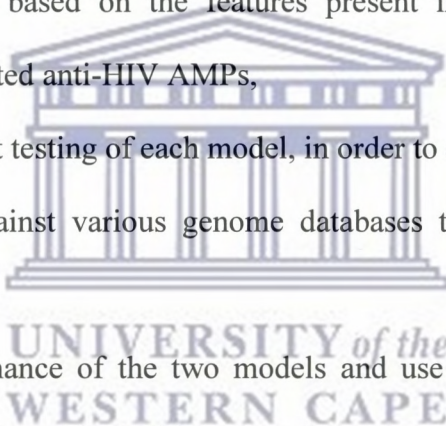


HIV infection and drug resistance to HAART. Consequently, many Antimicrobial Peptides have been shown to inhibit the activity of the HI Virus (Chang and Klotman, 2004; Wang *et al.*, 2008; Wang *et al.*, 2010).

- To date, a large body of molecular data on Antimicrobial Peptides have been generated and they had been accurately classified into different databases in respect to their biological activities. With the naissance of bioinformatics, proper classification of the AMPs in conjunction with available mathematical tools will assist in the discovery of putative Antimicrobial Peptides, and the construction of algorithm models for the designing of novel therapeutic compounds against the HI Virus. The *in silico* approaches are advantageous due to the fact that it is cost effective, less labour intense and less time consuming.
- Peptide modelling and clustering using SVM, HMMER, QSAR, GLAM2, LD, RF and ANN, combined with features selection have aided in the discovery of novel Antimicrobial Peptides (Torrent *et al.*, 2011; Torrent *et al.*, 2012). While SVM, QSAR, LD, ANN and RF require structure-activity relationship information in order to enhance their strength and performances, profile HMMER and GLAM2 only require the usage of peptide sequences to construct the models, which will displace their features based on the motifs of the input sequences. Conversely, the constructions of mathematical models such as profile HMMER and GLAM2 require the usage of naturally occurring, experimentally validated Antimicrobial Peptides, which have the desired activity against the specific target or pathogen. Both tools will be utilised in model creation to identify putative anti-HIV peptides.
- Thus, the aims of this project are to: (i) identify putative anti-HIV peptides using an *in silico* approach; (ii) Validate the putative anti-HIV peptides to confirm their therapeutic as well as the diagnostic abilities using an *in silico* approach.

In other to fulfil the gap in the development of putative Antimicrobial Peptides as therapeutic agent against HIV, the aims will be achieved as follow:

- ❖ Generate a final list of experimentally validated and inferred anti-HIV AMPs found in various databases (APD, CAMP, Bactibase, and UniprotKB). The experimentally validated anti-HIV AMPs retrieved from databases will be used as the training data set and the testing data set for model construction,
- ❖ Classify and separate these anti-HIV AMPs according to their super-families,
- ❖ Construct the predictive models representing each super-family using HMMER and GLAM2 algorithms, based on the features present in the primary sequences of experimentally validated anti-HIV AMPs,
- ❖ Carry out independent testing of each model, in order to assess their performances,
- ❖ Scan each model against various genome databases to retrieve putative anti-HIV peptides,
- ❖ Compare the performance of the two models and use the AMPs for further study generated by the model with the highest performance. The putative anti-HIV AMPs retrieve from genome scans of the various species will be ranked according to their E-values,
- ❖ Highest ranked putative anti-HIV AMP 3-D structures will be predicted and docked against the various HIV proteins (gp120, gp4, p24 and p17), to determine their binding affinities as well as orientation to these target HIV proteins to predict their activity for therapeutic purposes and their diagnostic capabilities.



## Chapter 2:

# Construction of HMMER and GLAM2 models for the identification of putative anti-HIV AMPs

### 2.1. Introduction

The first report of inhibition of HIV replication by synthetic guinea-pig, rabbit and rat  $\alpha$ -Defensins was in 1993 (Nakashima *et al.*, 1993). This study showed that  $\alpha$ -defensins could inhibit HIV-1 infection *in vitro* following viral entry into transformed CD4<sup>+</sup> T cells in the presence of serum (Nakashima *et al.*, 1993).

Several studies have investigated the anti-HIV activity of the defensins HNP1, HNP2 and HNP3 (Zhang *et al.*, 2002, Chang *et al.*, 2003). All three defensins showed similar activities against HIV primary isolates (Wu *et al.*, 2005) with at least two mechanisms proposed for their anti-HIV activity. Firstly, they can inhibit HIV-1 replication by direct interaction with the virus and secondly, by affecting the target cells (Chang *et al.*, 2003, Wang *et al.*, 2004). Electrostatic interaction used by HNPs, which are positively charged, directly binding to HIV virions might account for some of their direct inhibitory actions. The function of AMPs as lectins has also been reported by their binding to the HIV envelope glycoprotein gp120 and CD4<sup>+</sup> with high affinity, although their interference with the interaction between HIV gp120 and CD4<sup>+</sup> has not been well defined (Wang *et al.*, 2004).

Several other members belonging to the  $\alpha$ -defensin family of AMPs have also been investigated for their anti-HIV activity. Rhesus macaque myeloid  $\alpha$ -defensin-4 (RMAD4) blocks HIV replication at high concentrations however it is associated with cytotoxicity (Wang *et al.*, 2004). Members of the  $\beta$ -defensin AMP family, HBD2 and HBD3, showed dual anti-HIV activities similar to HNP 1, i.e. through direct interactions with the virus and indirectly by altering the target cell (Quinones-Mateu *et al.*, 2003, Sun *et al.*, 2005). Using electron microscopy, Quinones-Mateu *et al.*, 2003, showed the interaction between HBD2 and HBD3 with cellular membranes as well as HIV virions, although membrane disruption was not apparent. HBD2 does not affect cell-cell fusion but instead inhibits the formation of early reverse-transcribed HIV DNA products (Sun *et al.*, 2005). The  $\theta$ -defensins Retrocyclins, and RTD1, RTD2 and RTD3, function as lectins and can inhibit HIV entry (Munk *et al.*, 2003, Wang *et al.*, 2004). In addition they inhibit several HIV-1 X4 and R5 viruses, including primary isolates (Munk *et al.*, 2003, Wang *et al.*, 2004). Retrocyclin does not seem to inactivate the HIV virion directly unlike  $\alpha$ - and  $\beta$ -defensins, however, it binds to HIV gp120 as well as CD4<sup>+</sup> with high affinity. Most recently a peptide named Kn2-7, a derivative of BmKn2 cloned from the venom of the scorpion *Mesobuthus martensii* Karsch, was shown to be the most potent anti-HIV-1 peptide to date. Direct interaction was shown between Kn2-7 and the HIV-1 envelope (Chen *et al.*, 2012).

The driving force behind the development of newer anti-infectives has almost always been the inevitable emergence of bacterial resistance to antibiotics following widespread clinical, veterinary, and animal agricultural usage. This demand has been continuously met by the pharmaceutical industry by the modification of existing antibiotics and development of newer antibiotics in a timely fashion. Similarly, the development of effective anti-virals to eliminate important clinical viral pathogens (e.g. HIV, herpesviruses, and influenza) has also shown dramatic successes. Notwithstanding these advances in drug development, the rapid

emergence of resistance is even a greater problem for life-threatening viral infections. HIV remains one of the best examples where the rapid emergence of resistance to single drugs posed daunting clinical problems. The only effective solution to this problem would thus seem to develop combination therapy involving several antivirals with different mechanisms of inhibitory action.

Although AMPs generally exhibit lower potency against susceptible microbial targets compared to conventional low molecular weight antibiotic compounds, they hold several compensatory advantages including: (i) fast killing (ii) broad range of activity (iii) low toxicity and (iv) minimal development of resistance in target organisms (Jenssen *et al.*, 2006) For AMPs there are several different potential strategies for their general therapeutic application: (a) as single anti-infective agents, (b) in combination with conventional antibiotics or anti-virals to promote any additive or synergistic effects, (c) as immunostimulatory agents that enhance natural innate immunity, and (d) as endotoxin-neutralizing agents to prevent the potentially fatal complications associated with bacterial virulence factors that cause septic shock. In light of the increased resistance to antibiotics in pathogenic microorganisms, AMPs have drawn significant attention as possible sources of novel antimicrobial agents specifically against HIV/AIDS (Hancock and Sahl, 2006).

It stands to reason as the demand increases for the identification of AMPs, parallel technologies are developed to meet this demand. With the birth of Bioinformatics, a host of technologies using *in silico* approaches to identify AMPs has been developed. These methods have promised to be less time consuming, more cost effective and less labour intensive thus speeding up the discovery process. These *in silico* methods have been further boosted by the establishment of various AMP databases such as Bactibase (Hammami *et al.*, 2007), APD2 (Wang and Wang, 2004; Wang *et al.*, 2009), AMPer (Fjell *et al.*, 2007) and CAMP (Thomas *et al.*, 2009). These databases not only host a wealth of AMPs, but also incorporate many

embedded algorithms for AMP identification thus providing an indispensable knowledge base for both qualitative and quantitative activity prediction models using tools such as Support Vector Machine (SVM) (Thomas *et al.*, 2010), profile Hidden Markov Models (HMMER) (Brahmachary *et al.*, 2004; Fjell *et al.*, 2007), Gap Local Alignment of Motifs 2 (GLAM2) (Frith *et al.*, 2008), Quantitative Structure Activity Relationship (QSAR) (Fjell *et al.*, 2009), Linear Discriminant Analysis (LD) (Thomas *et al.*, 2010), Random Forest (RF) (Thomas *et al.*, 2010) and Sliding Window (SW) (Torrent *et al.*, 2012). The knowledge-based approach has allowed for the systematic mining of genomic expressed sequence tag data. The aim of which was to discover hitherto undescribed natural AMP sequence (Juretić *et al.*, 2011).

Prediction models can also be constructed independently, with the incorporation of properties related to the structure and the activity of existing peptides using, the same tools. Whilst SVM, QSAR, LD and RF require structure-activity relationship information in order to enhance their strength and performances (Torrent *et al.*, 2012), HMMER and GLAM2 only require the usage of peptide sequences in the construction of a model, which will display their features based on the motifs of the input sequences (Eddy, 1998; Frith *et al.*, 2008).

The constructions of mathematical models such as profile HMMER and GLAM2 require the usage of naturally occurring, experimentally validated Antimicrobial Peptides, which have the desired activity against the specific target or pathogen (for example anti-HIV). To construct these models, the experimentally validated peptide data set is divided into two sets, namely the training and the testing data set. The training data set is used in the construction of the model or training of the model, whereas the testing data set is used in validating the robustness of the model.

The aim of this chapter is to construct a sensitive and specific algorithm model for identification of putative anti-HIV AMPs. To accomplish this, the objectives of this chapter are as follows:

- ❖ To collect all experimentally validated anti-HIV AMPs from Antimicrobial Peptides Database (APD), Collection of Anti-Microbial Peptides (CAMP), Cybase and UniProt Knowledgebase (UniProtKB), eliminate duplicates from these lists and classify these peptides into two groups namely experimentally validated and predicted, followed by their classification according to their families or super-families;
- ❖ To use these groups of families or super-families of anti-HIV peptides in the construction of GLAM2 and HMMER profiles;
- ❖ To use these models to query various genome sequences in order to retrieve putative anti-HIV peptides.

Section 2.2.1 presents the retrieval of experimentally validated and predicted anti-HIV AMPs from various databases. In addition, the elimination of duplicates will be explained. Section 2.2.4 will present the procedure for constructing HMMER and GLAM2 models and finally section 2.2.5 gives the performance measures of the models created using HMMER and GLAM2.

## 2.2. Materials and methods

Experimentally proven AMPs showing activity towards HIV have been placed into several repository databases, and the most updated databases include APD, CAMP, Cybase and UniProtKB (Wang *et al.*, 2008; Wang *et al.*, 2009; Thomas *et al.*, 2010). These experimentally validated anti-HIV peptides will be used to construct stochastic models, using HMMER and GLAM2, which will aid in the scanning of various genome sequences to identify putative anti-HIV peptides. In this regard, peptides will be retrieved from these databases.

### 2.2.1. Data mining: Extraction of anti-HIV antimicrobial peptides from the various databases

#### 2.2.1.1. Antimicrobial Peptides Database (APD)

Antimicrobial Peptides Database (APD) is a manually curated database held by the Department of Pathology and Microbiology, at the University of Nebraska Medical Centre. The first version of this database was published in 2004 as APD and a later version in 2009 as APD2 (Wang and Wang, 2004; Wang *et al.*, 2009). The database has entries on Antimicrobial Peptides classified as antiviral, antifungal, anticancer/tumour, antibacterial, anti-protists, anti-parasital, insecticidal, spermicidal, anti-HIV and AMPs with chemotactic activity.

To retrieve the anti-HIV peptides, the link <http://aps.unmc.edu/AP/main.php> was opened. From the menu, the term “anti-HIV peptides” was selected and a new window containing all the peptides matching this term was displayed. Eighty-eight Antimicrobial Peptides matching

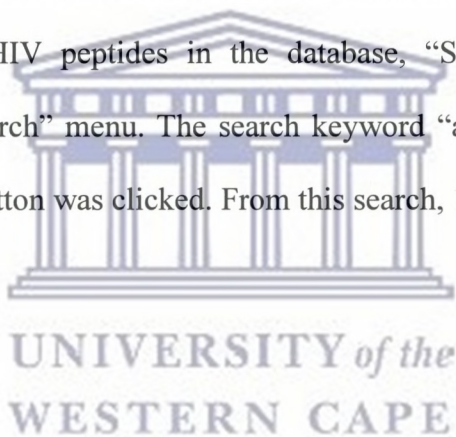


this search term was identified. The amino acid sequences of these AMPs were extracted and saved in the FASTA format as a text document.

#### **2.2.1.2. Collection of Anti-Microbial Peptides (CAMP)**

CAMP is a Comprehensive Anti-Microbial Peptide Database which incorporates experimental as well as predicted Antimicrobial Peptides, and contains various tools for AMP prediction. The database is hosted by the Biomedical Informatics Centre (BIC), which forms part of the National Institute for Research in Reproductive Health (NIRRH). The first version of the database was published in 2010 (Thomas *et al.*, 2010).

The link to the database <http://www.bicnirrh.res.in/antimicrobial/index.php> was opened. To retrieve the available anti-HIV peptides in the database, “Search” menu was selected, followed by the “Simple search” menu. The search keyword “antiviral” was entered in the search box and the submit button was clicked. From this search, 117 anti-HIV AMPs matched the keyword.



#### **2.2.1.3. Cybase**

Cybase is a database of cyclic protein sequences and structures, of which the first publication was in 2004 and its second version in 2008 (Mulvenna *et al.*, 2006; Wang *et al.*, 2008). The Cybase database is managed at the Institute of Molecular Bioscience (IMB), at the University of Queensland, Brisbane, Australia. The link of the database is <http://www.cybase.org.au/>.

The retrieval of experimentally validated anti-HIV peptide from the Cybase database was performed by going to the URL mentioned above. At the URL home page, under the category “Proteins”, the option “Search” was selected. Selection of this option opened the protein search page where the parameters were left as default, except for the parameter “Assay”

where the term “Anti-HIV” was entered. For the category “Result list column” multiple output options were selected, for example ID, Name, Sequence, References. Once all the search parameters were selected, the search option was chosen on this page. From the query keyword, 28 Antimicrobial Peptides were found to have anti-HIV activity.

#### **2.2.1.4. UniProt Knowledgebase (UniProtKB)**

UniprotKB is a Protein knowledge-base and it consists of two sections: UniProtKB/Swiss-Prot, which is manually annotated and reviewed; and UniProtKB/TrEMBL, which is automatically annotated and is not reviewed (Uniprot, 2009). UniProtKB website is <http://www.uniprot.org/>. For the retrieval of experimentally validated anti-HIV peptides, the UniProt/Swiss-Prot section of UniProtKB was used.

The collection of anti-HIV peptides was done by going to the URL mentioned above. On the home page of UniProtKB, the keyword “Antimicrobial peptides” was entered into the “query” box followed by selecting the “Search” menu. This was followed by choosing the “Advanced search” mode. Under the “Advanced search” option, the Boolean operator “AND” was selected as well as under the option “Field”, “All” was selected and under the option “Term”, the keyword “anti-HIV” was entered and the “Add and Search” option was selected. After the search, a list of 27 anti-HIV peptides which match the keyword was generated.

#### **2.2.2. Literature mining**

The reliability of the nature and the activity of the retrieved anti-HIV peptides can only be considered for model creation if they are proven and confirmed experimentally to have

activity against HIV. In addition, the source of the peptides activity had to be verified, that is, if they are experimentally validated, predicted or synthetic peptides. Literature mining was done using references from published articles, in relation to each anti-HIV peptide. This was done to verify that the reference stated in the databases was correct. Literature mining was done using Google scholar, Science direct and NCBI/PubMed. Results of the literature mining of all the experimentally validated Antimicrobial Peptides were arranged according to their respective database.

### **2.2.3. Elimination of duplicates and generation of final anti-HIV list**

The final refinement of the experimentally validated anti-HIV peptides was done by eliminating the duplicates. This elimination was made on the basis of their given name and not their ID, since the different databases have different ID but the same given anti-HIV name. The names which appeared twice or more in the combined list of all the databases were eliminated. This elimination was done manually since the numbers of anti-HIV peptides were not many.

Elimination was also done on the basis that the activity of an Antimicrobial Peptide is mainly coded for by the mature part of the peptide and not by the premature or precursor part. Further classification was made to group the experimentally validated and the predicted anti-HIV AMPs according to their families or super-families of origin. This classification of the anti-HIV AMPs resulted in the categorisation of 7 super-families of AMPs.

#### 2.2.4. Construction of Hidden Markov Models (HMMER) and Gap Local Alignment of Motifs2 models (GLAM2)

These two algorithms work by the same principle of motif profile description generated by sequences of the same family, however the two tools differ in their statistic scoring and probability prediction annotation. The models generated by these tools can however serve as templates to search for sequences of the same signature or the same sequences of selected families in databases. To achieve this, each super-family of anti-HIV AMPs were divided into a training set and a testing set. The training sets represented approximately three quarters of each super-family while the testing sets accounted for approximately one quarter of each super-family. The number of experimentally validated anti-HIV AMPs used for the training and testing of each super-family model are represented in **Table 2.1**.

**Table 2.1:** The number of anti-HIV AMPs used for the construction and testing of each super-family model built by HMMER and GLAM2

Super-families	Number of experimentally validated anti-HIV AMPs	Training sets	Testing sets
Amphibians	27	22	5
Microorganisms	4	3	1
Human Defensins	9	7	2
Fish and Crabs	5	4	1
Insects	6	5	1
Vertebrates	8	6	2
Plants	33	22	11

The training sets were used to construct the profiles and the testing sets were used to query the profiles. The sequences specific for each super-family used as the training sets for HMMER and GLAM2 models construction and testing sets of each super-family are listed in Table 2.5 and Table 2.6 of the results section of this chapter.

#### 2.2.4.1. Hidden Markov Models profiles

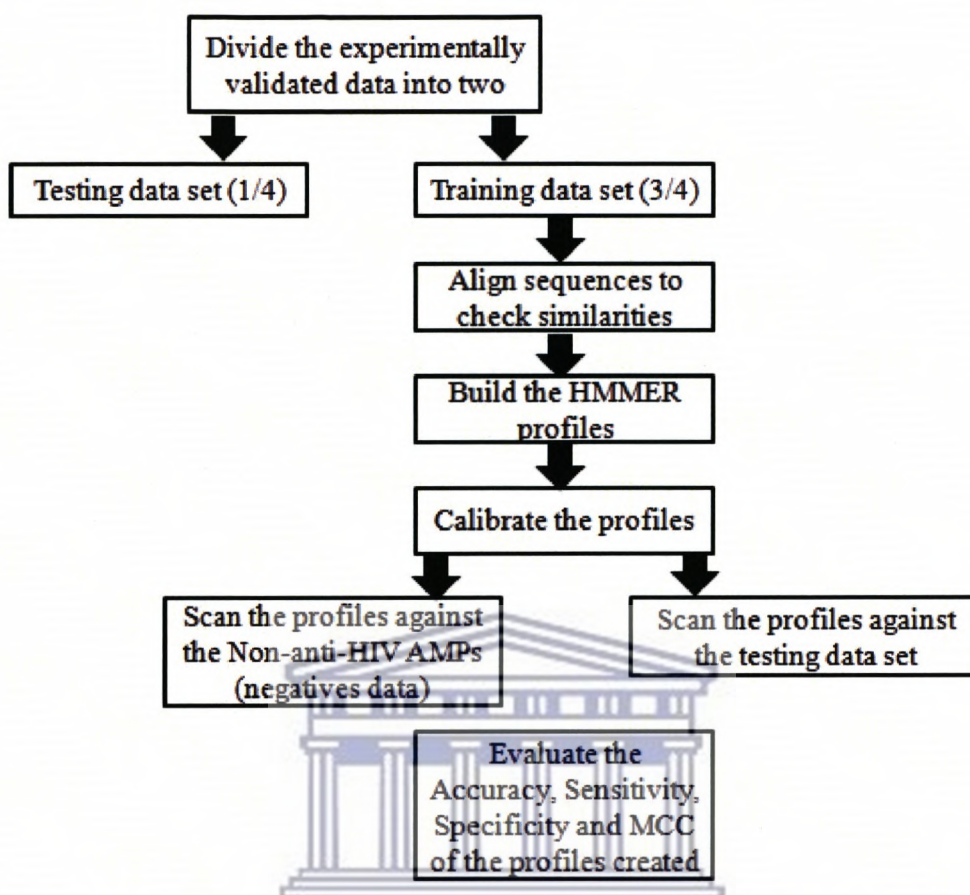


Figure 2.1: Architecture of the proposed method to build profiles using the profile HMMER algorithm

##### a) Construction of Hidden Markov Models profiles

Since the extracted experimentally validated anti-HIV AMPs were grouped as super-families (Table 2.1), the same steps were used for the profile construction of the individual HMMER models of each specific super-family. All the HMMER models were constructed on Ubuntu 12.04 LTS operating system, which is based on the Linux kernel and the Linux distribution Debian with HMMER 2.3.2 version.

The profile HMMER has multiple modules for it to perform properly. For the first step, the training set of each super-family were aligned using the ClustalW alignment tool. The alignment was performed using the command line

```
clustalw -align -output=gcg -case=upper -sequos=off -outorder=aligned -infile=family.fasta
```

The input sequences were in the FastA format as “family.fasta”. The language of the command line is simply stating <<do an alignment of the sequences which are in the upper case found in the input file “family.fasta” with the FastA, using ClustalW as multiple alignment tool and GCG Postscript output for graphical printing>>. The file generated from the alignments of the sequences or domains of each super-family was saved as “family.msf”. In particular, msf files are created using the above command. These files were used as input in the second step of HMMER for profile building, known as “Build profiles”. This step enables the creation of profiles/signatures of family sequences by showing the common motifs within the model. The “Build profiles” was run using the command

```
hmmbuild family.hmm family.msf
```

To enhance the sensitivity of the profile, the file generated from the profile building was calibrated by using the command line

```
hmmcalibrate family.hmm
```

The output profiles at this stage of the model construction is a calibrated model, which can be queried using the testing sets or can be used to search for additional anti-HIV AMPs if each family specific profile is used to query genome sequences within databases.

#### **b) Independent testing of Hidden Markov Models profiles**

To measure the performance of the constructed profiles of the specific super-families, the testing set of each super-family was used to query the profile of that super-family. The numbers of sequence used for the testing of each specific super-family are represented in **Table 2.1**.

The “Query profile” step helps to predict if the testing sets of a particular super-family truly belongs to the profile for that super-family. This is achieved by querying the model of the super-family with its testing set. In addition, this is to determine if the sequences of the testing set belongs to the constructed models which were built based on primary sequences of the training set. The query of profiles also confirmed that the testing and the training sets have anti-HIV activity since both sets were derived from the list of experimentally validated anti-HIV AMPs. The query profiles were carried out using the command line.

```
hmmsearch -E 5e-2 family.hmm familyquery (fasta format) > resultfile.txt
```

For each super-family, the cut-off E-value was 0.05. The profiles were also queried on a negative dataset consisting of 596 neuro-peptides sequences which are non anti-HIV peptides. The step helps to identify the number of true positive, false negative, true negative and false negative Antimicrobial Peptides, which enabled the performance calculation of the constructed models.

### **c) Querying genome sequence databases using Hidden Markov Models profiles**

The second step of the query consists of database queries know as “Query db”. More than 1059 genome sequences were queried by the respective super-families profiles with the list of all genome sequences searched was retrieved from the Ensembl database (<http://www.ensembl.org/index.html>) and the Uniprot database (<http://www.uniprot.org/>).

```
hmmsearch -E 1e-2 family.hmm familyquery (fasta format) > resultfile.txt
```

The cut-off E-value was set to be 0.01. The “Query db” was to identify peptides which had the same signatures/motifs and properties as the profiles of the various super-families. The

identified peptides were considered as putative and considered having activity against the HIV. The list of the putative anti-HIV AMPs was generated for further analysis.

#### 2.2.4.2. Gap Local Alignment of Motifs 2 profiles

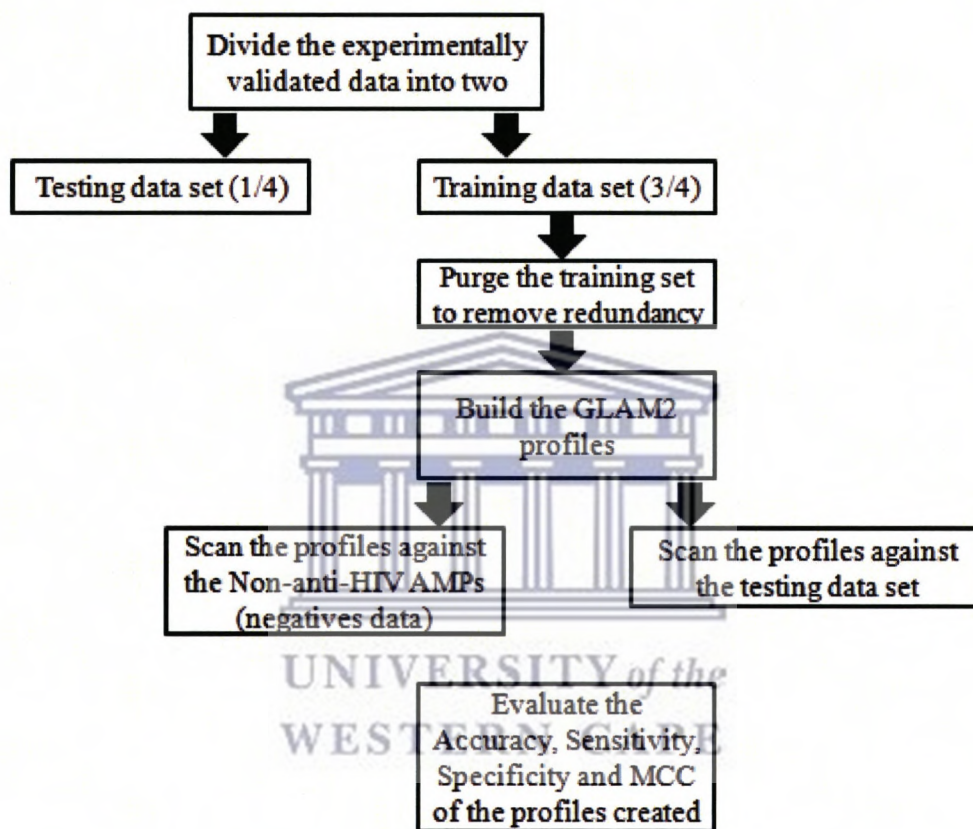


Figure 2.2: Architecture of the proposed method to build profiles using the GLAM2 algorithm

##### a) Construction of Gap Local Alignment of Motifs 2 profiles

All the GLAM2 profiles were also constructed on Ubuntu 12.04 LTS operating system, which is based on the Linux kernel and the Linux distribution Debian with the GLAM2 version 9999. To construct these super-family specific profiles, the same proportion of experimentally validated anti-HIV AMPs that were used for the model construction of the profile HMMER models were also used for the model construction of GLAM2 profiles



(Table 2.1). This was done in order to compare the performance of HMMER and GLAM2 in terms of accuracy measure and to see if the two tools can give the same result during the genome scans to search for additional and putative anti-HIV peptides. The same steps were used in model construction for each super-family. The construction of these profiles consists of two programs, which corresponds to two steps namely, the purge and the GLAM2 steps.

The construction of the profiles started with the “Purge” module, in which the training sets of each super-family were aligned in order to remove the redundant sequences within each super-family. The threshold for the alignment was chosen as 150 for each of the families as at this threshold, similar anti-HIV AMPs sequences were removed from the training sets. The “Purge” was to prevent highly similar sequences distorting the search for motifs. The purge was executed using the command line for each super-family.



```
glam2-purge family.fasta 150
```

The output files were saved as “family.fasta.b150”. After this step, the profile of each super-family was built using the alignment output files “family.fasta.b150” as the input files of the “Build profiles” step. The build profile also known as “glam2” was performed with the command line

```
glam2 -o family.glam p family.fasta.b150
```

The profiles of all seven super-families were generated. These profiles will be utilised in the search of putative peptides which have anti-HIV activity in genome sequence databases.

## **b) Independent testing of Gap Local Alignment of Motifs 2 profiles**

Though the profile GLAM2 construction differs to that of HMMER, both tools have the same module of query since both tools have a “Query profile” in which the testing sets are used to query the models created by using a training set and the “Query db” called query databases step, in which the models are scanned against various genome sequence databases. This was done once the profiles were created.

The strength of the models constructed with the GLAM2 algorithm for the specific super-families were also queried at the step “query profiles”. The testing set of each super-family was used to query the profile of that super-family (**Table 2.1**). The “Query profile” was done with the testing dataset on each profile specific super-family using the command line

```
glam2scan -o testfamily_result.glam2 p family.glam2 testfamily.fasta
```

Unlike HMMER, no E-value is required in this command line of the GLAM2 program. Also a negative dataset consisting of 596 neuro-peptides sequences, which are non anti-HIV peptides, were used to query the profiles to confirm the uniqueness of the profiles. Similar to HMMER, this step also helps to identify the number of true positive, false negative, true negative and false negative Antimicrobial Peptides, which enable the performance calculation of the constructed models.

## **c) Querying genome sequence databases using Gap Local Alignment of Motifs 2 profiles**

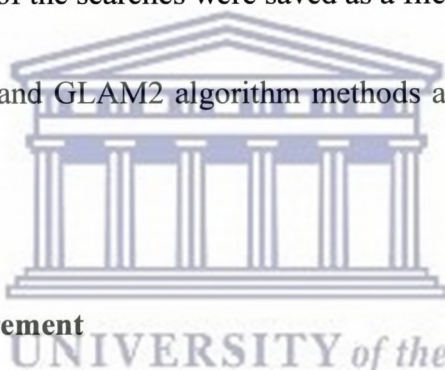
The database query known as “Query db” was performed to identify peptides having the same signatures/motifs and properties as the profiles. The database query of GLAM2 profiles

was performed with the same number of genome sequences as for HMMER which, the list of all the genome sequences searched was retrieved from the Ensembl database (<http://www.ensembl.org/index.html>) and the Uniprot database (<http://www.uniprot.org/>). The search was executed as

```
glam2scan -o family_result.glam2 p family.glam2 queryfamily.fasta
```

The identified peptides were assumed to have anti-HIV activity, hence can be considered as putative anti-HIV AMPs and a list of the putative anti-HIV AMPs was generated for further analysis. The usage of “-o” in the command line was to write the output to a file rather than to the screen thus all the results of the searches were saved as a file in the GLAM2 folder.

The main steps of HMMER and GLAM2 algorithm methods are summarised in **Figure 2.1** and **Figure 2.2** respectively.



#### 2.2.5. Performance measurement

To measure the performance of both tools: HMMER and GLAM2, the same performance measures were calculated to assess the strength of the constructed models. Such measures included the sensitivity, specificity, accuracy and Matthew’s Correlation Coefficient (MCC). The performances are described as follows:

- **Sensitivity** is the percentage of anti-HIV AMPs (testing sets) correctly predicted as anti-HIV AMPs (positive). The sensitivity (recall) is defined as:

$$\text{Sensitivity} = \left( \frac{TP}{TP + FN} \right) \times 100$$

- **Specificity** is the percentage of non-anti-HIV AMPs (negative sets) correctly predicted as non-anti-HIV AMPs (negative). The specificity is defined as:

$$\text{Specificity} = \left( \frac{TN}{TN + FP} \right) \times 100$$

- **Accuracy** is the percentage of correctly predicted peptides (anti-HIV AMPs and non-anti-HIV AMPs). The accuracy is defined as:

$$\text{Accuracy} = \left( \frac{TP + TN}{TP + FP + TN + FN} \right) \times 100$$

- **Mathew's correlation coefficient (MCC)** is a measure of both sensitivity and specificity. MCC = 0 indicates completely random prediction, while MCC = 1 indicates perfect prediction. It is defined as:

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

In this definition of the above performance measures, we define the followings:

*TP* (True positive) represents correctly predicted positive examples (anti-HIV AMPs),

*TN* (True negative) is correctly predicted negative examples (non-anti-HIV AMPs),

*FP* (False positive) is the number of non-anti-HIV AMPs examples wrongly predicted as anti-HIV AMPs,

*FN* (False negative) is the number of anti-HIV AMPs wrongly predicted as non-anti-HIV AMPs.

After performance calculation and scanning of various genome sequences databases duplicate sequences were removed to have a final list of putative anti-HIV AMPs.

## 2.3. Results

### 2.3.1. Data mining

The curation of the publicly available Antimicrobial Peptide databases has provided a number of experimentally validated anti-HIV AMPs, which are found across various plant and animal kingdoms, in eukaryotic as well as in prokaryotic organisms. These experimentally validated anti-HIV AMPs after retrieval gave us a final list of anti-HIV AMPs to be used in model construction as listed in **Table 2.2**.

**Table 2.2:** Results of all the anti-HIV AMPs entries found into the query databases

Databases	Query terms	Number of peptides from the query
APD	anti-HIV	88
CAMP	anti-viral	117
Cybase	anti-HIV	28
UniProtKB	Antimicrobial and HIV	27

### 2.3.2 Literature mining

Since the interest was in naturally occurring, experimentally validated and predicted anti-HIV peptides, confirmation of the activity of each anti-HIV was done through literature mining of each reference provided for the peptides by the respective AMP databases. Further literature mining was done for those Antimicrobial Peptides which did not have the relevant reference assigned to it. The purpose was to check if the peptides are naturally occurring or if they are synthetic peptides, and if they have proven anti-HIV activity. As such, the literature mining revealed that APD, CAMP, Cybase and UniProtKB had 79, 36, 23 and 27 experimentally validated anti-HIV peptides respectively as shown in **Table 2.3**. Some peptides were not included because they were not from natural sources but synthetic or had other antiviral activity.

**Table 2.3:** Number of experimentally validated and predicted anti-HIV peptides retrieve from the databases

Databases	Number of Peptides			
	Experimentally validated anti-HIV AMPs	Predicted or Inferred anti-HIV AMPs	Synthetic anti-HIV AMPs	Other anti-viral AMPs
APD	79	5	4	0
CAMP	36	3	0	78
Cybase	23	0	5	0
UniProtKB	27	0	0	0

### 2.3.3. Elimination of duplicates and generation of final anti-HIV list

The elimination process from the initial lists of anti-HIV peptides, has enabled the creation of a more refined final list of the naturally occurring, mature, experimentally validated anti-HIV peptides and predicted anti-HIV peptides. On this note, it was observed that the entire peptide from the amphibian sub-family Maximins 1, 3, 4 and 5 had the same therapeutic indices as the mature part of these peptides (Chen *et al.*, 2001; Lai *et al.*, 2001). Thus, only the mature Maximins class of the peptides was considered for use in model construction.

The final list comprised of 92 experimentally validated and 7 predicted anti-HIV peptides from the various databases after data and literature mining. The list of anti-HIV peptides was further classified into different families or super-families from which they originated, based on literature. The experimentally validated anti-HIV peptides were partitioned into families or super-families as shown in **Table 2.4**. The partitioning of the predicted anti-HIV peptides into different super-families resulted in 4, 1 and 2 anti-HIV peptides originating from Plants, Vertebrates and Microorganisms respectively.

The classification results of the 92 experimentally validated anti-HIV peptides and 7 predicted anti-HIV peptides according to different families or super-families are represented in **Table 2.4**, showing the diversity of Antimicrobial Peptides and also to enable a better

grouping of peptides as to construct a super-family specific model for each. The proof of anti-HIV activities of these peptides is shown in the **Supplementary material Table A.1**.

**Table 2.4:** Final list of all experimentally validated and predicted anti-HIV peptides retrieved from the databases

	Name of anti-HIV peptides	
Families or Super-families	Experimentally validated anti-HIV	Predicted or Inferred anti-HIV
<b>Amphibians</b>	Aurein 1.2, RANATUERIN 2P, Magainin 2, Dermaseptin-S4, Dermaseptin-S1, Maximins 1, Maximins 3, Maximins 4, Maximins 5, Caerin 1.1, Caerin 1.9, Caerin 4.1, Maculatin 1.1, Uperin 3.6, Uperin 7.1, RANATUERIN 6, RANATUERIN 9, Brevinin-2-related peptide, Maculatin1.3, Esculentin-2P, Dahlein5.6, Dermaseptin-S9, Ascaphin-8, Palustrin-3AR, Esculentin-1ARb, Temporin-LTc, Temporin-PTa	
<b>Insects</b>	Melittin, Cecropin A, Cicadin, Melectin, Ponericin L2, Spinigerin	
<b>Human Defensins</b>	HNP-4, HNP-3, HNP-2, HNP-1, SLPI, hBD-2, hBD-3, human Histatin 5, LL-37	
<b>Plants</b>	Sesquin, Ginkbilobin (GNL), Alpha-basrubrin, Beta-basrubin, Ascalin, Gymnin, Coccinin, Antifungal protein from coconut, Thaumatin-like protein, Thaumatin-like protein-Actc 2, palicourein, Circulin A, circulin B, circulin C, circulin D, circulin E, circulin F, cycloviolacin O12, cycloviolacin O13, cycloviolacin O14, cycloviolacin O24, cycloviolacin Y1, cycloviolacin Y4, cycloviolacin Y5, cycloviolin A, cycloviolin B, cycloviolin C, cycloviolin D, vhl-1, kalata B1, kalata B8, Lunatusin,	Tricyclon A, TPA_exp: DEFB4-like protein, TPA_exp: DEFB 103-like protein, Shepherdin II
<b>Vertebrates</b>	Indolicidin, L-amino-acid oxidase, BMAP-27, RTD-1, RTD-2, RTD-3, Protegrin 1, Alpha-MSH	Lactoferricin B,
<b>Microorganisms</b>	Gramicidin A, Siamycin I, NP-06, Agrocybin	RP 71955, Siamycin II
<b>Fish and Crabs</b>	Tachyplesin-1, Polyphemusin I, Polyphemusin II, Clavanin B, Piscidin 3	

#### 2.3.4. Sequences used to create and test the respective models

Besides classifying the experimentally validated anti-HIV AMPs according to their respective families, their amino acid sequences were also retrieved from the databases. The peptide sequences served as input data for the construction of super-family specific profiles. The sequences of these anti-HIV AMPs were randomly divided into the training set and the testing set for the purpose of model creation. The partitioning ( $\frac{3}{4}$ :  $\frac{1}{4}$  ratio) of each super-family for the training and testing sets is represented in **Table 2.1**. The sequences used for both the training and the testing sets are represented in **Table 2.5** and **Table 2.6**.





**Table 2.5:** Sequences of the experimentally validated anti-HIV AMPs used as the training set to create the family specific profiles

Super-families	Anti-HIV AMPs names	Anti-HIV AMPs Sequences
Amphibians	Magainin2	GIGKFLHSAKKFGKAFVGEIMNS
	Dermaseptin-S4	ALWMTLLKKVLAATAKALNAVLVGANA
	Dermaseptin-S1	ALWKTMLKKLGTMALHAGKAALGAAADTISQGTQ
	Caerin1.1	GLLSVLSVAKHVLPHVVPVIAEHL
	Caerin1.9	GLFGVLSIAKHVLPVVPVIAEK
	Caerin4.1	GLWQKIKSAAGDLASGIVEGIKS
	Maculatin 1.1	GLFVGVLAHVAAHVVPAIAEHF
	Uperin3.6	GVIDAAKKVNVVILKNLF
	Uperin7.1	GWFDVVKHIASAV
	RANATUERIN6	FISAIASMLGKFL
	RANATUERIN9	FLPPLITSFLSKVL
	Brevinin-2-relatedpeptide	GLLGLLSVWVSHVVPVAVGHF
	Maculatin1.3	GFSSIFRGVAKFASKGLGKDLARLGVNLVACKISKQC
	Esculentin-2P	GLLASLKGVFGGYLAEKLPK
	Dahlein5.6	GIWDTIKSMGKVFAGKILQNL
	Dermaseptin-S9	GLRSKIWLWLLMIWQESNKFKKM
	Ascaphin-8	GFKDLLKGAAKALVKTVLF
Palustrin-3AR	GIFPKIIGKIVNGIKSLAKGVGMKVFAGLNNIGNTGNNRDEC	
Esculentin-1ARb	GLFPKFNKKKVKTGIFDIKTVGKEAGMDVLRGTGIDVIGCKIKGEC	

	Temporin-LTc	SLSRFLSFLKIVYPPAF
	Temporin-PTa	FFGSVLKLIKIL
	Maximin_1	GIGTKILGGVKTALKGALKELASTYAN
<b>Microorganisms</b>	GramicidinA	VGALAVVWLWLWLW
	NP-06	CLGVGSCNDFAGCGYAIVCFW
	Agrocybin	ANDPQCLYGNVAAKF
<b>Human Defensins</b>	hNP-2	CYCRIPACIAGERRYGTCIYQGRLWAFCC
	hNP-1	ACYCRIPACIAGERRYGTCIYQGRLWAFCC
	SLPI	SGKSFKAGVPPKSAQCLRYKKECQSDWQCPGKKRCCPDTCGIKCLDPVDTNPTRRK PGKCPVTYGOCLMLNPPNFCMDGQCKRDLKCCMGKSCVSPVKA
	hBD-1	DHYNCSVSSGGQCLYSACPIFTKIQTGYRGKAKCCK
	hBD-2	GIGDPVTCLKSGAICHVPFCPRRYKQIGTCGLPGTKCCKKP
	hNP-3	DCYCRIPACIAGERRYGTCIYQGRLWAFCC
	LL-37	LLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLRVPTES
	humanHistatin5	DSHAKRRHHGYKRRKFHEKHHSHRGY
<b>Fish and Crabs</b>	PolyphemusinII	RRWCFRVCYKGFYRKCRCR
	ClavaninB	VFQFLGRIIHVGNFVHGFSHVF
	Piscidin1	FFHHIFRGIVHVGKTIHRLVTG
<b>Insects</b>	Melittin	GIGAVLKVLTGTLPALISWIKRKRQQ
	Cecropin A	KWKLFFKIEKVGQNIIRDGIKAGPAVAVVGQATQIAK
	Spinigerin	HVDKVVADKVVLLKQLRIMRLLTRL

	Ponericin L2	LLKELWTKIKGAGKAVLGKIKGLL
	Melectin	GFLSILKKVLPKVMAMHK
<b>Vertebrates</b>	RTD-1	GFRCLCRRRGVCRCICTR
	Protegrin1	RGGRLCYCRRRRFCVCVGR
	RTD-3	RCICTRGFRCICTRGFC
	Alpha-MSH	SYSMEHFRWGKPV
	Indolicidin	ILPWKWPWPWRR
	L-amino-acidoxidase	MNVFEMFSLFLAALGSCADDRNPLEECFRETDEEFLIARNGLKATSNPKHVIVGAGMSG LSAAYVLAGAGHEVTVLEASERAGGRVRTYRNDEEGWYANLGPMLPEKHRIVREYIRKFNLQ LNEFSQENDNAWHFVKNIIRKTVGEVKKDPGVLKYPVKPSEEGKSAEQLYEESLRKVEKELKRITN CSYILNKYDPTSTKEYLKEGNLSPGAVDMIGDLMNEDAGYYVSFIESMKHDDIFAYEKRFDEIVD GMDKLPMSMYRAIBEKVHFNAAQVIKIQKNAEEVTVTYHTPEKDTSFVTADYVIVCTTSRAARRIKF EPPLPKKAHALRSVHYRSGTKIFLCTCKKFREDEGIHGGKSTDDLPSRFIYYPNHNF TSGVGVIIAY GIGDDANFFQALDLKDCGDIVNDLSLHQLPREEIQTFCYPSMIQKWSLDKYAMGGITFTTPYQFQ HFSEALTSHVDRIYFAGEYTAHAHGWDSSIKSGLTAARDVNRASENPSGIHLSNDDDEL
<b>Plants</b>	Sesquin	KTCENLADTV
	Ginkbilobin(GNL)	ANTAFVSSAHNTQKIPAGAPFNRLRAMLADLRQNAAFAG
	Alpha-basrubrin	GADFQECMKEHSQKQHQQG
	Ascalin	YQCGQGG
	Gymnin	KTCENLADY
	Thaumatococin-likeprotein,Actc2	ATFNFINNCPFTVWAAAVPG
	Palicourenin	GDPTFCGETCRVIPVCTYSAAALGCTDDRRSDGLCKRN
	CirculinA	GIPCGESCWIPICISAAALGCCKNKVCYRN
	CirculinD	KIPCGESCWIPCVTSIFNCKCENKVCYHD

	CirculinE	KIPCGESCWWIPCLTSVFNCKENKVCYHD
	CirculinF	AIPCGESCWWIPCISAAIGCSCKNKVCYR
	CycloviolacinO12	GLPICGETCVGGTCNTPGCSCSWPVCTR
	CycloviolacinO13	GIPCGESCWWIPCISAAIGCSCKSKVCYRN
	CycloviolacinO24	GLPTCGETCFGGTCNTPGCTCDPWPVCTHN
	CycloviolacinY1	GGTIFDCGETCFLGTCTYTPGCSCGNYGFCYGTN
	CycloviolacinY4	GVPCGESCVPICITGVIGSCSSNVCYLN
	CycloviolinA	GVIPCGESCVPIPCISAAIGCSCKNKVCYRN
	CycloviolinB	GTACGESCYYLPQFTVGGCTCTSSQCFKN
	CycloviolinC	GIPCGESCVPICLITVAGCSCKNKVCYRN
	Vhl-1	SISCGESCAMISFCFTEVIGCSCKNKVCYLN
	KalataB1	GLPVCGETCVGGTCNTPGCTCSWPVCTR

**Table 2.6:** Sequences of the experimentally validated anti-HIV AMPs used as the testing set for the family specific profiles

Super-families	Anti-HIV AMPs names	Anti-HIV AMPs Sequences
<b>Amphibians</b>	Aurein1.2	GLFDIIKKIAESF
	RANATUERIN2P	GLMDTVKNVAKNLAGHMLDKLKCKITGC
	Maximin_3	GIGGKILSGLKTALKGAAKELASTYLH
	Maximin_4	GIGGVLLSAGKAALKGLAKVLAEKYAN
	Maximin_5	SIGAKILGGVKTFKKGALKELASTYLQ
<b>Microorganisms</b>	SiamycinI	CLGVGSCNDFAGCGYAVVCFW
<b>Human Defensins</b>	hNP-4	VCSRLVFCRRTEL RVGNLIGGVSFYCCTRV
	hBD-3	GIINTLQKYYCRVGGRC AVL SCLPKEEQIGKCTRGRKCCRRKK
<b>Fish and Crabs</b>	TachyplesinI	KWCFRVCYRGICYRRCR
	PolyphemusinI	RRWCFRVCYRGFCYRKKCR
	Cicadin	NEYHGFVDKANNENKRRKKQQGRDDFVVKPNNFANRRRKKDDYNNYYDDVDAAADV
<b>Vertebrates</b>	RTD-2	RCLCRRGVCRCLCRRGV
	BMAP-27	GRFKRFRKKFKKLFKLLSPVPLLLHLG
	Beta-basrubin	KIMAKPSKFYEQLRGR
<b>Plants</b>	Coccinin	KQTENLADTY
	Antifungalproteinfromcoconut	EQCREEEEDDR
	Thaumatococin-like protein	AKITFTNNHPRTIWP
	circulinB	GVIPCGESC VFIPICISTLLGCSCKNKVCYRN

	CirculinC	GIPCGESCVPFIPCITSVAGCSCKSKVCYRN
	CycloviolacinO14	GSIPACGESCFKGYTPGCSCSKYPLCAKN
	CycloviolacinY5	GIPCAESCVPWIPCTVTALVGCSCSDKVCYN
	CycloviolinD	GFPCGESCVFIPCISAAIGCSCKNKVCYRN
	KalataB8	GSVLNCGETCLLGTCTYTTGCTCNKYRVCTKD
	Lunatusin	KTCENLADTFRGPCFATSNC



### 2.3.5. Independent testing, performance measure and discovery of putative anti-HIV AMPs

#### 2.3.5.1. Independent testing results

The profile query step was to determine if the build profiles were of good quality and able to recognize peptides of similar properties within genome sequences which had the same pattern as the constructed models. The results of the independent testing of the HMMER models are shown in **Table 2.7**. Figures generated by querying the HMMER models using the testing data sets are presented in **Appendix A, Supplementary Figure A.1, Figure A.2, Figure A.3, Figure A.4, Figure A.5, Figure A.6, Figure A.7**.

The query of the HMMER profiles against specific super-families using the 596 non-anti-HIV AMP sequences are shown in **Table 2.7**.

On the other hand, the profiles query of the GLAM2 models during the independent testing of these models using the testing data set are represented in **Table 2.8**. Figures generated by querying the GLAM2 models using the testing data sets are presented in **Appendix A, Supplementary Figure A.8, Figure A.9, Figure A.10, Figure A.11, Figure A.12, Figure A.13, Figure A.14**.

The query of the GLAM2 profiles against specific super-families using the 596 non-anti-HIV AMP sequences are shown in **Table 2.8**.

The results showed that there were inconsistencies with the two tools. The number of true positive anti-HIV AMPs is almost identical for the models constructed by HMMER and GLAM2. However, the query of the profiles with the negative AMPs showed a huge difference in numbers between the HMMER models and the GLAM2 models.

The differences in number of false positive sequences between the tools after the scanning of the negative sets against the created models showed that HMMER is more specific than GLAM2.

**Table 2.7:** Results obtained by querying the HMMER profile

Families or AMPs profile	True Positive (TN)	False Negative (FN)	True Negative (TN)	False Positive (FP)
Amphibians	3	2	577	19
Microorganisms	1	0	596	0
Human Defensins	2	0	587	9
Fish and Crabs	2	0	594	2
Insects	0	1	593	3
Vertebrates	1	1	590	6
Plants	6	5	565	31

**Table 2.8:** Results obtained by querying the GLAM2 profile

Super-families	True Positive (TN)	False Negative (FN)	True Negative (TN)	False Positive (FP)
Amphibians	4	1	263	333
Microorganisms	1	0	561	35
Human Defensins	2	0	471	125
Fish and Crabs	0	2	457	139
Insects	0	1	265	331
Vertebrates	1	1	381	215
Plants	6	5	564	32



### 2.3.5.2. Performance measurement

The qualities of both the HMMER models and the GLAM2 models were assessed by calculating their performances. Since certain parameters such as the TP, TN, FP and FN numbers were obtained by querying the various profiles with the testing and the negative sets, these parameters were used to calculate the sensitivity, specificity, accuracy and Matthew's Correlation Coefficient (MCC) of the models as represented in **Table 2.9** for the HMMER models and **Table 2.10** for the GLAM2 models.

**Table 2.9:** Performance measurements generated for each super-family using the Model created by HMMER profile

Super-families	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
Amphibians	60	96.8	96.5	0.27
Microorganisms	100	100	100	1
Human Defensins	100	98.48	98.49	0.42
Fish and Crabs	100	99.66	99.66	0.71
Insects	0	99.5	99.33	-2.9e-3
Vertebrates	50	98.99	98.82	0.26
Plants	54.54	94.79	94.06	0.28

**Table 2.10:** Performance measurements generated for each super-family using the Model created by GLAM2 profile

Super-families	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
Amphibians	80	44.13	44.13	0.044
Microorganisms	100	94.13	94.13	0.16
Human Defensins	100	79	79.10	0.11
Fish and Crabs	0	76.68	76.41	-0.032
Insects	0	44.45	44.40	-0.046
Vertebrates	50	63.93	63.88	0.016
Plants	54.54	94.63	93.90	0.27

### 2.3.5.3. Databases query and discovery of putative anti-HIV AMPs

After testing and confirming the accuracy of each model, having calculated their performances, the profiles were used to scan various genome sequence databases. As such, the scanning of the databases, using the profiles generated by HMMER and GLAM2, was conducted in order to identify AMPs which can be utilised for the diagnosis and treatment of HIV/AIDS.

#### a) Using the Hidden Markov Models profiles

The scanning of the various genome sequence databases with the super-families specific to the HMMER profiles identified a number of peptides. The results of the genome sequences scan for the HMMER model are represented in **Table 2.11**.

**Table 2.11:** Name and number of genome sequences predicted to have anti-HIV AMPs after scanning the genome sequence databases using the HMMER Model

Family name	Species name	Number of anti-HIV AMPs
<b>Amphibians</b>	<i>Amolops jingdongensis</i> , <i>Amolops lifanensis</i> , <i>Amolops loloensis</i> , <i>Amolops mantzorum</i> , <i>Litoria caerulea</i> , <i>Phyllomedusa sauvagei</i> , <i>Rana amurensis</i> , <i>Rana chensinensis</i> , <i>Rana ornativentris</i> ( <b>Species names from the download amphibians sequences in UniprotKB</b> )	19
<b>Human Defensins</b>	<i>Homo sapiens</i>	6
<b>Insects</b>	<i>Bombyx mori</i>	9
	<i>Danaus plexippus</i>	6
	<i>Heliconius melpomene</i>	2
<b>Vertebrates</b>	<i>Ictidomys tridecemlineatus</i>	3
	<i>Taeniopygia guttata</i>	1
	<i>Pelodiscus sinensis</i>	1
<b>Plants</b>	<i>Zea mays</i>	2
	<i>Setaria italica</i>	4

The sequences of putative AMPs identified by the HMMER models were composed of single and multiple domains. The total number of identified single domain peptides was 48 (**Appendix A, Supplementary Figure A.15, Figure A.16, Figure A.17, Figure A.18, Figure A.19, Figure A.20, Figure A.21; Appendix A, Supplementary Table A.2**), and after removal of duplicate sequences from this list, the final list contained 30 putative anti-HIV AMPs which are presented in **Table 2.12** with the name of the species, the gene name, the putative anti-HIV AMP sequence and their ranking according to the E-values.

The E-values and the scores of the different domains of the putative anti-HIV AMPs multiple domains identified by the genome scanning of the vertebrates model are presented in the **Appendix A, Supplementary Figure A.22, Figure A.23, Figure A.24**.



**Table 2.12:** Final list of all putative anti-HIV AMPs after removal of duplicate sequences and their classification according to the E-values

Name of the Species	Gene name	Sequence of putative AMPs	E-values
<i>Homo sapiens</i>	ENSP00000342082	CLRYKKPECQSDWQCQPGKKRCCPDTCGIKCLDPDTPNPTRRKPGKCPVTYGQCLMLNPPNFCMDGQCKRDLKCCMGM	1.40E-54
<i>Danaus plexippus</i>	EHJ64257	KWKIFKKIEKVRNVRDGIKAGPAVQVVGGQATSIK	3.10E-15
<i>Bombyx mori</i>	BGJBMGA006280-TA	RWKLFFKKIEKVRNVRDGLIKAGPAIAVIGQAKSLGK	1.10E-13
<i>Danaus plexippus</i>	EHJ64256	KWKFFKKIEKVRNVRDGIKAGPAVQVLGEAKAIGK	7.30E-13
<i>Danaus plexippus</i>	EHJ71827	RWKLFFKKIEKVRNVRDGIKAGPAVGVVGGQATSIYK	3.00E-12
<i>Danaus plexippus</i>	EHJ68082	KWKPFKKLEKIGORVRDGIKAGPAVQVVGEAAIILK	2.90E-11
<i>Bombyx mori</i>	BGJBMGA000024-TA	RWKIFKKIEKVRNVRDGIKAGPAIEVLGSAKAIGK	2.70E-10
<i>Homo sapiens</i>	ENSP00000303532	CLKSGAICHVPFCPRR YKQJGTGGLPGTKCCKKP	3.50E-10
<i>Setaria italica</i>	Si023829m	TPCGESCLIPCITAAIGCSCKDKVCY	3.30E-08
<i>Heliconius melpomene</i>	HMEL010650-PA	WNPFKELEKAGQVRDatisAKFAVDVVGGQATAIJK	8.60E-08
<i>Setaria italica</i>	Si024202m	IPCGESCFILPCVTTAAIGSCQDRVCY	9.80E-08
<i>Litoria caerulea</i>	trfQ800R8 Q800R8_LITCE	GLFGILGSVAKHVLPHVVPVIAEH	2.40E-07
<i>Setaria italica</i>	Si023827m	ISCGTCLMLPCEHQAIGCRCKNKICY	3.40E-06
<i>Heliconius melpomene</i>	HMEL008612-PA	RWKFWKKEVEHAGQNRDGIKAGPAVAVKCGKVSFVRLIR	4.40E-06
<i>Litoria caerulea</i>	trfQ800R9 Q800R9_LITCE	GLFSVLGSVAKHVVPVVPVIAEH	5.00E-06
<i>Danaus plexippus</i>	EHJ68081	WNPFKELEKAGQVRDAISAAPAVEVVGQASSILK	6.30E-06
<i>Homo sapiens</i>	ENSP00000296435	CDKDNKRFAILLGDFFRKSKEKIGKEFKRIVQRIKDFLRNL	6.80E-06
<i>Homo sapiens</i>	ENSP00000297439	CVSSGGQLYSACPIFTKIQTGTCYRGKAKCCK	4.20E-05
<i>Setaria italica</i>	Si023843m	IQCGQTCFWIPCLDLGCSCKDNICY	4.40E-05

<i>Bombix mori</i>	BG BMGA000020-TA	KRKVFKIIIEKIGRNVGGVITAGPAVVVVVQAAASVGM	6.80E-05
<i>Zea mays</i>	GRMZM2G032198_P01	ISGESCVIIPCVCSTLLGRCENKLCV	0.0002
<i>Zea mays</i>	GRMZM2G374405_P01	VPCFESCVPVPCISSVVGRCENNVCV	0.00065
<i>Amolops jingdongensis</i>	tr G3ETP8 G3ETP8_AMOJI	GLFSIFKTAAKFVGKNNLLKQAGK	1.30E-003
<i>Amolops jingdongensis</i>	tr G3ETP4 G3ETP4_AMOJI	GIFSLFKTAAKFGKNNLLKEAGK	1.60E-003
<i>Amolops mantz-orum</i>	tr E1B242 E1B242_9NEOB	GIFSLIKTAAKFGKNNLLKQAGK	2.40E-003
<i>Rana ornativentris</i>	tr D1MYB8 D1MYB8_9NEOB	GLFNVFKGALKTAGKHVAGSLLNQ	5.00E-003
<i>Rana chersinensis</i>	tr F1AEM1 F1AEM1_9NEOB	GLLSVFKGVLLKTAGKNVAKNVAGS	6.30E-003
<i>Phyllomedusa sauvogeti</i>	tr Q1EN15 Q1EN15_PHYSA	GLRSKIWLWVLLMIWQESNKFKKM	7.30E-003
<i>Rana amurensis</i>	tr A0AAR9 A0AAR9_RANAM	GLLSVFKGVLLKGVGKNVAGSLLDQ	9.30E-003
<i>Rana amurensis</i>	tr A0AAS0 A0AAS0_RANAM	GLFSVVKGVLLKGVGKNVAGSLLDQ	9.40E-003



## b) Using the Gap Local Alignment of Motifs 2 profiles

The scanning of the various genome sequence databases with the respective super-families created by the GLAM2 software was also conducted. Besides the fact that all the genome sequences produced putative anti-HIV AMPs, the scanning of each genome sequence presented a limitation of 25 putative anti-HIV AMPs. Thus, about 25175 putative anti-HIV AMPs were obtained after the scanning of the genomes of the various species (**Table 2.13**). Common peptides sequences were found within the results obtained from the two tools. **Table 2.13** presents the number of putative anti-HIV AMPs observed for each tool and the number of common putative anti-HIV AMPs found within the two tools.

**Table 2.13:** Number of putative anti-HIV AMPs obtained per model for both tools and the common putative anti-HIV AMPs found within the two tools

Super-families	Number of putative anti-HIV AMPs obtained with HMMER Models	Number of putative anti-HIV AMPs obtained with GLAM2 Models	Common putative anti-HIV within the two tools
Amphibians	19	50	10
Microorganisms	0	22450	0
Human Defensins	6	25	0
Fish and Crabs	0	225	0
Insects	17	750	0
Vertebrates	5	1100	3
Plants	6	575	6

## 2.4. Discussion

The use of Antimicrobial Peptides as alternative sources for drug design has encouraged a massive explosion in the research area of these biomolecules. Many scientific approaches have been used to identify and comprehend their mechanism of action and to predict their activities. These AMPs exhibit certain characteristics and properties namely net positive charges, hydrophobicity, high specificity towards microorganisms and low microbial resistance which is exploited in research and their use as novel drug compounds.

Many of these peptides have been proven to have activity towards various gram-positive and gram-negative bacteria, protozoa, cancer, fungi as well as viruses, and some active Antimicrobial Peptides have advanced into clinical trial (Wang and Wang, 2004; Mulvenna *et al.*, 2006; Wang *et al.*, 2008; Wang *et al.*, 2009; Thomas *et al.*, 2010; Fjell *et al.*, 2012). Some of these AMPs have been shown to possess anti-HIV activity, with the most potent anti-HIV to date named Kn2-7 (Chen *et al.*, 2012). However, AMPs are also expressed in human and have shown to have anti-HIV activity. Examples include members of the defensin peptide family which have been shown to possess anti-HIV activity by inhibiting the interaction of the HIV with the target cell (Sun *et al.*, 2005). Also, the retrocyclins, RTD1, RTD2 and RTD3 can inhibit the entry of the virus into the T cells (Munk *et al.*, 2003; Wang *et al.*, 2004).

Different approaches have been used to identify and validate the activity of these peptides, either using molecular techniques or computational tools. Molecular methodologies were used as early tools for Antimicrobial Peptides identification, but due to the race between scientific research and technology, dramatic advances have been made in scientific approaches. This race and progression of scientific technologies have brought about the implementation of computational biology in various aspects of research (Hogeweg, 2011;

Ouzounis, 2012). These computational approaches have facilitated easy identification of Antimicrobial Peptides because they are less time consuming, less laborious and affordable.

The use of computational biological knowledge and the available model construction algorithms such as SVM, ANN, SW, LD, SW have shown to have some limitations. These shortcomings are because the models ought to be constructed in combination with tools that calculate structure-activity relationship information in order to enhance their strength and performances. The HMMER algorithm has been used in a previous study to classify and search for new AMPs (Brahmachary *et al.*, 2004; Fjell *et al.*, 2007). The HMMER and GLAM2 were utilised to identify putative anti-HIV AMPs in this study due to the fact that they use the sequences of the experimentally validated AMPs and take into consideration the individual motifs of the amino acids within sequences as features for model construction.

As part of model creation, the peptide activities were verified through literature mining after retrieval from the curated databases. This was done to confirm the anti-HIV activity of these peptides. This is an essential step as it aids in the construction of a model of a particular family or super-family with specific activity. These models would be useful in the identification of putative Antimicrobial Peptides having the same activity as the input sequences used in model creation.

The performance indications of the respective algorithms will be individually discussed in each subsequent paragraph. Performance indications of the respective models can be achieved by calculating the sensitivity, accuracy, specificity and the MCC. In this section, the sensitivity calculated for each model is discussed. It was observed that both tools have near identical sensitivity scores. Whilst 60%, 100%, 100%, 100%, 0%, 50% and 54.54% were obtained for the Amphibian, Microorganism, Human defensin, Fish and Crab, Insect, Vertebrate and Plant super-families respectively using HMMER models, 80%, 100%, 100%,



0%, 0%, 50% and 54.54% were obtained for for the Amphibian, Microorganism, Human defensin, Fish and Crab, Insect, Vertebrate and Plant super-families respectively with GLAM2.

The difference in the performances of both tools is due to the calibration of the models constructed by HMMER algorithm which was not done for GLAM2 models. Although it is said that the calibration step in HMMER increases the sensitivity of the models during the genome sequence scanning of databases (Eddy, 1998; Eddy, 2003), the performance of the calibrated model is only rated on its accuracy and specificity performances. All the accuracies of the HMMER models are above 95% except for the Plants super-family which is 94.79%. This proves that the models have more than 95% confidence to predict a peptide as a putative anti-HIV AMP. On the other hand, only the Microorganism and Plant super-families showed at least 90% confidence for GLAM2. Thus the models of GLAM2 algorithm will be insensitive to predict the above anti-HIV AMP.

Whilst the MCC ranges from -1 to +1, MCC equal to zero indicates a completely random prediction, MCC equal to 1 indicates perfect prediction and MCC equal to -1 indicated a total disagreement between the prediction and the observation (Baldi *et al.*, 2000). It was also observed that the MCC of the individual super-family constructed by both tools varied. In general, the MCC of HMMER super-family are greater than that of the GLAM2 super-family (**Table 2.9 and Table 2.10**).

The performance of the models is calculated based on the interrogation of these models with the testing sets of the corresponding models and a negative set (non-anti-HIV AMPs). The models built by using HMMER were sensitive to the testing sets and the negative set. Conversely, the models built by using the GLAM2 algorithm were less sensitive to the testing and the negative set. Therefore, the predicted false negative anti-HIV AMPs were less for

HMMER models than GLAM2 models whilst the predicted true negative for HMMER models were higher and the predicted true negatives for GLAM2 models were low. Thus, minimizing the discriminatory power of the GLAM2 algorithm. Consequently, high performances were observed for HMMER models as compare to the GLAM2 models.

The calibration of HMMER models also played an important role in the scanning of genome sequence databases. It was observed that the number of putative anti-HIV AMPs predicted by HMMER were not random. This was due to the high sensitivity of the HMMER models and the use of a low E-value cut-off during the scanning. Besides the sensitivity enhanced by the calibration of the models, the scoring system which combines a score and the E-value of the predicted AMP are better explained in HMMER (Eddy, 1998; Brahmachary *et al.*, 2004; Fjell *et al.*, 2007). This E-value gives more information about the probability of that predicted peptide to be a true anti-HIV AMP.

While the profile HMMER is based on statistical evaluation and probability assignment of amino acid position of the sequences, the GLAM2 tool is only based on statistical evaluation of the interrogated sequences thus their results presentations differ.

The HMMER results of the “profile query” and “database query” steps resulted in a ranked list of the best scoring domains in the order of occurrence of the sequence and alignments for the highest scoring domains. The matches of the query profiles against the genome sequences are shown with scores (bits) and E-values. The E-value, which is calculated from the bits score, shows the number of false positives that is expected to be seen at or above this bit score. Therefore an E-value of 0.01 and 0.05 indicates that there is only a 1% and 5% chance respectively that the hit is false or has come up by chance. Hence, a low E-value is considered appropriate with the lowest E-value appearing at the top of the result list. The high conserved residues for both the query sequence and the consensus pattern of the profile of a

family are shown in capital letters for the results. The result annotation of the “Query profile” is also represented as that of the “Query db” but contrary to the “Query profile”, the “Query db” searches for putative anti-HIV peptides can be added into AMP libraries to create further profiles.

After the scanning of genome sequence databases to identify putative anti-HIV AMPs, an observation was made that the peptides belonging to the Amphibian, Human Defensin, Insect and Plant super-families, were single domain peptides (**Appendix A, Figure A.15, Figure A.16, Figure A.17, Figure A.18, Figure A.19, Figure A.20, Figure A.21**). However, the putative anti-HIV AMPs identified using the Vertebrate super-family model, were multiple domain peptides i.e. some sections (domains) of the proteins were predicted to have anti-HIV activity but not the entire protein sequence (**Appendix A, Figure A.22, Figure A.23, Figure A.24**). Due to the fact that most active Antimicrobial Peptides have sequences which range from 10 to 100 amino acids in length (Hancock and Sahl, 2006), AMPs with single domains from the list of identified putative anti-HIV AMPs were considered for the continuation of this project, in the 3-D structure determination and docking studies. Additionally, the individual domains within the multiple domain proteins were predicted as putative anti-HIV AMPs, with E-values which were higher than the cut-off E-values, which was set at 0.01.

The GLAM2 result representation just gives the sequences that match the created profiles; the starting position and the end position of the matched sequences; and the score of the matched sequences. No E-values is attached to these results so as to better describe the probability that the peptides are false positives or truly predicted as putative anti-HIV AMPs. However, the “Query profile” helped to calculate the performance of GLAM2 in order to assess its overall quality.

The list of the putative anti-HIV AMPs generated by HMMER super-families specific models truly shows the diversity of the Antimicrobial Peptides. Different AMPs from different species were predicted to have the same activity on a specific pathogen. Whilst the best anti-HIV AMPs have an E-value of  $1.4e-54$ , meaning that there is only a  $1.4e-54\%$  chance for the peptide to be a false predicted anti-HIV AMP, the lowest score observed has an E-value of  $9.4e-3$ , meaning that there is only a  $9.4e-3\%$  chance for the peptide to be predicted as a false anti-HIV AMP. Hence, all the putative anti-HIV AMPs predicted by HMMER had higher probability scores to be considered true anti-HIV AMPs.

Conversely, the results given by GLAM2 models predicted 25 putative anti-HIV AMPs per genome scan. This can be due to the fact that the models were not calibrated since the GLAM2 algorithm does not allow this step in its program. Also, it is not possible to set a cut-off value with GLAM2 algorithm so as to eliminate non-specific peptide identification within the genome sequence. In addition, no statistical explanations of the results are provided, but just a score which gives no information about the percentage of a predicted peptide to be a false positive (Frith *et al.*, 2008). Thus, the results generated by the GLAM2 algorithm could not give the same confidence scoring for that the peptides predicted and their potential as anti-HIV AMPs.

The results given by the GLAM2 algorithm proved difficult to interpret and to classify the peptides. This is due to the fact that the query result of the algorithm only provides the sequence starting amino acid and ending amino acid, and also gives a score number without an E-value to explain the probability of the peptide being a putative anti-HIV AMP (Frith *et al.*, 2008). Additionally, there is no cut-off value or parameter to restrict the quality of the peptides predicted by the GLAM2. Due to this problem, the results obtained during the genome scan of the GLAM2 models were not considered for the rest of the study.

The assurance that the HMMER algorithm gives a more accurate prediction with a higher performance than the GLAM2 algorithm makes the number of predicted anti-HIV AMPs irrelevant but rather gives an indication of the potential of a predicted peptide to have anti-HIV activity.

Though the number of experimentally validated anti-HIV AMPs found in the repository databases was not big enough, future work will be to increase the samples size of the experimentally validated anti-HIV AMPs which will be used as the training and testing sets, so as to have more robust models of the various super-families.

## 2.5. Conclusion

In order to create models to identify putative Antimicrobial Peptides, experimentally validated peptides were required. These experimentally validated anti-HIV AMPs were used to create the models that scanned for putative anti-HIV AMPs from genome sequence databases.

The super-family specific models were successfully constructed based on the HMMER and GLAM2 algorithms. Whilst the performances of the models built using HMMER showed higher accuracies, the models built using GLAM2 algorithm showed the lowest accuracies. This meant that the HMMER algorithm is a better tool for the construction of models based on the sequences of AMPs and taking into consideration the motifs within these amino acid sequences. Additionally, HMMER algorithm is an appropriate tool to better interpret the results of the genome scanning, as it allows the user to choose a cut-off value with a good confidence interval. Also, it gives the probability of the predicted peptide to be a false positive based on the E-value obtained. The GLAM2 algorithm although suffering from

several limitations, should not be ignored. The tool has predicted some peptides which were also predicted by the HMMER algorithm.

The strength of HMMER is its performance and the detail with which the algorithm score the predicted putative anti-HIV AMPs, justifying the use of these AMPs for the next stage of the project. The study will be continued with the interaction prediction between the putative anti-HIV AMPs identify by HMMER and the HIV proteins gp120, gp41, p24 and p17.



## Chapter 3:

# 3-D structure prediction and anti-HIV AMPs interaction study

### 3.1. Introduction

Antimicrobial Peptides are natural occurring biomolecules made up of proteins and non-protein amino acids which contributes to the activity of this class of peptides. These biomolecules act on various pathogens and microbes and have shown to have activity against gram-positive and gram-negative bacteria, protozoa, fungi, cancer cells as well as viruses (Powers and Hancock, 2003). The interaction of the Antimicrobial Peptide with the microbes' membrane is a consequence of the positive charge of the AMP and the negative charge of the microbe. The result of which is the electrostatic attraction between the Antimicrobial Peptide and the microbe (Zasloff, 2002).

However, the activity, function and the mode of action of these biomolecules towards microorganisms are not restricted to their amino acids composition and peptide charge alone, but also to the peptides structure. These AMPs exhibit different classes of 3-D structures, which can be grouped into four major classes: the  $\beta$ -sheets,  $\alpha$ -helices, loop and extended peptides (Hancock and Lehrer, 1998).

The study of Antimicrobial Peptide 3-D structures and their interaction with the pathogens membrane protein's, could help resolve their mode of action and function. On the other hand,

most Antimicrobial Peptides do not have solved 3-D structures (Biaroch *et al.*, 2005). Whilst molecular methods including X-ray crystallography, NMR spectroscopy, dual polarisation interferometry, circular dichroism and cryo-electron microscopy, have been used to solve the 3-D structures of most AMPs, techniques such as affinity electrophoresis, phage display, Surface Plasmon Resonance (SPR), Isothermal Titration Calorimetry (ITC) and Fluorescence Resonance Energy Transfer (FRET) have been utilised for the interaction studies of these peptides to their target proteins. However, the molecular methods have shown to be time consuming, need specialised personnel and expensive equipment, hence the need for alternative approaches.

Computational methods can either be used for the prediction of a protein's 3-D structure or to study protein-protein interactions including cytosolic membrane bound proteins. Current methods for *in silico* prediction of proteins and peptides 3-D structure include (1) the “fold recognition and threading” methods, (2) the “integrative” or “hybrid” methods, (3) the “comparative” or “homology” modelling approach and (4) the “*de novo*” or “*ab initio*” methods (Schwede *et al.*, 2008).

The Iterative Threading ASSEMBLY Refinement (I-TASSER) server is an example of a *de novo* structure prediction method (Wu *et al.*, 2007; Zhang, 2008; Roy *et al.*, 2010). The *de novo* method uses the principles of physics that governs protein folding and/or using information derived from known structures but without relying on any evolutionary relationship to known folds. This method can be ideal for predicting the 3-D structure of an unknown protein sequence such as the putative anti-HIV AMPs. The modelling steps to predict the 3-D structure of the peptides and the protein molecules include multiple alignments of the target sequences, followed by iterative structural assembly simulations.



The predictions of the molecule 3-D structures are evaluated by examining the C-score, the TM-score and the Root Means Square Deviation (RMSD). The C-score given by the I-TASSER software is a confidence score for estimating the quality of predicted models and ranges from -5 to 2. The C-score is a scoring function based on the relative clustering structural density and the consensus significance score of multiple threading templates used to estimate the accuracy of the I-TASSER predictions. A C-score cutoff  $> -1.5$  indicates that the model has a correct fold. The TM-score is a scale for measuring the structural similarity between the predicted 3-D structure and the template structure. A TM-score  $> 0.5$  indicates a model of correct topology and a TM-score  $< 0.17$  means a random similarity (Roy *et al.*, 2010). The Root Means Square Deviation (RMSD) measures the distance between atoms of superimposed proteins hence there is a strong correlation between the TM-score and the RMSD. The number of Decoys represents the number of structural decoys that are used in generating each model and the cluster density is defined as the number of structure decoys at a unit of space in the SPICKER cluster (Zhang, 2008; Roy *et al.*, 2010).

Although the prediction of the AMPs 3-D structure from the amino acid sequence is a challenge, having a 3-D structure allows for the extrapolation on the mechanism (s) on how these peptides gain entry into the microbes' membrane and thus the function of these peptides (Schwede *et al.*, 2008).

The PatchDock server is an example of *in silico* prediction tool of protein-protein interactions, and it is based on a rigid-body geometric hashing algorithm, which works on the principle of molecular shape complementarities between the 3-D structures of the two proteins involved in the complex formation. The tool runs with high efficiency to perform fast transformations which is driven by local feature matching to search the six-dimensional transformational spaces created by the formation of the complex (Schneidman-Duhovny *et al.*, 2005).

An *in silico* approach comprising of the 3-D structure prediction of the putative anti-HIV AMPs followed by docking studies of these predicted structures against various HIV protein's will be carried out in this chapter to further characterize these AMPs. Thus, the aim of this chapter is to predict the best suitable interaction of several HIV proteins and the putative anti-HIV AMPs which could be used for the diagnostics and therapeutics of the HIV.

To achieve this, the 3-D structures of ten of the 30 putative anti-HIV AMPs listed in chapter 2 will be predicted using the algorithm I-TASSER and *in silico* docking will be performed, using PatchDock between the 3-D structures of the putative anti-HIV AMPs and those for the HIV proteins gp120, gp41, p24 and p17.

## **3.2. Materials and methods**

### **3.2.1. Selection of putative anti-HIV AMPs and HIV proteins**

The Antimicrobial Peptides used in this chapter are those which were identified by scanning genome sequences located within databases using HMMER. The final list of putative anti-HIV AMPs contained 30 peptide sequences. We selected the 10 top ranked putative Antimicrobial Peptides based on the lowest E-value scores, since a lower score meant these had the highest probability to be peptides with anti-HIV activity. A positive and negative control was also included into the 3-D modelling as well as the docking studies to compare the results obtained for the putative AMPs. The kn2-7 was selected as the positive control as it has been reported to be the most potent anti-HIV AMP to date. Mucroporin S1 was selected as the negative control since it was also used as the negative control in the study of kn2-7 to establish its anti-HIV activity and was inactive against HIV-1 pseudotyped virus (HIV-1 PV) (Chen *et al.*, 2012).

The HIV proteins gp120, gp41, p24 and p17 will be used for the docking study. The gp120 and gp41 was selected since the virus binds to the human CD4+ T cells with the help of the membrane surface protein molecule gp120 upon infection and, this allows entry of the virus into the T cells and the propagation to other host cells (Kwong *et al.*, 1998). The HIV glycoprotein gp41 favours the attachment of the virus to human T cells and injection of viral RNA into the human cells, after which infection occurs. Targeting the gp120 and gp41 proteins might be an additional way to prevent HIV propagation within the human T cells and consequently prevent a decrease in the T cell population.

The core proteins p24 and p17 were selected because the HIV capsid is composed of these proteins and they are released into the human blood stream during replication. These proteins serve as tools for early detection of the disease instead of using HIV-specific antibodies, which are expressed at the latter stage of the disease. Although p24 is routinely used in diagnostic systems, it has some shortcomings such as low sensitivity (Monaco-Malbet *et al.*, 2000; Sutthent *et al.*, 2003), emphasising the need for a more sensitive diagnostic tool based on the detection of the HIV associated proteins/particles. The sequences of the putative anti-HIV AMPs and HIV proteins used for the 3-D structure prediction and protein-protein interaction studies are showed in (**Supplementary materials B, Table B.1 and Table B.2**).

### **3.2.2. Physicochemical characterisation of the putative anti-HIV AMPs and the HIV proteins**

The aim was to characterise the predicted AMPs based on their physicochemical properties to ensure that they conform to known AMPs and specifically to AMPs with proven anti-HIV activity. Also, the physicochemical properties of the 4 HIV proteins were determined, to ascertain if the properties of these proteins will favour interaction with the putative AMPs. It

is a well known fact that Antimicrobial Peptides and microbes interact by electrostatic attraction and is a fundamental principle for the activity of AMPs. The calculation of several parameters such as: (i) the number of basic residues, (ii) acidic residues, (iii) net charge, (iv) the Isoelectric point, (v) the Boman Index (or protein binding potential), (vi) Hydrophobic residues, (vii) the instability index of the proteins (Wang and Wang, 2004); (viii) the number of Arginine (Arg) or Lysine (Lys) residues, (ix) the presence of Cysteine (Cys) residue (Wang *et al.*, 2010; Wang *et al.*, 2011) will be measured as they are fundamental to the activity of AMPs. Additional parameters for the prediction of a good anti-HIV AMPs interaction and activity may include the peptide length, sequence order and structure of the peptides (Wang *et al.*, 2010).

The following parameters of the putative anti-HIV AMPs and the HIV proteins were calculated: the charge, Boman index, the instability index and many more using the prediction interface of the Bactibase and Antimicrobial Peptides Database with the amino acid sequences of the peptides and the proteins as input (<http://bactibase.pfba-lab-tun.org/physicochem> and [http://aps.unmc.edu/AP/design/design\\_improve.php](http://aps.unmc.edu/AP/design/design_improve.php)) (Hammami *et al.*, 2007; Wang and Wang, 2009; Hammami *et al.*, 2010).

### **3.2.3. *De novo* structure predictions of the putative anti-HIV antimicrobial peptides and HIV proteins**

A number of computational algorithms have been developed to accommodate the high demand of protein 3-D structure determination and to predict their possible interaction with other proteins so as to assign functions to these proteins.

The prediction of the 10 putative anti-HIV AMPs and the HIV proteins 3-D structures were performed using I-TASSER (Iterative Threading ASSEMBLY Refinement) server which is an example of a *de novo* method of structure prediction (Schwede *et al.*, 2008). This method was chosen because the 10 putative anti-HIV AMP sequences were considered novel and thus had no homologous structures available for comparative modelling.

I-TASSER server is a free on-line tool which computationally predicts the 3-D structure of protein or peptide from its amino acid sequence. The server can be found at the URL (<http://zhanglab.ccmb.med.umich.edu/I-TASSER/>) and it is held at the University of Michigan, USA. The server implements various mathematical algorithms to predict 3-D structure of proteins and peptides.

The 3-D structures of the anti-HIV AMPs and HIV proteins were predicted by uploading each sequence onto the I-TASSER website. The user enters their email address to which the results link will be send. After, naming the uploaded sequence, the menu “Run I-TASSER” was selected.

The output from the I-TASSER server includes: a full-length secondary and tertiary structure prediction as well as functional annotations on ligand-binding sites, Enzyme Commission numbers, Gene Ontology terms and most importantly provides an estimate of accuracy scoring of the predicted peptides and the proteins 3-D structures based on the C-score, TM-score and RMSD (Roy *et al.*, 2010). The visualisations of the 3-D structures were done using the PyMOL 1.3. Software as PDB files.

In addition to the above results presented by I-TASSER server, a default superimposition result was given at the bottom of the result page. This superimposition result is a structural alignment and comparison of the predicted 3-D structures with known solved 3-D structures found in the Protein Data Bank (PDB). The website result page presented the PDB ID

templates which were used to create the superimpositions, TM-score and the RMSD and the percentage similarity with the template PDB of the determined 3-D structures with their template structures (Roy *et al.*, 2010).

#### **3.2.4. Docking and interaction analysis of the putative anti-HIV antimicrobial peptides and HIV proteins**

Docking is a method used to predict the most favourable orientation of protein-protein or protein-ligand complexes (Lengauer, and Rarey, 1996). The docking of the 10 putative anti-HIV AMPs to the HIV proteins gp120, gp41, p24 and p17 were done using PatchDock Beta 1.3 version. PatchDock is a free online web-server that allows for protein-protein and protein-small ligand molecule docking and is available at <http://bioinfo3d.cs.tau.ac.il/PatchDock/>.

Docking was done by uploading the respective PDB files of the HIV proteins and the putative anti-HIV AMPs onto the PatchDock server website, after which the user enters an email address. The cluster RMSD was set to 4.0 Å and the complex type was selected as “protein-small ligand”. The task was submitted by selecting “Submit Form”. The docking results are sent via an email notification, containing the web link to the docking results. The result provides the highest scoring complexes between the HIV protein and the anti-HIV AMP as a PDB output file (Schneidman-Duhovny *et al.*, 2005).

Besides the geometric scoring system given in the result section of PatchDock, additional information includes the Atom Contact Energy (ACE), the area covered between the two molecules, the transformation coordinates during the molecular interaction and the PDB file of the complex formed as a ball and stick structure (Schneidman-Duhovny *et al.*, 2005). Interaction analysis of the complex formation between the HIV protein and the putative anti-

HIV AMP was done using PyMOL 1.3. Software and the distance between the interacting residues calculated.

### **3.3. Results**

#### **3.3.1. Physicochemical characterisation of the putative anti-HIV AMPs and HIV proteins**

The physicochemical properties of the identified putative anti-HIV AMPs were studied as to ensure that these putative peptides shared similar features to all classes of AMPs and specifically to anti-HIV AMPs. The physicochemical properties and the parameters for the putative anti-HIV AMPs and HIV protein characterisations are presented in **Table 3.1** and **Table 3.2** respectively.



**Table 3.1:** Physicochemical properties and parameters for the 10 putative anti-HIV AMPs, the positive control (Kn2-7) and the negative control (Mucroporin S1)

	Mass	Most common amino acids and %	Lysine %	Arginine %	Cysteine %	Isoelectric point	Net charge	Total hydrophobic ratio	Instability Index	Protein-binding Potential (Boman Index)	Half Life in Mammals	Sequence similarity with other molecules and percentage
Molecule1	8903.716 Da	Cys: 16	11.39	6.33	16	8.37	6	34 %	43.71	2.17 kcal/mol	1.2 hour	SLPI: 68.22 %
Molecule2	4028.831 Da	Lys: 18.92	18.92	5.41	0.00	11.49	7	43 %	7.98	1.26 kcal/mol	1.3 hour	Cecropin A: 86.48 %
Molecule3	4040.889 Da	Lys: 18.92	18.92	8.11	0.00	11.86	8	43 %	27.80	1.37 kcal/mol	1 hour	Hyphancin IIIF: 81.08 %
Molecule4	4088.926 Da	Lys: 21.62	21.62	5.41	0.00	11.25	7	43 %	20.12	1.26 kcal/mol	1.3 hour	Cecropin B: 83.78 %
Molecule5	4077.906 Da	Lys: 18.92	18.92	8.11	0.00	11.48	8	40 %	23.94	1.39 kcal/mol	1 hour	Papiliocin: 76.92 %
Molecule6	4031.883 Da	Lys: 18.92	18.92	5.41	0.00	11.17	6	45 %	1.40	1.03 kcal/mol	1.3 hour	Cecropin A: 72.97 %
Molecule7	4073.94 Da	Lys: 18.92	18.92	8.11	0.00	11.46	7	43 %	61.88	1.45 kcal/mol	1 hour	Cecropin B: 94.59 %
Molecule8	3670.552 Da	Cys: 17.65	14.71	5.88	17.65	9.60	8	38 %	48.28	1.07 kcal/mol	1.2 hour	hBD2: 82.92 %
Molecule9	2780.401 Da	Cys: 22.22	7.41	0.00	22.22	6.03	0	51 %	32.31	-0.11 kcal/mol	7.2 hour	Cliotide T1: 76.66 %
Molecule10	3908.564 Da	Ala: 16.67	11.11	5.56	0.00	10.33	2	47 %	7.89	1.33 kcal/mol	2.8 hour	Cecropin D: 80.55 %
Kn2-7	1674.152 Da	Arg&Ile:23	15	23	0.00	12.81	5	61 %	36.06	1.8 kcal/mol	1.1 hour	Bmk2: 62.5 %
Mucroporin S1	1091.39 Da	Ser&Leu:18	9	0	0.00	9.70	1	54 %	6.65	-1.19 kcal/mol	1.9 hour	Mucroporin: 64.7 %

**Table 3.2:** Characterisation of the different physicochemical properties and parameters for the four HIV proteins

HIV proteins	Mass	Isoelectric point	Net charge	Total hydrophobic ratio	Instability Index	Half Life in Mammals
gp120	35098.29 Da	7.52	+6	37 %	41.74	1 hour
gp41	8357.17 Da	4.64	-2	38 %	66.08	2.8 hour
p24	23402.43 Da	5.69	0	39 %	44.98	100 hour
p17	14983.77 Da	9.46	+7	32 %	42.75	3.5 hour



### 3.3.2. Prediction of the putative anti-HIV AMPs and HIV proteins 3-D structures

The results for I-TASSER showed that the C-score of the 3 HIV proteins (p24, p17 and gp120) were 1.41, 1.70 and 2.00 respectively, whilst the C-score of gp41 was -0.17. Additionally, the TM-score of p24, p17, gp120 and gp41 were 0.91, 0.95, 0.99 and 0.69 respectively. All proteins had a TM-score above 0.5. The Root Means Square Deviation (RMSD) of p24, p17, gp120 and gp41 were 2.7, 1.4, 1.7 and 3.6 respectively. The RMSD of p17 and gp120 were less than 2Å. Although the RMSD is not less than 1Å, the topologies of the predicted 3-D structures are a consequence of them having a TM-score above 0.5, since there is a strong correlation between the RMSD and the TM-score of a predicted 3-D protein structure (Roy *et al.*, 2010).

The prediction of the putative anti-HIV AMP 3-D structures gave C-score values which ranged from -1.83 to 0.95. All the peptides had C-score values which were higher than -1.5, except for molecule 1, which had a C-score of -1.83. The TM-score of the 10 putative anti-HIV AMPs ranged from 0.49 to 0.84 and were all above 0.5 except again for molecule 1 which had a TM-score of 0.49. Whilst molecules 8 and 9, the positive and the negative controls had RMSD scores which were less than 1Å, AMP molecules 2, 3, 4, 5, 6, 7 and 10 predicted 3-D structures have RMSD scores higher than the positive and negative controls, ranging between 2.0 to 2.2Å. The RMSD value of molecule 1 was above 4Å and was reported to be 7.3Å.

The quality assessment scores obtained for the prediction of the 3-D structures for the HIV proteins and putative anti-HIV AMPs by the I-TASSER server are summarised in **Table 3.3** and **Table 3.4**.

**Table 3.3:** Quality assessment scores of the predicted 3-D structures of the putative anti-HIV AMPs, the positive and negative controls

Putative AMPs	C-score	Exp. TM score	Exp. RMSD (Å)	No of Decoys	Cluster Density
Molecule1	-1.83	0.49 ± 0.15	7.3 ± 4.2	4345	0.0784
Molecule2	-0.06	0.71 ± 0.12	2.2 ± 1.7	8654	0.3641
Molecule3	0.03	0.72 ± 0.11	2.0 ± 1.6	8718	0.3975
Molecule4	-0.05	0.71 ± 0.12	2.2 ± 1.7	8719	0.3653
Molecule5	-0.01	0.71 ± 0.11	2.1 ± 1.7	8890	0.3874
Molecule6	0.04	0.72 ± 0.11	2.0 ± 1.6	9191	0.4041
Molecule7	-0.00	0.72 ± 0.11	2.1 ± 1.7	9019	0.3844
Molecule8	0.95	0.84 ± 0.08	0.5 ± 0.5	10040	0.8868
Molecule9	0.73	0.81 ± 0.09	0.5 ± 0.5	10167	0.6645
Molecule10	0.06	0.72 ± 0.11	2.2 ± 1.7	7895	0.4324
+ve control (Kn2-7)	0.14	0.73 ± 0.11	0.5 ± 0.5	10200	1.2500
-ve control (MucroporinS1)	0.28	0.75 ± 0.10	0.5 ± 0.5	10199	1.2499

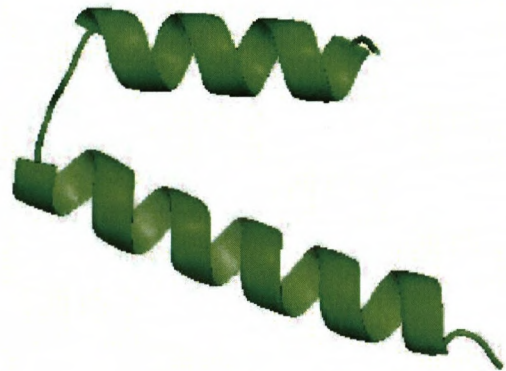
**Table 3.4:** Quality assessment scores of the predicted 3-D structures for the HIV proteins

HIV proteins	C-score	Exp. TM score	Exp. RMSD (Å)	No of Decoys	Cluster Density
p24	1.41	0.91 ± 0.06	2.7 ± 2.0	8301	0.8658
p17	1.70	0.95 ± 0.05	1.4 ± 1.3	10155	0.9134
gp120	2.00	0.99 ± 0.03	1.7 ± 1.5	8424	1.2500
gp41	-0.17	0.69 ± 0.12	3.6 ± 2.5	7065	0.3724

The anti-HIV AMPs represented different secondary structures. Molecule1 represented an extended or loop structure whilst molecules2, 3, 4, 5, 6, 7 and 10, the positive and negative each had  $\alpha$ -helical conformation. Molecule9 represented a very short  $\alpha$ -helical structure with some coil structure. Only molecule8 had an anti-parallel  $\beta$ -sheeted secondary structure arrangement, mix with a loop structure.



(A: Molecule1)

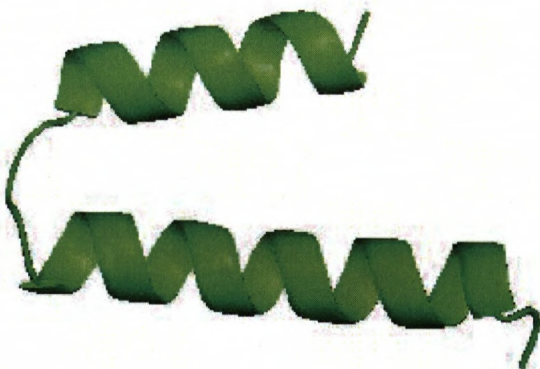


(B: Molecule2)

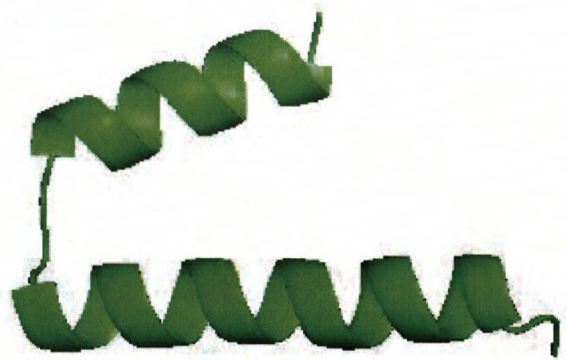


(C: Molecule3)

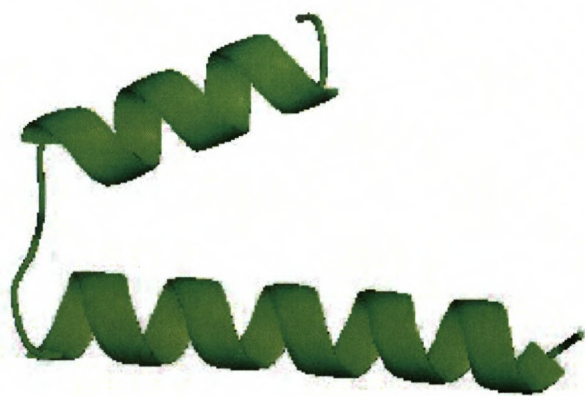
(D: Molecule4)



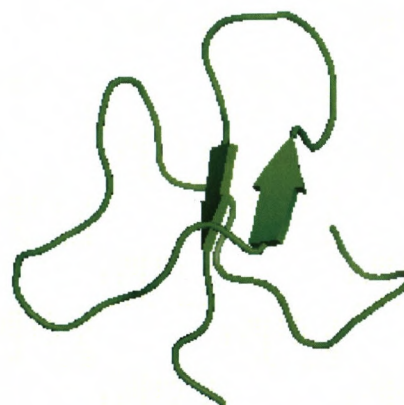
(E: Molecule5)



(F: Molecule6)



(G: Molecule7)



(H: Molecule8)



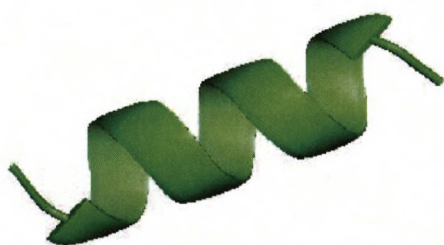
(I: Molecule9)



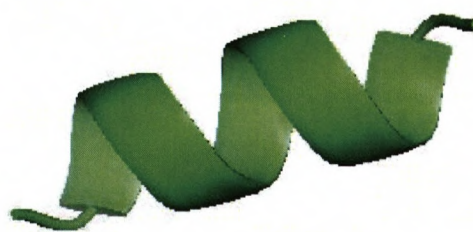
UNIVERSITY of the  
WESTERN CAPE



(J: Molecule10)

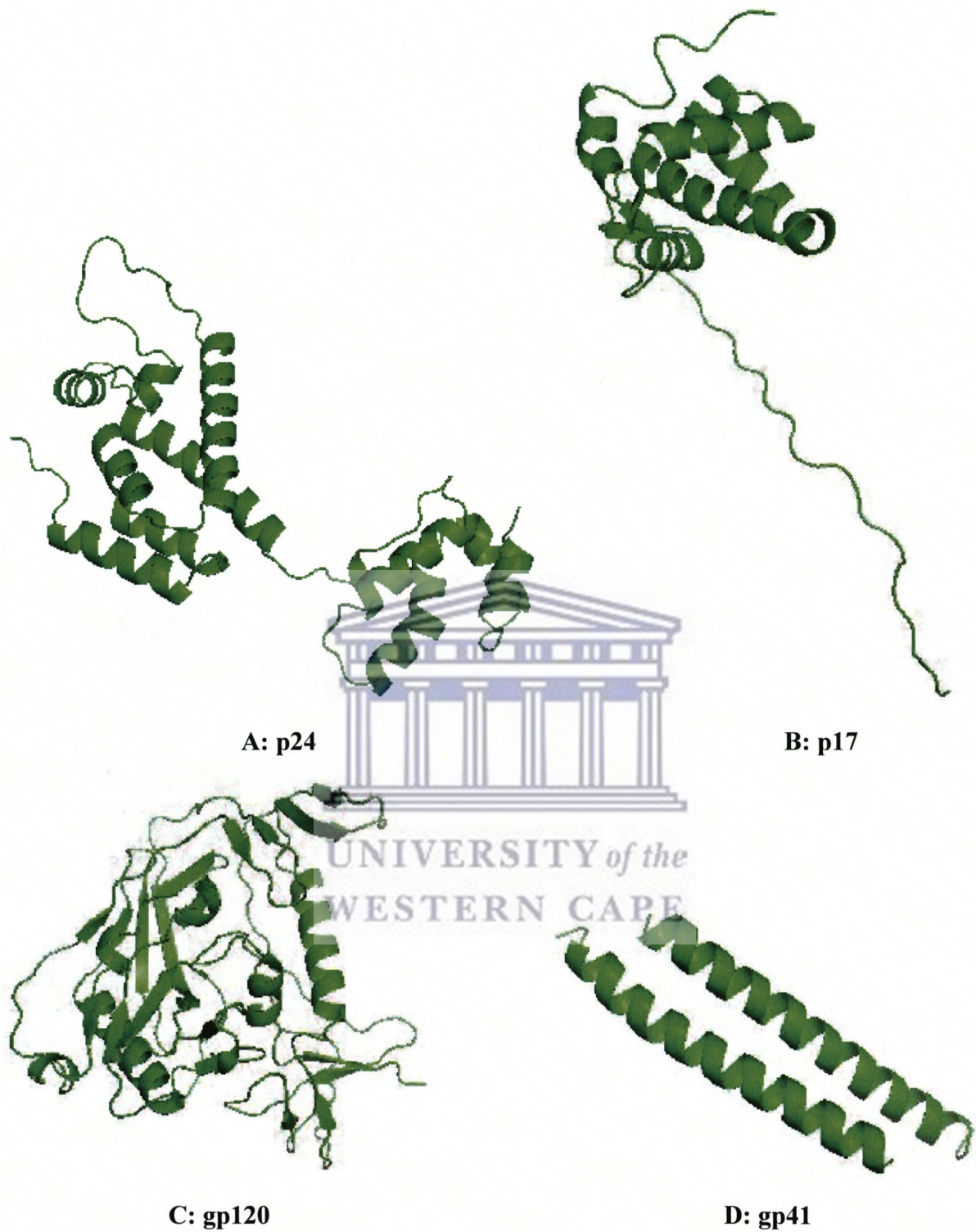


(K: kn2-7)



(L: Mucroporin S1)

**Figure 3.1:** Cartoon representations of the 3-D structures predicted for 10 putative anti-HIV AMPs by the I-TASSER server. A: molecule1, B: molecule2, C: molecule3, D: molecule4, E: molecule5, F: molecule6, G: molecule7, H: molecule8, I: molecule9, J: molecule10, K: Kn2-7 and L: Mucroporin S1 are represented by PyMOL 1.3. Software



**Figure 3.2:** Cartoon representations of the 3-D structures of the four HIV proteins predicted by the I-TASSER server. **A:** p24, **B:** p17, **C:** gp120 and **D:** gp41 are represented by PyMOL 1.3. Software

The HIV proteins presented different structures. The gp41 is formed by two monomers which has an  $\alpha$ -helix secondary structure. The gp120 protein is composed of a mixture of  $\alpha$ -helices, anti-parallel  $\beta$ -sheets and loop structures. The p24 proteins are composed of 12  $\alpha$ -helices and 12 loop structures and p17 protein is composed of 6  $\alpha$ -helices, 10 loop and 3  $\beta$ -sheet secondary structures. The cartoon representation of the 3-D models of the 10 putative anti-HIV AMPs and the HIV proteins are presented in **Figure 3.1 A-L** and **Figure 3.2 A-D**.

### **3.3.3. Superimposition of the predicted molecule 3-D structures with the known structures**

As part of the results output delivered by the I-TASSER server, the 3-D structural models of each molecule predicted was confirmed by superimposing the secondary structure of the predicted molecule to the closest similar protein 3-D structure found in the Protein Data Bank. The PDB ID of the known 3-D structures which were used for the superimposition was attached to the I-TASSER result page. The images of the superimposition are presented in (**Figure 3.3 A-L, Figure 3.4, Figure 3.5, Figure 3.6 and Figure 3.7**).

The superimposition is well presented with some statistical explanation for the accuracy of these structures. Besides molecule1 which had a TM-score less than 0.5, all the superimposed 3-D structures had TM-scores which was greater than 0.5, meaning that the structures had structural similarity with the templates which were used for their prediction. Also, all the 3-D structures except molecule1, had RMSDs less than 2Å, thus the superimposition of the 3-D structures and the templates are closely related and there is little atomic deviation between the predicted molecules and their templates.



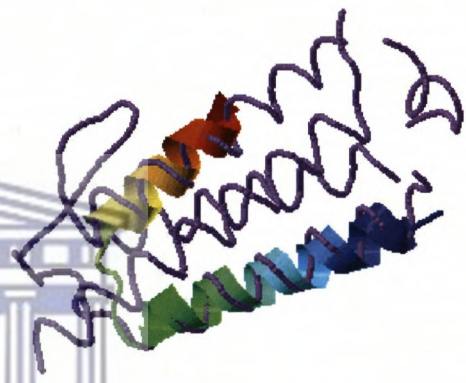
(A: Molecule1≈2afbB)



(B: Molecule2≈3kasA)



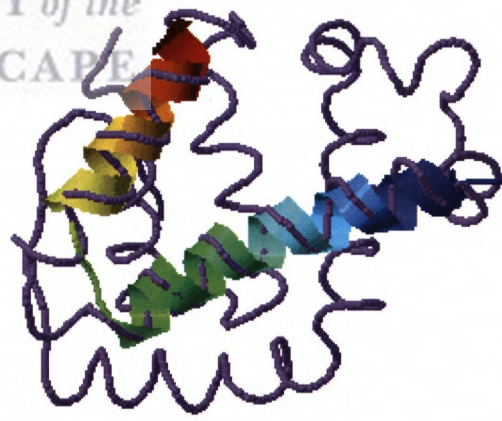
(C: Molecule3≈3s4wA2)



(D: Molecule4≈2yqyB)



(E: Molecule5≈2yqyB)



(F: Molecule6≈1uc3L)



UNIVERSITY of the  
WESTERN CAPE



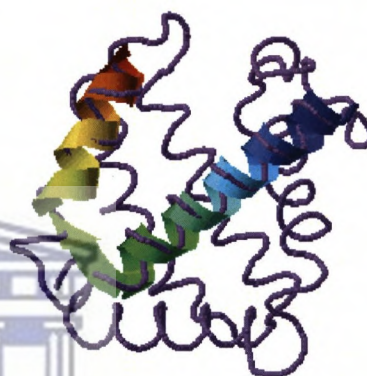
(G: Molecule7≈3lhbH)



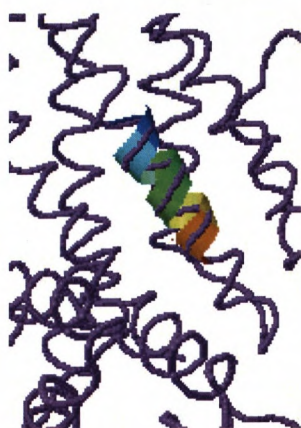
(H: Molecule8≈1fd3A)



(I: Molecule9≈1za8A)



(J: Molecule10≈2d2mD)



(K: kn2-7≈3s4wA)

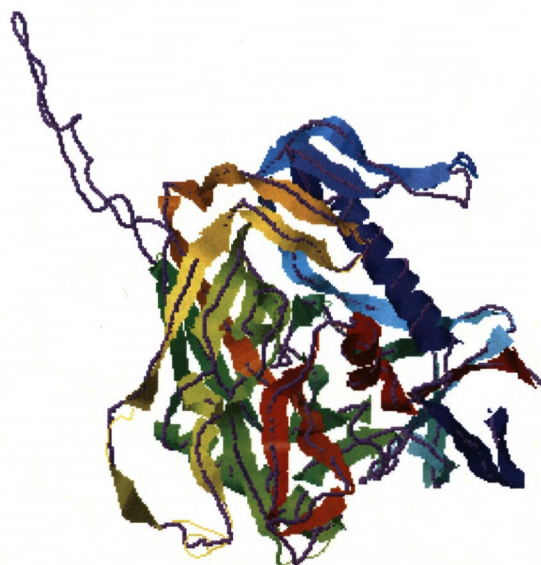


(L: Mucroporin S1≈3uboA)



**Figure 3.3:** Superimposition of each putative anti-HIV AMP, the positive and negative controls with the closest solve 3-D structure found in the Protein Data Bank. Each molecule is in parenthesis with the PDB ID from which the structure was superimposed and are closely related and are closely related and are closely related. The colour cartoon representation of the predicted 3-D structures while the line structure is the known solve structure used for the superimposition.

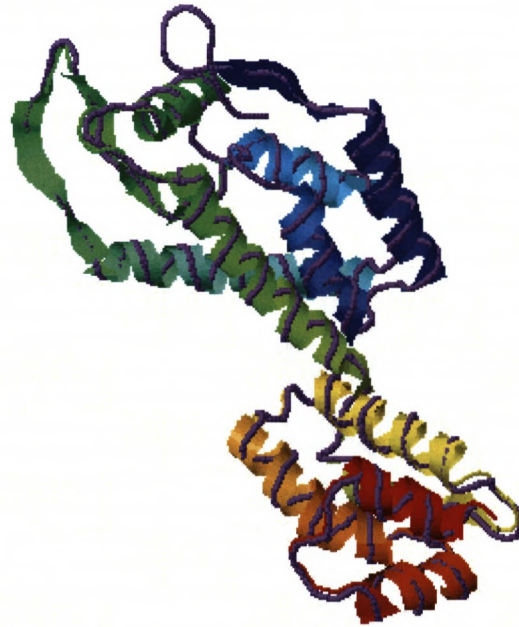




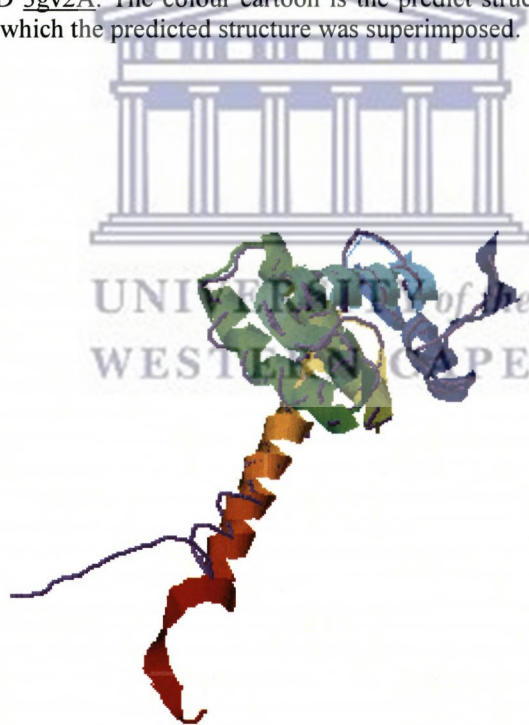
**Figure 3.4:** HIV protein gp120 3-D structure predicted by I-TASSER superimposed with the known solved 3-D structure having the PDB ID 2b4cG. The colour cartoon is the predicted structure while the line structure is the known solve structure from which the predicted structure was superimposed.



**Figure 3.5:** HIV protein gp41 3-D structure predicted by I-TASSER superimposed with the Known solved 3-D structure having the PDB ID 2cmrA. The colour cartoon is the predicted structure while the line structure is the known solve structure from which the predicted structure was superimposed.



**Figure 3.6:** HIV protein p24 3-D structure predicted by I-TASSER superimposed with the Known solved 3-D structure having the PDB ID [3gv2A](#). The colour cartoon is the predict structure while the line structure is the known solve structure from which the predicted structure was superimposed.



**Figure 3.7:** HIV protein p17 3-D structure predicted by I-TASSER superimposed with the Known solved 3-D structure having the PDB ID [116nA1](#). The colour cartoon is the predict structure while the line structure is the known solve structure from which the predicted structure was superimposed.

The PDB file ID of the known solved structure is also shown in **Table 3.5** on which the putative AMP predicted structures were based on. The statistical explanation of each superimpose molecule to the closest similar protein structure are represented in **Table 3.5**.

After the prediction of the 3-D structures, the putative anti-HIV AMPs, the positive control (kn2-7), the negative control (Mucroporin S1) were docked to the HIV proteins gp120, gp41, p24 and p17.

**Table 3.5:** Statistical explanation of the similarity between the predicted 3-D structure using the I-TASSER server and the 3-D structure of known protein found in the PDB. The TM score and the RMSD give information on how close they proteins are related structurally

	PDB ID of the superimposed template	TM-score	RMSD (Å)	% Identity with PDB ID template
Molecule1	2afbB	0.469	3.52	0.083
Molecule2	3kasA	0.775	1.77	0.027
Molecule3	3s4wA2	0.731	1.53	0.000
Molecule4	2yqyB	0.776	1.45	0.108
Molecule5	2yqyB	0.766	1.67	0.162
Molecule6	1uc3L	0.816	1.15	0.054
Molecule7	3lhbH	0.788	1.70	0.108
Molecule8	1fd3A	0.693	1.33	0.912
Molecule9	1za8A	0.568	1.05	0.720
Molecule10	2d2mD	0.798	1.05	0.083
Kn2-7	3s4wA	0.845	1.17	0.154
MucroporinS1	3uboA	0.888	1.60	0.091
gp120	2b4cG	0.967	0.84	0.868
gp41	2cmrA	0.867	1.30	1.000
p24	3gv2A	0.989	0.43	0.976
p17	1l6nA1	0.877	1.40	0.931

### 3.3.4. Protein-protein interaction study of the anti-HIV AMPs bound to the HIV proteins

The results from the PatchDock docking study indicated that all the putative anti-HIV AMPs as well as positive and the negative control, binds to the four HIV proteins gp120, gp41, p24 and p17. The geometric scores of the binding affinities of the putative anti-HIV AMPs and HIV proteins ranged from 14926 to 6710. It was observed that the 10 putative anti-HIV AMPs have higher binding affinity scores than the positive and negative controls. Most importantly, the 10 putative anti-HIV AMPs geometric scores are higher than the positive control kn2-7 that has been demonstrated to possess potent anti-HIV activity (Chen *et al.*, 2012). Molecule1 showed a very strong binding affinity geometric score to all the HIV proteins. This confirms the probability of this peptide to be a putative anti-HIV peptide as it also showed the lowest E-value prediction score and fulfils all the physicochemical property requirements of a good AMP. However, Molecule9 has the lowest binding affinity for all the HIV proteins.

The binding affinity of protein-protein interaction study of the 10 putative anti-HIV AMPs bound to the HIV proteins (gp120, gp41, p24 and p17) using PatchDock server results are recapitulated in **Table 3.6**.

**Table 3.6:** Geometric scores of the binding affinity obtained for docking the anti-HIV AMPs to HIV proteins.

AMPs	Binding affinity geometric scores			
	gp120	gp41	p24	p17
Molecule1	14926	13992	14708	14492
Molecule2	12650	9588	12114	11510
Molecule3	13686	9872	12310	11832
Molecule4	11968	9368	11188	11570
Molecule5	12708	9766	11974	12324
Molecule6	13104	10062	12534	12150
Molecule7	13648	10262	11930	11462
Molecule8	11086	8410	9418	12960
Molecule9	9186	7098	8618	9352
Molecule10	12208	10010	11560	11396
+ve control (Kn2-7)	8158	6082	8062	8732
-ve control (MucroporinS1)	6912	5066	7008	6710

In addition to the binding affinity geometric scores which are given in the result page of PatchDock, the area between the molecules forming the binding complexes, the atomic contact energy (ACE) and the transformations were also calculated. The results showed that the area covered by molecule1 has the largest surface coverage with gp120, gp41, p24 and p17. This is due to its large molecular weight (Table 3.1). In addition, the positive and negative controls had smaller surface coverage with the HIV proteins since both peptides have smaller sizes. This may also be due to the fact that their net charges are less than that of the putative anti-HIV AMPs. Also, the Boman index of the negative control is smaller than zero (Table 3.1). The data on the area, the ACE and the transformations are represented in the Appendix B, Table B.3 and Table B.4.

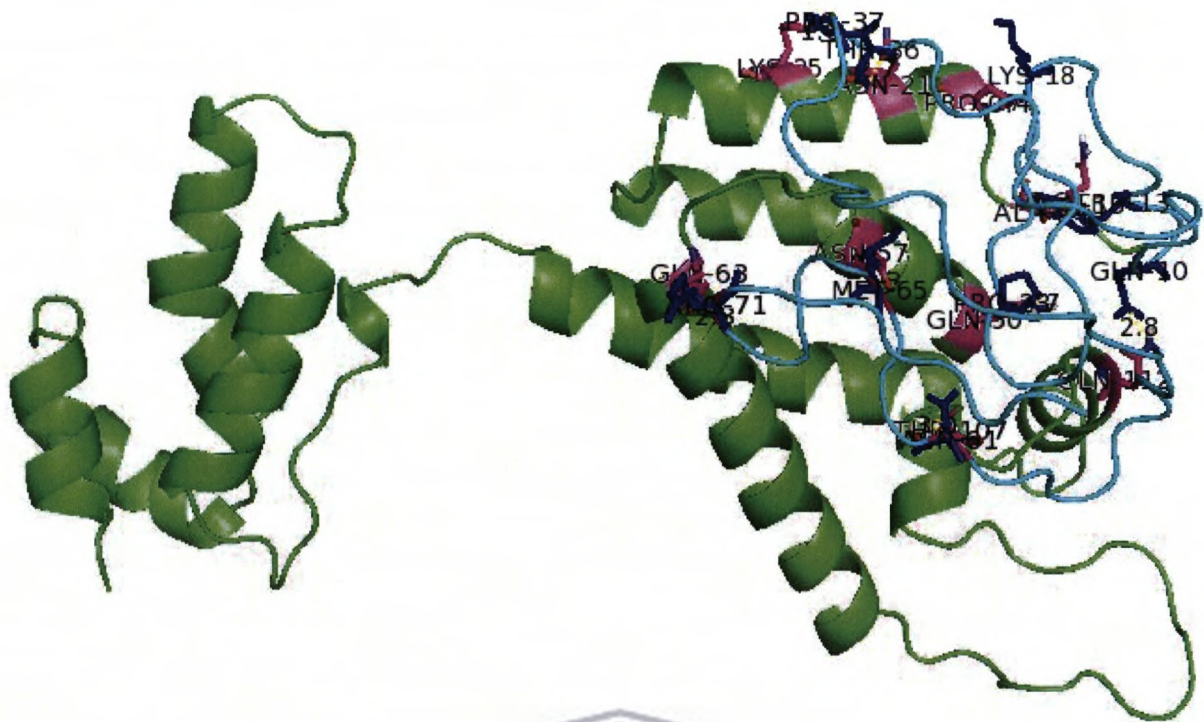
The main objective for studying the structure complex formation between the putative anti-HIV AMPs and the HIV proteins was not only to predict the interaction of the proteins to the AMPs, but also to show that the AMPs bind gp120 and p24 at a specific site on these HIV target proteins. PyMOL 1.3. Software were utilised to view the cartoon presentation of the

amino acid interaction between the HIV proteins (gp120 and p24) and the putative anti-HIV AMPs.

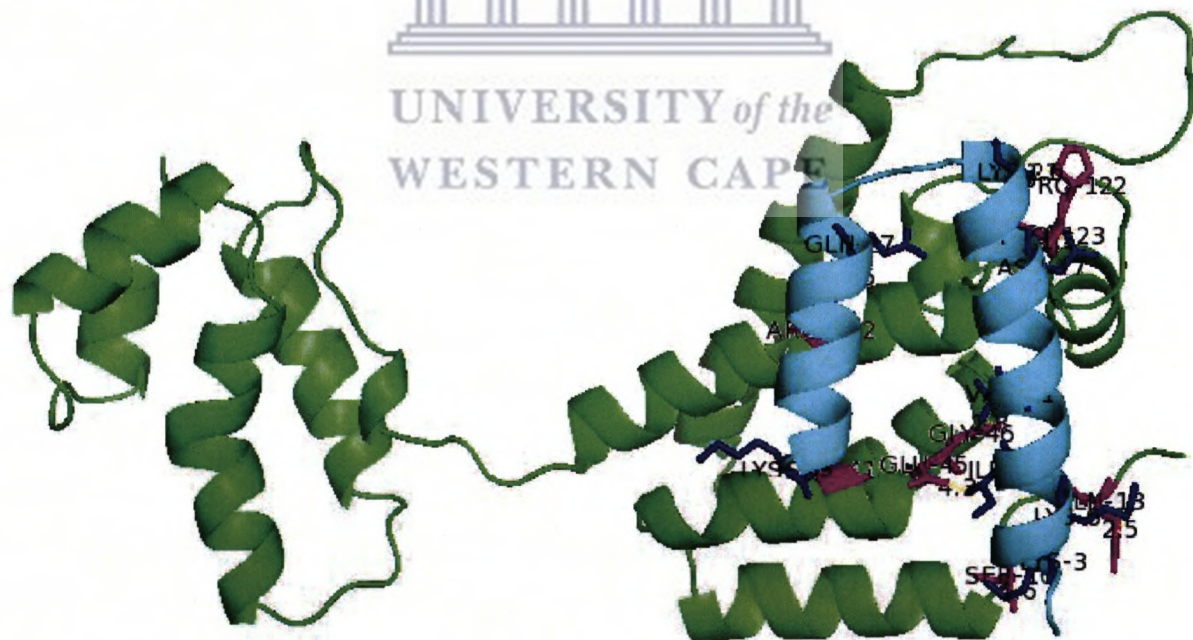
Only molecules 1, 3, 8 and 10 bind at the V1/V2 domain of the gp120 protein and at the point of interaction between the gp120 and CD4+ T cells. The complex formation of molecule 7 and gp120 was shown to have a high binding affinity geometric score, even though this putative anti-HIV AMP does not bind to the V1/V2 domain of gp120 (**Figure 3.9a, Figure 3.9b, Figure 3.9c, Figure 3.9d and Figure 3.9e**). Conversely, all the putative anti-HIV AMPs bind to the large surface interaction of the N-terminal of the p24 protein. However, only the molecules 1, 2, 3, 5, 6 and 8 were presented and used for further analysis because they had the highest binding affinity geometric scores (**Figure 3.8a, Figure 3.8b, Figure 3.8c, Figure 3.8d, Figure 3.8e and Figure 3.8f**).

The complex formed between the proteins during the p24 protein interaction of molecules 1, 2, 3, 5, 6 and 8 are shown in **Figure 3.8**. **Figure 3.9** shows the cartoon representation of the highly ranked complexes formed between the putative anti-HIV AMPs (molecules 1, 3, 7, 8 and 10) and the HIV protein gp120.

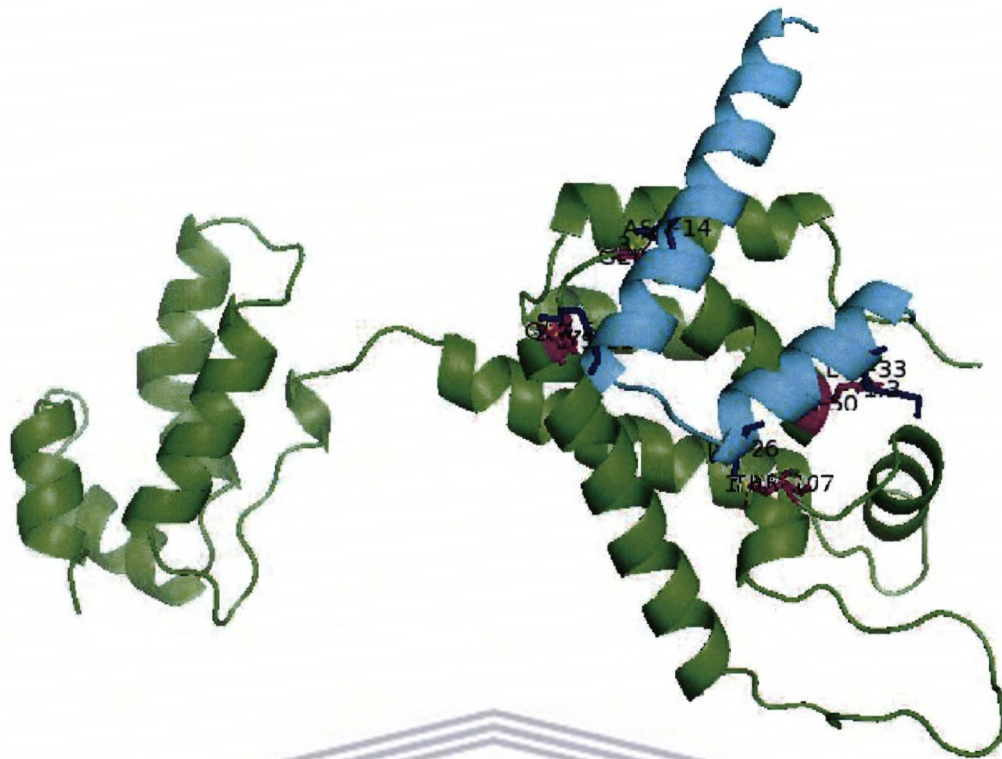
The gp41 and p17 protein results were not further considered since gp41 is a transmembrane protein and it only aids to attach the virus to the host cell after the virus has bound to the cell using the gp120 protein (Chinen and Shearer, 2002). Although p17 forms part of the HIV capsid membrane, the main protein that forms the virus capsid is the p24 protein which is utilised to diagnose the HIV infection (Sutthent *et al.*, 2003; Buttò *et al.*, 2010). Hence, the main focus was on the gp120 protein since it is the attachment point to the T cells and the p24 protein which is used as a diagnostic marker for HIV infection.



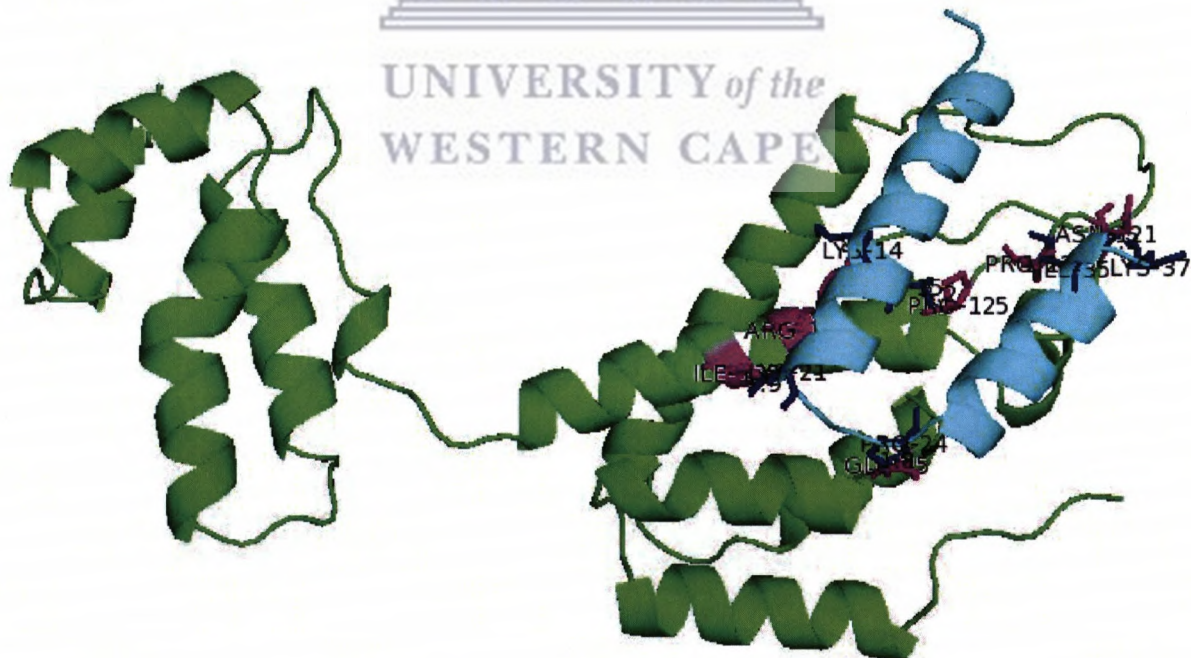
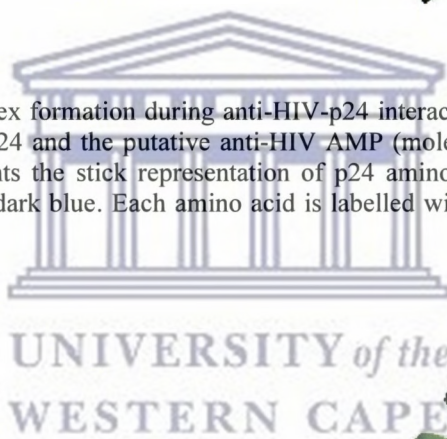
**Figure 3.8a:** p24-molecule1 complex formation during anti-HIV-p24 interaction. The cartoon representation in green colour is the HIV proteins p24 and the putative anti-HIV AMP (molecule1) is represented in light blue colour. The purple colour represents the stick representation of p24 amino acids interacting with molecule1 amino acid stick representation in dark blue. Each amino acid is labelled with the position of their amino acid for both proteins.



**Figure 3.8b:** p24-molecule2 complex formation during anti-HIV-p24 interaction. The cartoon representation in green colour is the HIV proteins p24 and the putative anti-HIV AMP (molecule2) is represented in light blue colour. The purple colour represents the stick representation of p24 amino acids interacting with molecule2 amino acid stick representation in dark blue. Each amino acid is labelled with the position of their amino acid for both proteins.

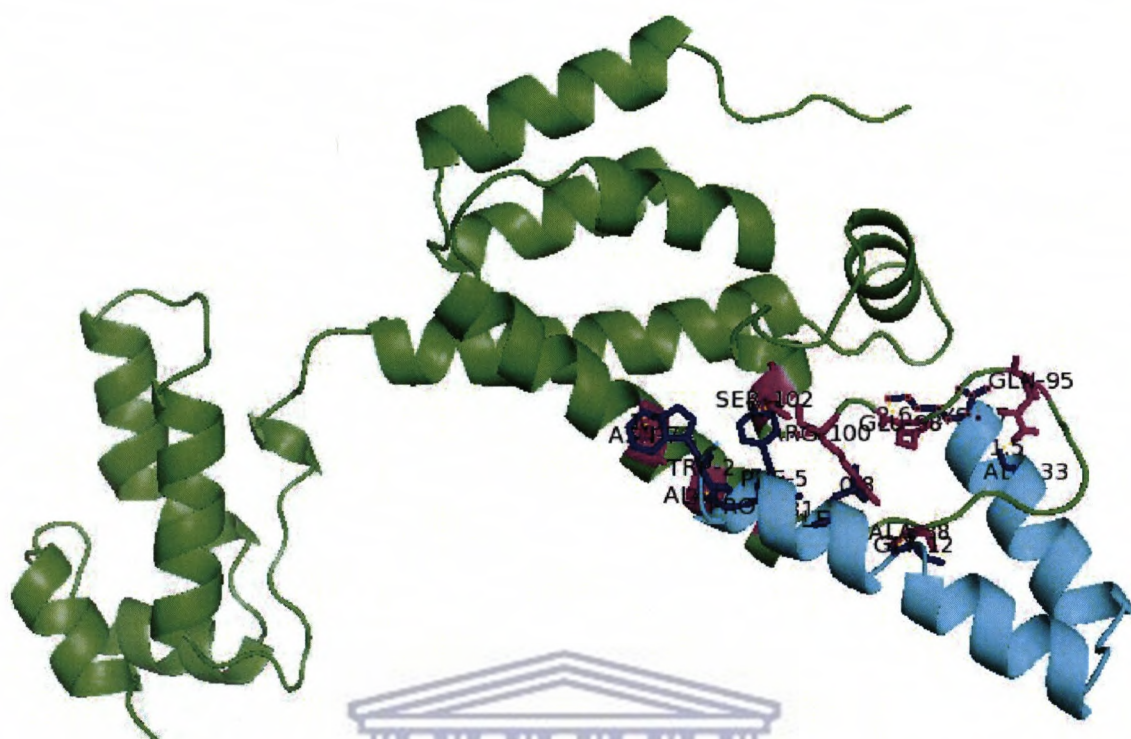


**Figure 3.8c:** p24-molecule3 complex formation during anti-HIV-p24 interaction. The cartoon representation in green colour is the HIV proteins p24 and the putative anti-HIV AMP (molecule3) is represented in light blue colour. The purple colour represents the stick representation of p24 amino acids interacting with molecule3 amino acid stick representation in dark blue. Each amino acid is labelled with the position of their amino acid for both proteins.

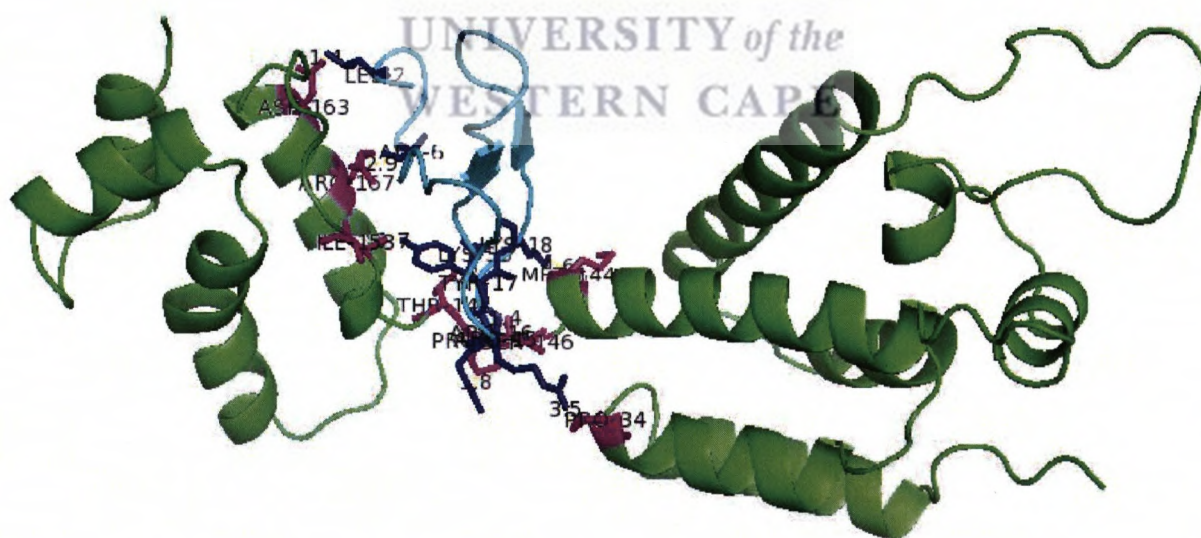


**Figure 3.8d:** p24-molecule5 complex formation during anti-HIV-p24 interaction. The cartoon representation in green colour is the HIV proteins p24 and the putative anti-HIV AMP (molecule5) is represented in light blue colour. The purple colour represents the stick representation of p24 amino acids interacting with molecule5 amino acid stick representation in dark blue. Each amino acid is labelled with the position of their amino acid for both proteins.

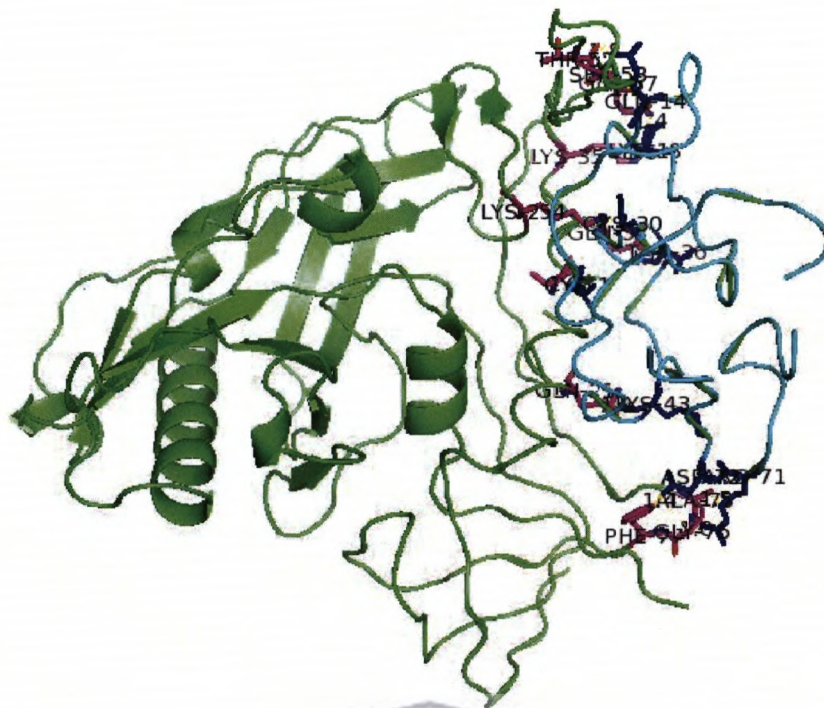




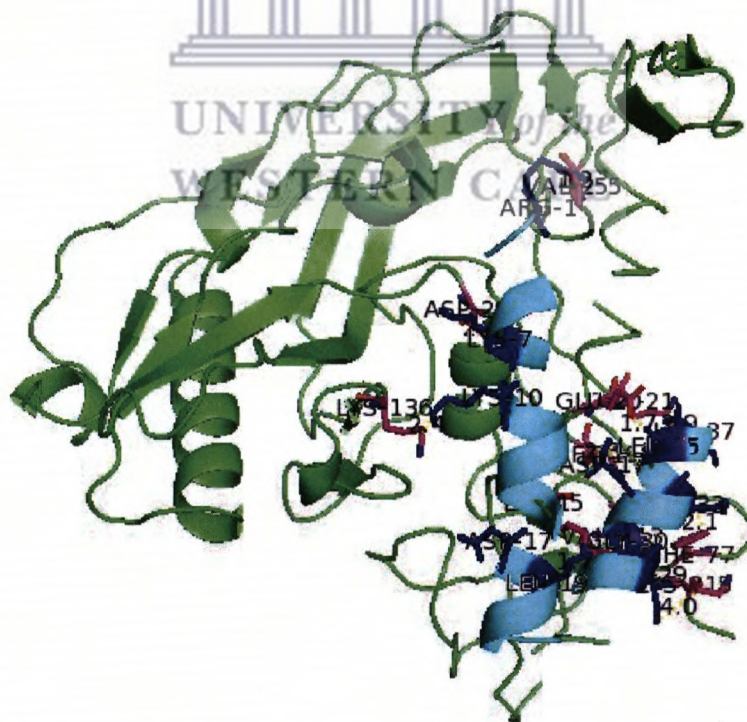
**Figure 3.8e:** p24-Molecule6 complex formation during anti-HIV-p24 interaction. The cartoon representation in green colour is the HIV proteins p24 and the putative anti-HIV AMP (molecule6) is represented in light blue colour. The purple colour represents the stick representation of p24 amino acids interacting with molecule6 amino acid stick representation in dark blue. Each amino acid is labelled with the position of their amino acid for both proteins.



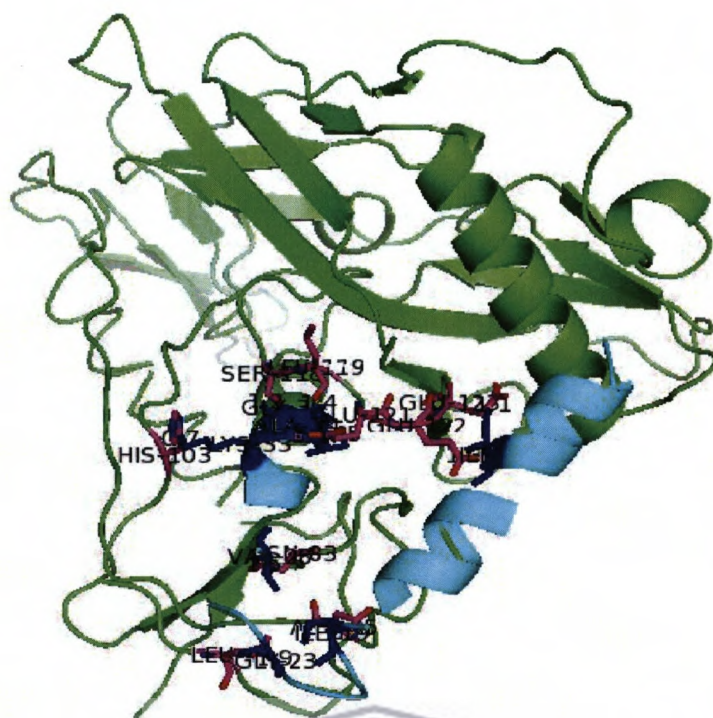
**Figure 3.8f:** p24-Molecule8 complex formation during anti-HIV-p24 interaction. The cartoon representation in green colour is the HIV proteins p24 and the putative anti-HIV AMP (molecule8) is represented in light blue colour. The purple colour represents the stick representation of p24 amino acids interacting with molecule8 amino acid stick representation in dark blue. Each amino acid is labelled with the position of their amino acid for both proteins.



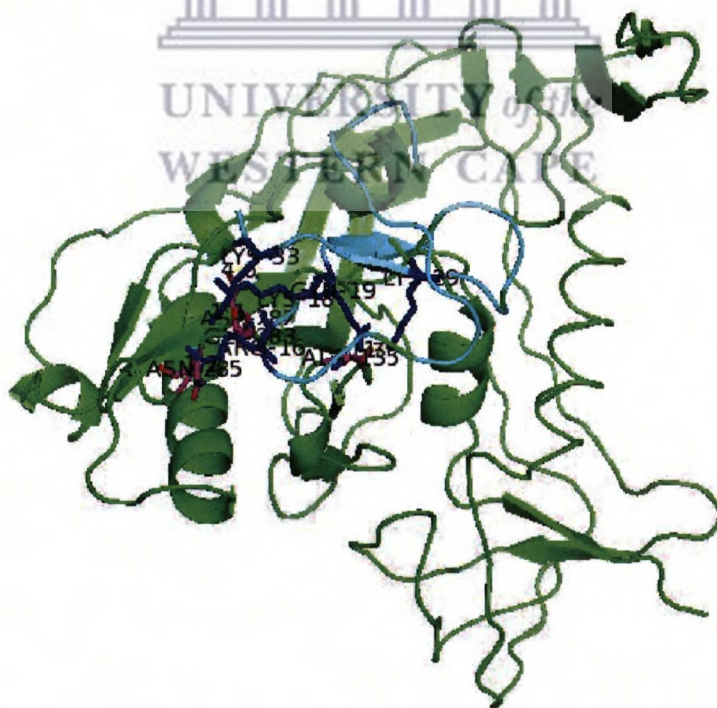
**Figure 3.9a:** gp120-molecule1 complex formation during anti-HIV-gp120 interaction. The cartoon representation in green colour is the HIV proteins gp120 and the putative anti-HIV AMP (molecule1) is represented in light blue colour. The purple colour represents the stick representation of gp120 amino acids interacting with molecule1 amino acid stick representation in dark blue. Each amino acid is labelled with the position of their amino acid for both proteins.



**Figure 3.9b:** gp120-molecule3 complex formation during anti-HIV-gp120 interaction. The cartoon representation in green colour is the HIV proteins gp120 and the putative anti-HIV AMP (molecule3) is represented in light blue colour. The purple colour represents the stick representation of gp120 amino acids interacting with molecule3 amino acid stick representation in dark blue. Each amino acid is labelled with the position of their amino acid for both proteins.



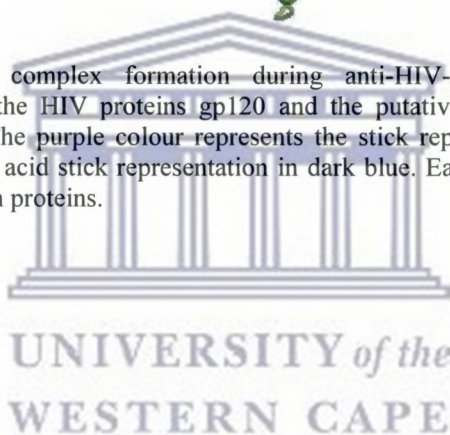
**Figure 3.9c:** gp120-molecule7 complex formation during anti-HIV-gp120 interaction. The cartoon representation in green colour is the HIV proteins gp120 and the putative anti-HIV AMP (molecule7) is represented in light blue colour. The purple colour represents the stick representation of gp120 amino acids interacting with molecule7 amino acid stick representation in dark blue. Each amino acid is labelled with the position of their amino acid for both proteins.



**Figure 3.9d:** gp120-molecule8 complex formation during anti-HIV-gp120 interaction. The cartoon representation in green colour is the HIV proteins gp120 and the putative anti-HIV AMP (molecule8) is represented in light blue colour. The purple colour represents the stick representation of gp120 amino acids interacting with molecule8 amino acid stick representation in dark blue. Each amino acid is labelled with the position of their amino acid for both proteins.



**Figure 3.9e:** gp120-molecule10 complex formation during anti-HIV-gp120 interaction. The cartoon representation in green colour is the HIV proteins gp120 and the putative anti-HIV AMP (molecule10) is represented in light blue colour. The purple colour represents the stick representation of gp120 amino acids interacting with molecule10 amino acid stick representation in dark blue. Each amino acid is labelled with the position of their amino acid for both proteins.



UNIVERSITY of the  
WESTERN CAPE

The individual amino acids of the anti-HIV AMP and the HIV protein that interact during the complex formation are represented in **Table 3.6** and **Table 3.7**. The distance measurements between each amino acid of each putative anti-HIV and the amino acid of the HIV proteins (gp120 and p24) are shown in brackets with blue colour in **Table 3.6** and **Table 3.7**. Each table gives the position of the amino acids as well as their one letter abbreviation name of the putative anti-HIV AMP and the HIV proteins p24 and gp120 that interacted during the docking.

**Table 3.7:** The individual residues interaction of the putative anti-HIV AMPs binding to the HIV protein p24. For the nomenclature, the number and letter is the position and one letter abbreviation of an amino acid on the putative anti-HIV AMP whilst, the next number and letter is the position and one letter abbreviation of an amino acid on the p24. The distance (in Amstrong) between the interacting amino acids is in bracket and has a blue colour.

AMPs/HIV Proteins	p24
Molecule1	37P-15K (1.0Å), 71R-53Q (2.3Å), 61N-97T (3.0Å), 10Q-102Q (2.8Å), 65M-47N (1.3Å), 23P-40Q (1.7Å), 18K-7P (0.4Å), 13W-4A (1.5Å), 13W-3Q (1.6Å), 36T-11N (2.7Å)
Molecule2	3K-6S (2.5Å), 6K-3Q (1.6Å), 8I-35E (4.3Å), 11V-36G (2.8Å), 16D-113P (1.5Å), 21K-112P (1.6Å), 26Q-122R (2.5Å), 37K-31S (1.3Å)
Molecule3	21K-53Q (2.7Å), 14N-50G (3.2Å), 26I-97T (1.3Å), 33K-40Q (1.3Å)
Molecule5	21K-125I (1.9Å), 24P-35E (1.8Å), 14K-122R (1.2Å), 15V-115P (4.2Å), 36K-111N (2.2Å), 35I-113P (1.2Å)
Molecule6	2W-64N (1.5Å), 4P-68A (2.4Å), 4P-71D (2.9Å), 5F-92S (1.7Å), 8L-90R (0.8Å), 37K-88E (2.6Å), 33A-85Q (1.5Å), 12G-78A (2.4Å)
Molecule8	15R-24P (3.5Å), 16R-137P (1.8Å), 18K-134M (4.6Å), 33K-136S (1.4Å), 33K-138T (1.3Å), 17Y-143I (1.7Å), 2L-153D (1.4Å), 6A-157R (2.9Å)

The abbreviation of the 20 amino acids used for the interaction study is as follow: Arginine: R, Lysine: K, Aspartic acid: D, Glutamic acid: E, Alanine: A, Isoleucine: I, Leucine: L, Phenylalanine: F, Valine: V, Proline: P, Glycine: G, Glutamine: Q, Asparagine: N, Histidine: H, Serine: S, Threonine: T, Tyrosine: Y, Cysteine: C, Methionine: M, Tryptophan: W

**Table 3.8:** The individual residues interaction of the putative anti-HIV AMPs binding to the HIV protein gp120. For the nomenclature, the number and letter is the position and one letter abbreviation of an amino acid on the putative anti-HIV AMP, whilst the next number and letter is the position and one letter abbreviation of an amino acid on the gp120. The distance (in Amstrong) between the interacting amino acids is in bracket and has a blue colour.

AMPs/HIV Proteins	gp120
Molecule1	71R-75A (1.5Å), 71R-77G (1.0Å), 18K-58Q (1.4Å), 72D-76F (1.4Å), 1C-27I (1.9Å), 30C-254K (2.2Å), 18K-36K (1.6Å), 43K-21Q (1.1Å), 36T-32Q (1.0Å), 14Q-52T (1.8Å), 14Q-53S (1.2Å)
Molecule3	35L-20E (1.7Å), 19L-15K (1.7Å), 10K-136K (2.4Å), 17D-17M (2.3Å), 37K-21Q (2.9Å), 7K-299D (2.4Å), 1R-255V (1.2Å), 29I-315K (4.0Å), 33K-77F (2.1Å), 30G-313V (1.8Å)
Molecule7	8I-123E (1.1Å), 8I-122E (1.3Å), 19I-95N (1.9Å), 23G-4L (1.9Å), 33K-103H (0.7Å), 34A-121E (1.1Å), 36G-119L (1.4Å), 35I-118S (1.2Å), 28V-83N (1.5Å)
Molecule8	16R-285N (1.2Å), 33K-282D (4.6Å), 29K-135A (1.5Å), 19Q-135A (1.7Å), 33K-283G (1.1Å)
Molecule10	28V-288N (2.1Å), 13R-189S (1.6Å), 1W-253Q (3.6Å), 3P-299D (2.4Å), 36K-132T (3.0Å)

The abbreviation of the 20 amino acids used for the interaction study is as follow: Arginine: R, Lysine: K, Aspartic acid: D, Glutamic acid: E, Alanine: A, Isoleucine: I, Leucine: L, Phenylalanine: F, Valine: V, Proline: P, Glycine: G, Glutamine: Q, Asparagine: N, Histidine: H, Serine: S, Threonine: T, Tyrosine: Y, Cysteine: C, Methionine: M, Tryptophan: W

### 3.5 Discussion

The naissance of computational biology and bioinformatics has opened doors into fields such as drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression, protein-protein interactions, and genome-wide association studies (Ouzounis, 2012). This has reduced the time spent to accomplish scientific advances since computational biology is less time consuming, cost effective and less labour intensive. Applying computational approaches, it was possible to characterise the putative anti-HIV

AMPs identified by the HMMER algorithm, by predicting their 3-D structures using the I-TASSER software and their interaction with the HIV proteins; gp120, gp41, p24 and p17 using PatchDock.

The characterisation of the putative anti-HIV AMPs based on their physicochemical properties showed that most peptides have a positive net charge ranging from +2 to +8, except for molecule9 which has a zero net charge. This positive net charge is contributed to by the presence of Arginine and Lysine amino acids. The positive net charge helps in directing the Antimicrobial Peptides to the target pathogen through electrostatic attraction. From the results shown in **Table 3.1**, molecules1, 8 and 9 contained the Cysteine (Cys) amino acid residue, molecules 2-7 and 10 however did not have the Cysteine (Cys) amino acid residue in their respective sequences. Particularity was placed on the presence of Cysteine amino acids within the AMPs sequence which is said to enhance the anti-HIV activity of the AMPs containing this amino acid, since it aids in proper folding. However, the effect of this amino acid can only be proven during the molecular examination of the anti-HIV activity of these peptides (Wang *et al.*, 2011). Additionally, the hydrophobicity of the putative anti-HIV peptides was all higher than 30%, which is the minimum hydrophobicity value for an AMP (**Table 3.1**). This property may also be a contributor to the activity towards pathogenic organisms (Hancock and Diamond, 2000; Hancock and Sahl, 2006).

Additionally, the Boman index which is an estimate of the potential of peptide to bind to different receptors such as membrane proteins, and is defined as the sum of the free energies of the amino acid residue side chains divided by the total number of amino acid residues (Boman, 2003). For Antimicrobial Peptides, a lower Boman index value  $\leq 1$  indicates that the peptide is likely to have a higher antimicrobial activity without many side effects and peptides with Boman indices less than zero only have antibacterial activity, whereas a higher index value (2.50-3.00) indicates that the peptide is multifunctional with hormone-like

activities (Boman 2003). Most putative anti-HIV AMPs had a good potential to bind to other proteins based on their Boman indices that were less than 2.5 kcal/mol. An exception was observed for molecule9, with a Boman index less than zero, meaning that molecule9 may be a good antibacterial peptide. The Boman index of molecule9 and the negative control MucroporinS1 were less than zero. No sequence similarity was shown between the putative anti-HIV AMP sequences with other known peptide sequences, thus these AMPs have not yet been implicated as anti-HIV peptides and they were therefore regarded as putative AMPs.

The net charge of the HIV proteins ranged from -2 to +7, with p17 having the highest net positive charge of +7, whilst gp41 has the lowest charge of -2. Eventhough, the net charge of gp120 is +6, this protein can still be targeted by the putative AMPs having a similar or higher net charge. Most importantly, the isoelectric point of the target HIV proteins gp120 and p24 are lower compare to molecules 1-8 and 10. This is ideal for the interaction of the AMPs and the HIV proteins since it will favour the electrostatic attraction of the AMPs towards the HIV proteins. Although the half life of gp120, gp41 and p17 are shorter, the half life of p24, which is 100 hours, is an important characteristic since the protein is used for diagnosis, and can be detected in serum by the AMPs, after being released by the virus.

The prediction of the 3-D structures of the HIV proteins as well as the putative anti-HIV AMPs, which were identified by the HMMER algorithm showed good results. The HIV proteins p24, p17, gp41 and gp120 all had C-scores higher than -1.5 suggesting that the 3-D structures of these proteins are of the correct fold (Roy *et al.*, 2010). Also, p24, p17 and gp120 3-D structures had the highest TM-scores with more than 90% similarity with their templates, whilst gp41 had a 69% similarity with its template. The TM-score is a scale for measuring the structural similarity between the predicted 3-D structure and the template structure. A TM-score > 0.5 indicates a model of correct topology and a TM-score < 0.17 means a random similarity (Roy *et al.*, 2010). All HIV proteins have TM-scores greater than



0.5, meaning that the proteins have correct topology or structural shape (Roy *et al.*, 2010). Though the C-score and the TM-score are different, it can be observed that there is direct correlation between these two parameters, although it is not always the case for all these structures. Reciprocally, there is also a strong correlation between the RMSD and the TM-score of a predicted protein or peptide (Zhang, 2008).

The prediction of the 10 putative anti-HIV AMPs, the positive and the negative controls gave C-score values which were higher than -1.5 meaning that the predicted structures have the correct fold. However, molecule1 had a C-score of -1.83, which is smaller than -1.5. This scoring may mean that the structure of molecule1 was randomly predicted and there was not enough information available for an accurate 3-D prediction (Roy *et al.*, 2010). Additionally, the TM-score and the RMSD of the 10 putative anti-HIV AMPs and the controls have values that were within the lowest ranges. All the TM-score were above 0.5 and the RMSD were around 2Å. The TM-scores being more than 0.5 signified that molecule 2-10, the positive and the negative controls have structure similarity with the templates that were used for their prediction (Zhang, 2008; Roy *et al.*, 2010). The RMSD being around 2Å meant that there were less distance between atoms of the putative peptides and the templates which were used for their 3-D structure prediction (Wei *et al.*, 1999; Carugo and Pongor, 2001). Observation can be made that there is good correlation between the TM-score and the RMSD of the predicted AMP structures (Roy *et al.*, 2010). Molecule1 still had the lowest TM-score and the highest RMSD value. Though molecule1 does not reach the expected threshold of these parameters, its E-value suggests that it could be a putative anti-HIV peptide.

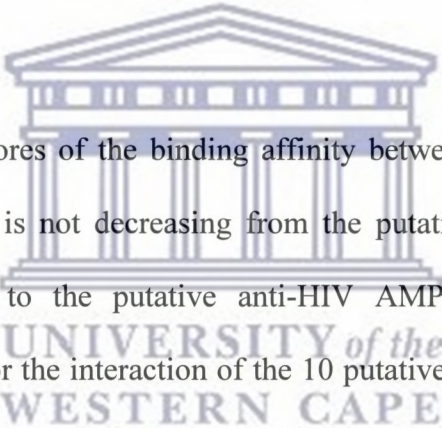
The different secondary structures exhibited by the 10 putative anti-HIV AMPs which included the  $\alpha$ -helical, the parallel  $\beta$ -sheet, anti-parallel  $\beta$ -sheet, the extended and the loop structures showed that these peptides are diverse since the secondary structure can be used for

their classification. Nevertheless, Antimicrobial Peptides with different secondary structures can still perform the same activity.

The predicted 3-D structure molecules were also superimposed with known 3-D structures by a default setting in I-TASSER. The superimposition of the predicted 3-D structures with the templates from which they were derived still showed that 9 putative anti-HIV AMPs, the controls and the HIV proteins have TM-score higher than 0.5 (expected TM-score  $> 0.5$ ), meaning that the predicted 3-D structures have the correct fold and have similar structures with the superimposed templates (Zhang, 2008; Roy *et al.*, 2010). Additionally, the RMSD of these molecules is still below 2Å, thus the molecules have less atomic deviation from the templates molecules used to predict their structures (Wei *et al.*, 1999; Carugo and Pongor, 2001; Zhang, 2008). However, 3-D structure prediction accuracy of molecule1 is still confirmed by its superimposition with the template. Eventhough its TM-score was less than 0.5 and the RMSD value higher than 2Å. Though molecule1 has a RMSD of 3.52 as a superimposition to its template, a RMSD of 4Å can still be considered acceptable (Park, and Levitt, 1996; Park *et al.*, 1997).

The activity of biomolecules such as Antimicrobial Peptides can only be effective if they interact with their target molecules. The interaction of the 10 anti-HIV AMPs and the HIV proteins gp120, gp41, p24 and p17 proved to have high interaction affinities, with their binding affinity scores greater than that of the positive and the negative controls. This means that these putative anti-HIV AMPs might have higher anti-HIV activity compared to the Kn2-7. The lack of anti-HIV activity by the negative control can be caused by its net charge of 1, which reduces the electrostatic attraction with the target microbes and a protein-binding potential lower than zero (Hancock and Diamond, 2000; Boman, 2003).

From the 10 putative anti-HIV AMPs, molecule1 had the highest geometric binding affinity eventhough its 3-D structure had low C- and TM-scores, whilst molecule9 had the lowest binding affinity out of the 10 peptides regardless of the HIV protein it was docked to (**Table 3.6**). The high binding affinity of molecule1 may be justified by its positive net charge of + 6 and a good Boman index of 2.17 kcal/mol. In additional, molecule1 had the lowest E-value, thus it has the highest potential to have anti-HIV activity. On the other hand, molecule 9 low binding affinity can be contributed to the fact that it has a zero net charge and a protein binding potential less than zero. The binding potential of this molecule was predicted from its physicochemical characterisation and has shown a strong correlation between the charge of the peptide and his binding potential during the docking study (Hancock and Sahl, 2006; Giuliani *et al.*, 2007).



Conversely, the geometric scores of the binding affinity between the 10 putative anti-HIV AMPs and the HIV proteins is not decreasing from the putative anti-HIV AMP with the lowest E-value (molecule1) to the putative anti-HIV AMP with the highest E-value (molecule10). For example, for the interaction of the 10 putative anti-HIV AMPs and gp120, molecule1 has the highest score and molecule3 has the second highest binding affinity geometric score. Surprisingly, molecule2 did not have the second highest binding affinity geometric score as it had the second lowest E-value prediction score. This might be caused by the variable parameters that contribute to the binding of the peptides to their targets, and not just by the physicochemical properties such as the positive net charge, the Boman index and the number of Cysteine and Lysine or Arginine residues present in the AMP (Wang *et al.*, 2011). On average, all the putative anti-HIV AMPs binds stronger to the gp120 protein than the gp41 protein. The binding affinity of the 10 putative anti-HIV AMPs with the HIV proteins p24 and p17 is also stronger for p24 compared to that of p17 for molecules1, 2, 3, 6, 7 and 10.

Since the interest of this project was to identify putative anti-HIV AMPs for HIV diagnosis and therapy, the specific binding of the putative anti-HIV AMPs to their targets proteins gp120 and p24 had to be taken to consideration besides the high binding affinity these peptides exhibited.

The consideration of p24 was due to its current use in HIV diagnostic assays. The p24 assay is much more simple compared to RNA or DNA based HIV assay since it does not require the use of expensive equipment to perform the assay as well as well trained personnel. The testing of HIV using the p24 protein as marker is appropriate for district clinics as it can be used in a point-of-care device (Wang *et al.*, 2010). Also, p24 protein assay is more stable and is less affected by variation in time and physical conditions during transportation and genetic diversity (Sutthent *et al.*, 2003).

Applying the principle of specific binding and the usage of p24 as a diagnostic biomarker, all 10 molecules bound to the N-terminal rather than the C-terminal of the p24 protein of the HIV. These results look promising as previous studies have shown that the antibody Fab13B5 used to detect p24 binds to the C-terminal of the protein during their complex formation. The binding of Fab13B5 to p24 antigen at its C-terminal, results in a small surface interaction between the two molecules. The implication of such a small surface interaction between p24 antigen and Fab13B5 antibody results in low binding affinity between the two molecules during their interaction (Monaco-Malbet *et al.*, 2000).

Furthermore, testing of HIV-1 patients with a regular p24 antigen assay bioMeriux showed low sensitivity in detecting the p24 antigen since the free p24 antigen may be already bound to p24 antibody thus the test can only detect about 50% of asymptomatic patients to be HIV positive (Goudsmit *et al.*, 1987). Though this detection method (bioMeriux) was upgraded by a booster step, the plasma and the serum samples to be analysed ought to be preheated before

the assay is performed in order to measure the p24 antigen (Sutthent *et al.*, 2003). Thus, the bioMerix diagnostic method still requires the usage of heating equipment and storage facilities hence is not appropriate for a point-of-care device. Since the binding of a diagnostic antibody for the detection of HIV have not yet be implicated in the binding of p24 antigen N-terminal, the putative anti-HIV AMPs would be important detection molecules as they can bind freely to p24 antigen and can favour the early diagnosis of HI Virus with high specificity.

The choice to select only molecules 1, 3, 8 and 10 was because these putative anti-HIV AMPs bind to gp120, at the V1/V2 domain and the point of interaction between gp120 and CD4+ T cell during virus entrance into the host cells. Studies have shown that the point of contact between HIV membrane protein gp120 with the CD4+ molecule and the chemokine receptor on T cells is crucial for the HIV virus to gain entry into the human T cells (Kowalski *et al.*, 1987; Lu *et al.*, 1995). Additionally, the V1/V2 domain is a fragment on the gp120 protein which has been shown to be a good binding spot for most HIV-1 neutralizing antibodies (Kwong *et al.*, 1998; McLellan *et al.*, 2001; Zhou *et al.*, 2007). If considering this hypothesis, molecules 1, 3, 8 and 10 could have good therapeutic abilities since these peptides bind to the gp120 V1/V2 domain and to the point of interaction of gp120 with CD4+ T cells. Hence, molecules 1, 3, 8 and 10 would be vital for the prevention of the HIV viral entry by restricting infection of T cells and thus replication of the virus (Kowalski *et al.*, 1987).

### **3.6. Conclusion**

The selection of the 10 putative AMPs with putative anti-HIV activity identified by the HMMER algorithm, was further backed by their performance, based on their physicochemical characterisation of the parameters employed. Additionally, all the AMPs

had a 3-D structure with good topology (TM-score) and little atomic deviation (RMSD) from their templates except for molecule1, which C-score, TM-score and RMSD were not within the expected ranges. Nonetheless, molecule1 had the highest probability to be an anti-HIV AMP based on its low E-value. Furthermore, these putative anti-HIV AMPs were used for docking and interaction studies with the HIV proteins gp120, gp41, p24 and p17.

Although all the 10 putative anti-HIV AMPs had higher geometric binding affinities than the positive and the negative controls, only molecules1, 3, 8 and 10 were considered as putative anti-HIV AMPs since they have specific binding affinity with the gp120 protein. Conversely, all 10 AMPs binds to the N-terminal of p24 protein but only molecules1, 2, 3, 5, 6 and 8 were however selected for future work since they had a high geometric binding affinity with the p24 protein. Since molecules1, 3, 8 and 10 have specific binding on the VI/V2 domain and the point of interaction with CD4+ T cells, these molecules should be considered amongst the top ranking candidates in the list of the putative anti-HIV AMPs that could be used for HIV therapy. On the other hand, due the binding of molecules1, 2, 3, 5, 6 and 8 at the N-terminal of p24 protein, these AMP molecules could be utilised as HIV diagnostic tools.

These studies are therefore fundamentally important because gp120 has been implicated in HIV propagation within patients T cells (Kwong *et al.*, 1998) while p24 HIV protein has been used as a diagnostic tool but with low sensitivity (Monaco-Malbet *et al.*, 2000; Sutthent *et al.*, 2003), hence the findings could be beneficial and represent additional markers for early diagnostics and treatment of HIV infection. Further molecular work still remains to be done, to evaluate the functions of these putative anti-HIV AMPs.

## Chapter 4:

# General discussion and conclusion

### 4.1. General discussion

Antimicrobial Peptides have been shown to have a broad range of activity against gram-positive and gram-negative bacteria, fungi, cancer, protozoa as well as viruses (Andreu and Rivas, 1998; Brodgen, 2005; Wang *et al.*, 2011). This class of peptide has also shown some anti-HIV activity (Quiñones-Mateu *et al.*, 2003; Ireland *et al.*, 2007; Wang *et al.*, 2008; Wang *et al.*, 2010). However, many Antimicrobial Peptides specifically anti-HIV AMPs remains unidentified and a great number of genome sequences remains unexplored to identify more AMPs of this class. This is due to the fact that the identification of these peptides was mostly done by using molecular techniques, which is time consuming and costly. With the help of scientific advancement and the implementation of computational biology, it is now possible to identify and predict Antimicrobial Peptides exhibiting different activities, using mathematical predictive algorithms and the information gained from experimentally validated Antimicrobial Peptides (Torrent *et al.*, 2012).

Using this approach, the identification of anti-HIV AMPs from genome databases was performed. HMMER and GLAM2 algorithm were used to build models against specific super-families, after retrieving 92 experimentally validated anti-HIV AMPs from various

databases, and classifying them according to their super-families which included: Amphibians, Microorganisms, Human defensins, Fish and crabs, Insects, Vertebrates and Plants. Whilst the sensitivity of HMMER super-family models were 60%, 100%, 100%, 100%, 0%, 50% and 54.54% respectively for the different super-families, the GLAM2 super-family models had sensitivities of 80%, 100%, 100%, 0%, 0%, 50% and 54.54% respectively. The sensitivities of the models built by both tools showed very similar results.

However, the specificity and the accuracy of each of the 7 super-family models built with the HMMER algorithm were higher than their 7 counterpart models built with the GLAM2 algorithm. The accuracies of the HMMER models were all above 95% except for the Plant super-family which was 94.79%. This implies that the models created with HMMER had more than 95% confidence to predict a peptide as a putative anti-HIV AMP. On the other hand, only the Microorganism and Plant super-family models each had specificities and accuracies above 90% confidence, for the GLAM2 algorithm.

The high performance of the HMMER models is due to a calibration step included in the construction of these models. This step is said to increase the sensitivity of the search during the query of the models with the testing set and the negative set (non-anti-HIV AMPs) (Eddy, 1998; Eddy, 2003). Hence, the prediction of false negative anti-HIV AMPs was less likely for HMMER models compared to the GLAM2 models, whilst the true negative anti-HIV AMPs predicted by HMMER models were higher than those predicted by the GLAM2 models (**Table 2.7 and Table 2.8**).

The difference in performance measurement of the HMMER algorithm as compared to the GLAM2 algorithm indicates that the HMMER algorithm is a more suitable tool for the computational modelling of proteins and/or peptides sequences since it showed higher sensitivity in the identification of putative anti-HIV AMPs. Conversely, the models of the



GLAM2 algorithm can be considered insensitive to predict putative anti-HIV AMPs when querying genome databases. Furthermore, it was also observed that the MCC of the individual super-family constructed by HMMER was greater than those built by GLAM2 algorithm of the same super-families (**Table 2.9 and Table 2.10**).

Confirming the supremacy of the HMMER algorithm in terms of its specificity and accuracy during the querying of genome sequence in identifying putative anti-HIV AMPs can be considered non-random. Whilst the HMMER models only returned the sequences which had E-value lower than the cut-off of 0.05, the GLAM2 models return all the sequences that were recognised by its models without any cut-off value restriction. This makes it very difficult to confirm the authenticity of peptides identified by GLAM2 models to be putative AMPs with anti-HIV activity.

Additionally, the scoring system of the HMMER algorithm provides more statistical and probability explanation about the identified AMP since it combines the score and the E-value of the predicted AMP (Eddy, 1998; Brahmachary *et al.*, 2004; Fjell *et al.*, 2007). This E-value gives information about the probability of that predicted peptide to be a true anti-HIV AMP. The GLAM2 algorithm does not allowed a cut-off value so as to eliminate non-specific peptides identified within the genome sequences during the database queries by the models, since there is no statistical explanation of these results, but rather a score which gives no information about the percentage of a predicted peptide to be a false positive, hence the results are not that reliable (Frith *et al.*, 2008).

The identification of putative anti-HIV AMPs by HMMER models showed that the best AMP had an E-value of  $1.4e-54$ , meaning that there is only a  $1.4e-54\%$  probability for the peptide to be falsely predicted. The AMP with the lowest score had an E-value of  $9.4e-3$ , meaning that there is only a  $9.4e-3\%$  probability for the peptide to be predicted as a false

anti-HIV AMP (Eddy, 1998; Eddy, 2003). Hence, all the putative AMPs predicted by HMMER had a higher probability to be potential anti-HIV AMPs compared to those predicted by GLAM2.

The 10 best selected putative anti-HIV AMPs, based on their E-values, for subsequent use in the docking studies were firstly characterised based on their physicochemical properties. Nine out of the 10 putative anti-HIV AMPs had a net positive charge. However, only molecule9 had a zero net charge. The zero net charge was contributed to by the low percentage of the positively charged amino acids Lysine and Arginine. These amino acids are the primary molecules which trigger the electrostatic interaction of the Antimicrobial Peptide to the pathogen (Hancock and Chapple, 1999). The zero net charge of molecule9 was shown to have a direct affect on its protein binding potential (Boman index) and its binding capacity to other proteins. The Boman index of molecule9 was lower than zero. A protein binding potential value of  $\leq 1$  indicates that the peptide will likely have increased antimicrobial activity without many side effects. Antimicrobial Peptides with Boman indices less than zero have been shown to only have antibacterial activity, and Antimicrobial Peptides with a higher index value (2.50-3.00) indicates that the peptide is multifunctional with hormone-like activities (Boman 2003). On this note, all the putative anti-HIV AMPs could have anti-HIV activity except for molecule9.

It is a well known fact that the most potent anti-HIV AMPs have a high presence of Cysteine residues (Wang *et al.*, 2011), only a few AMPs (molecules1, 8 and 9) were shown to have a percentage of Cysteine higher than 16% (**Table 3.1**). This amino acid residue may be a contributor to proper folding of the peptide in order for the peptide to fit at the binding point where gp120 interacts with CD4+ molecules of T cells. The hydrophobic content of the individual putative anti-HIV AMP was all above 30%, which is the expected value of Antimicrobial Peptide hydrophobicity. This will contribute to the affinity binding of the

putative anti-HIV AMPs to the virus and destruction of the viral membrane (Hancock and Sahl, 2006; Giuliani *et al.*, 2007).

The 3-D structures of the ten putative anti-HIV AMPs was determined using I-TASSER. The results showed that the C-scores of all the predicted 3-D structures of molecules 2-10 and the HIV proteins were above the limiting value of -1.5 especially the C-scores of gp120, p24 and p17 for which the 3-D structures have already been solved. This lends credibility to this tool for the prediction of these 3-D structures. However, it is difficult to comprehend why the C-score of gp41 is not closer to 2 since the 3-D structure of gp41 has already been solved. This may be due to some shifts in the atomic orientation during the process of modelling its 3-D structure. The C-scores observed for the 10 putative anti-HIV AMPs could be as a result of the fact that there is not available or correct templates to base the modelling and prediction of the structures of these peptides (Roy *et al.*, 2010). The C-scores of molecules 2-10 were above the limit of -1.5 except for molecule 1 which had a C-score of -1.83 and could indicate that the molecule did not have an available template for its modelling (Roy *et al.*, 2010).

Similar to the C-score, the TM-score of the predicted molecules were also higher than the cut-off value of 0.5, except for molecule1. Their TM-score being higher than 0.5 signified that molecules 2-10, the HIV proteins, the positive and the negative controls have structural similarity with the templates that were used to predict their structures (Zhang, 2008; Roy *et al.*, 2010). The result of molecule1 does not always achieve the threshold value imposed by each parameter of I-TASSER, thus it can be concluded that there is a strong correlation between the Root Mean Square Deviation (RMSD) and the TM-score of a predicted protein or peptide (Zhang, 2008). Although there is not a defined RMSD value for 3-D structure prediction, a RMSD value of 2-4 Å is considered good and a  $\text{RMSD} \leq 1 \text{ \AA}$  is considered ideal. On this note, molecules 2-10 have less distance and atomic deviation between the superimposed peptides and the templates which were used for their 3-D structure prediction

(Park, and Levitt, 1996; Wei *et al.*, 1999; Carugo and Pongor, 2001). Still, molecule1 had the highest RMSD whilst it had the lowest C-score and TM-score.

The predicted 3-D structures of the AMPs were subsequently docked with the 4 HIV proteins using PatchDock. The binding affinity of the putative anti-HIV AMPs corroborates the hypothesis that peptides with a net positive charge less than +2 would have less electrostatic attraction towards the pathogenic organisms, thus less binding affinity to these organisms. Molecule9 which had a zero net charge, presented the lowest binding affinity across its interaction with the HIV proteins gp120, gp41, p24 and p17. The zero net charge of this molecule was shown to have the lowest Boman index as well, which was confirmed with the docking study. Hence, the deduction could be made that the positive charges of an Antimicrobial Peptide influence the binding of that peptide to its target (Hancock and Sahl, 2006; Giuliani *et al.*, 2007). However, molecule1 highest binding affinity to the respective HIV proteins is contributed by its positive net charge of + 6, a good Boman index, the presence of Cysteine amino acids and the fact that it had the lowest E-value. Conversely, the binding affinities between the 10 predicted AMPs and the different HIV proteins did not show a parallel decrease from the AMPs with the lowest E-value (most probable to be an anti-HIV AMP) to the putative anti-HIV AMP with the highest E-value (less probable to be an anti-HIV AMP). The observation might be contributed to by the various parameters that contribute to a good anti-HIV peptide not taken into consideration by this study during model prediction.

The specific interaction of the 10 putative anti-HIV peptides to the p24 and gp120 proteins proved to be promising. This is as a result of all the putative anti-HIV AMPs binding to the N-terminal rather than the C-terminal of p24 HIV protein. Binding of Fab13B5 (used for the p24 assay for HIV detection) at the p24 C-terminal has shown to be of a low affinity due to the small surface interaction between the two molecules. However, the N-terminal of p24

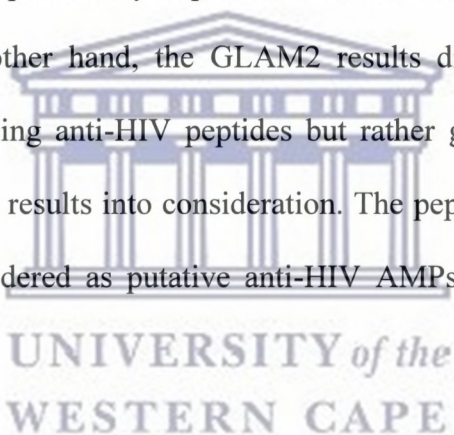
offers a larger surface interaction with other molecules (Monaco-Malbet *et al.*, 2000). Additionally, in 50% of the cases testing for presence of HIV with the p24 assay system, the C-terminal of p24 is already occupied by the antibodies produced in response to the HIV infection. Thus, the binding of the Fab13B5 to the already occupied C-terminal is prevented and can indicate a false negative test. With these AMPs binding to the unoccupied N-terminal with high specificity, can result in 100% accuracy of the p24 assay even if the C-terminal is already occupied (Goudsmit *et al.*, 1987). Although the bioMerieux assay was upgraded, the diagnosis process still requires a preheating step of the patient sample, making the method not ideal for use in poor clinic facilities and implementation as a point-of-care device.

On the other hand, the specific binding of molecules 1, 3, 8 and 10 to the V1/V2 domain of gp120 and, the point of interaction of gp120 and the CD4<sup>+</sup> of T cells, could be of great interest since this point of contact is crucial for HIV viral entry into the human T cells (Kowalski *et al.*, 1987; Lu *et al.*, 1995). Furthermore, the V1/V2 domain is a fragment on the gp120 protein which has been shown to be a good binding interacting spot for most HIV-1 neutralizing antibodies (Kwong *et al.*, 1998; McLellan *et al.*, 2001; Zhou *et al.*, 2007). The binding of molecules 1, 3, 8 and 10 could be a new strategy to target the HIV-1 and prevent its entry into the human T cells and this prevents propagation through contamination of the T cell population through competing for this target with the HI virus (Kowalski *et al.*, 1987).

## 4.2. Conclusion

Several Antimicrobial Peptides have been shown to have anti-HIV activity, thus the aim of this project was to identify putative anti-HIV AMPs using computational biology and validating the use of these AMPs as putative anti-HIV agents with *in silico* approaches. To identify this class of AMPs, experimentally validated anti-HIV AMPs were retrieved from

repository databases and super-family specific models were built using the GLAM2 and HMMER algorithms. Results showed that all the super-families specific models were successfully built using both algorithms. However, the models built by HMMER exhibited a higher performance in term of accuracy and specificity compared to models built with the GLAM2 algorithm. The use of the models constructed by HMMER and GLAM2 to query genome sequence databases showed that the HMMER models were more sensitive to identify putative anti-HIV AMPs and allowed the user to select a cut-off value. Conversely, GLAM2 models did not allow the selection of a cut-off value and were considered less sensitive. In addition, the sequences identify by the HMMER models had a score and E-value which provided more statistical and probability explanations about these sequences being putative anti-HIV peptides. On the other hand, the GLAM2 results did not give any probability prediction of the peptides being anti-HIV peptides but rather gives a simple score, which made it difficult to take these results into consideration. The peptides identified by using the HMMER models were considered as putative anti-HIV AMPs of which at least 30 were identified.



The interaction study of the putative anti-HIV AMPs with the HIV proteins was imperative to ascertain which of the peptides have specific binding to the HIV proteins. The 10 putative anti-HIV AMPs 3-D structures were predicted using the I-TASSER server, which works on a *de novo* method. The 10 AMPs and the HIV proteins had the correct 3-D structures. However, molecule1 did not have the correct topology and structural similarity with its template molecule. The binding study of the putative anti-HIV AMPs to gp120, gp41, p24 and p17 prove to be satisfactory. Nevertheless, only results for gp120 and p24 were considered useful as gp120 is crucial for the interaction of the virus to CD4+ T cells and since p24 is already used as a diagnostic biomarker in the detection of the HI Virus. Molecules1, 3, 8 and 10 binds at the point of interaction between gp120 protein and the

CD4+ molecule of T cells and the V1/V2 domain of gp120. This knowledge may be promising to develop additional therapies to prevent T cells invasion by HI Virus. All the 10 putative anti-HIV AMPs binds at the N-terminal of p24 protein as compared to the C-terminal used by the Fab13B5 antibody. However, molecules 1, 2, 3, 5, 6, and 8 had the highest binding geometric score to p24 at its N-terminal providing a possible improvement to the existing p24 diagnostic assay.

### 4.3. Future works

The abilities of these putative anti-HIV AMPs would only find their application in the prevention and/or treatment and diagnosis of HIV/AIDS if torrent examinations is made about their function during the testing of their activity on the target organism and associated molecules (gp120 and p24).

The way further with this project will include the following:

- Validate the activity of these anti-HIV AMPs as diagnostic and therapeutic agents
- Solve the crystal structure of the interaction between the anti-HIV AMPs and the HIV proteins p24 and gp120

To fulfil these aims, the objectives will be:

- ❖ Study the molecular interaction between the putative anti-HIV AMPs and the HIV proteins gp120 and p24 using Surface Plasmon Resonance (SPR),
- ❖ Study the effective concentration of the anti-HIV AMPs on the HI virus,
- ❖ Study the activity of the anti-HIV AMPs on different HIV strains,
- ❖ Study the toxicity of these anti-HIV AMPs on the host cells,
- ❖ Study the mode of action of these putative anti-HIV AMPs to the HIV-1,

- ❖ Measure the anti-HIV activity of the peptides against replication-competent HIV-1 virus,
- ❖ Improve the sensitivity and specificity detection of p24 by conjugating nano-particles to putative anti-HIV AMPs,
- ❖ Study and solve the crystal structure interaction between the anti-HIV AMPs and the HIV proteins gp120 and p24 in order to confirm the functions of the peptides in the diagnosis and treatment of HIV.





## References:

[No authors listed] (2012). HIV exposure through contact with body fluids. *Prescrire Int.*, **21** (126): 100-1, 103-5.

Abbas A. K. and Lichtman A. H., *Functions and Disorders of the Immune System*, second edition, SAUNDERS, ISBN: 0-7216-0241-X.

Aboudy, Y., Mendelson, E., Shalit, I., Bessalle, R. and Fridkin, M. (1994). *Int J Pept Protein Res.*, **43**: 573-582.

Adessi, A. and Soto, C. (2002). Converting a peptide into a drug: strategies to improve stability and bioavailability. *Curr. Med. Chem.*, **9**: 963-978.

Alimonti, J. B., Ball, T. B. and Fowke, K. R. (2003). Mechanisms of CD4+ T lymphocyte cell death in human immunodeficiency virus infection and AIDS. *J. Gen. Virol.*, **84** (7): 1649-1661.

Altman, L. K. (May 11, 1982). New homosexual disorder worries health officials. *The New York Times*, Retrieved August 31, 2011.

Andersson, M., Boman, A. and Boman, H. G. (2003). *Ascaris nematodes* from pig and human make three antibacterial peptides: isolation of cecropin P1 and two ASABF peptides. *Cell. Mol. Life Sci.*, **60**: 599-606.

Andreu, D. and Rivas, L. (1998). Animal Antimicrobial Peptides: An Overview. *Biopoly.*, **47**: 415-433.

Anglemyer, A., Rutherford, G. W., Egger, M. and Siegfried, N. (2011). Antiretroviral therapy for prevention of HIV transmission in HIV discordant couples. *Cochrane Database of Systematic Reviews*, Issue 5.

Bachar, O., Fischer, D., Nussinov, R. and Wolfson, H. J. (1993). A computer vision based technique for 3-D sequence-independent structural comparison of proteins. *Protein Eng.*, **6**: 279-288.

Back, N. K., Smit, L., De Jong, J. J., Keulen, W., Schutten, M., Goudsmit, J. and Tersmette, M. (1994). An N-glycan within the human immunodeficiency virus type-1 gp120 V3 loop affects virus neutralization. *Virology*, **199** (2): 431-438.

Baghian, A., Jaynes, J., Enright, F. and Kousoulas, K. G. (1997). *Peptides*, **18**: 177-183.

Bailes, E., Gao, F., Bibollet-Ruche, F., Courgnaud, V., Peeters, M., Marx, P. A., Hahn, B. H. and Sharp, P. M. (2003). Hybrid Origin of SIV in Chimpanzees. *Science*, **300** (5626): 1713.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., D. A., O'Donovan, C., Redaschi, N. and Yeh, L. S. L. (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Research*, **33** (Database issue): D154-D159.

Bajorath, J. (2002). Integration of virtual and high-throughput screening. *Nature Rev. Drug Discov.*, **1**: 882-894.

Baker, D. and Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, **294** (5540): 93-96.

Baker, M. A., Maloy, W. L., Zasloff, M. and Jacob, L. S. (1993). Anticancer efficacy of magainin2 and analogue peptides. *Canc. Res.*, **53**: 3052-3057.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F. and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**: 412-424.

Banerjee, A., Pramanik, A., Bhattacharjya S. and Balaram, P. (1996). Omega amino acids in peptide design: incorporation into helices. *Biopolymers*, **39**: 769-777.

Barbaro, G. and Barbarini, G. (2011). Human immunodeficiency virus and cardiovascular risk. *The Indian journal of medical research*, **134** (6): 898-903.

Barcellini, W., Colombo, G., La Maestra, L., Clerici, G., Garofalo, L., Brini, A. T., Lipton, J. M. and Catania, A. (2000).  $\alpha$ -Melanocyte-stimulating hormone peptides inhibit HIV-1 expression in chronically infected promonocytic U1 cells and in acutely infected monocytes. *Journal of Leukocyte Biology*, **68** (5): 693-699.

Barre-Sinoussi, F., Chermann, J., Rey, F., Nugeyre, M., Chamaret, S., Gruest, J., Dauguet, C. and Axler-Blin, C. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, **220** (4599): 868-871.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D. and Sonnhammer, E. L. (1999). Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.*, **27** (1): 260-262.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourn P. E. (2000). The Protein Data Bank. *Nucl Acids Res*, **28** (1): 235-242.

Blancou, P., Vartanian, J. P., Christopherson, C., Chenciner, N., Basilico, C., Kwok, S. and Wain-Hobson, S. (2001). Polio vaccine samples not linked to AIDS. *Nature*, **410** (6832): 1045-6.

Blundell, T. L., Jhoti, H. and Abell, C. (2002). High-throughput crystallography for lead discovery in drug design. *Nature Rev. Drug Discov.*, **1**: 45-54.

Boman, H. G. (1991). Antibacterial peptides: key components needed in immunity. *Cell*, **65**: 205-207.

Boman, H. G. (1995). Peptide antibiotics and their role in innate immunity. *Annu. Rev. Immunol.*, **13**: 61-92.

Boman, H. G. (2000). Innate immunity and the normal microflora. *Immunol. Rev.*, **173**: 5-16.

Boman, H. G. (2003). Antibacterial peptides: basic facts and emerging concepts. *J. Intern. Med.*, **254** (3): 197-215.

Boman, H. G., Agerberth, B. and Boman, A. (1993). Mechanisms of action on *Escherichia coli* of cecropin P1 and PR-39, two antibacterial peptides from pig intestine. *Infect. Immun.*, **61**: 2978-2984.

Bommarius, B., Jenssen, H., Elliott, M., Kindrachuk, J., Pasupuleti, M., Gieren, H., Jaeger, K. E., Hancock, R. E. and Kalman, D. (2010). Cost-effective expression and purification of antimicrobial and host defense peptides in *Escherichiacoli*. *Peptides*, **31**: 1957-1965.

Bourinbaiar, A. S. and C. F. Coleman, C. F. (1997). The effect of gramicidin, a topical contraceptive and antimicrobial agent with anti-HIV activity, against herpes simplex viruses type 1 and 2 in vitro. *Arch. Virol.*, **142**: 2225-2235.

Bourinbaiar, A. S. and Lee-Huang, S. (1994). Comparative in vitro study of contraceptive agents with anti-HIV activity: gramicidin, nonoxynol-9, and gossypol. *Contraception*, **49**: (2): 131-137.

Bourinbaiar, A. S., Krasinski, K. and Borkowsky, W. (1994). Anti-HIV effect of gramicidin in vitro: potential for spermicide use. *Life Science*, **54** (1): PL5-PL9.

Braff, M., Hawkins, M. A., Di-Nardo, A., Lopez-Garcia, B., Howell, M. D., Wong, C., Lin, K., Streib, J. E., Dorschner, R., Leung, D. Y. M. and Gallo, R. L. (2005). Structure-function relationships among human cathelicidin peptides: dissociation of antimicrobial properties from host immunostimulatory activities. *J. Immunol*, **174**: 4271-4278.

Brahmachary, M., Krishnan, S. P. T., Koh, J. L. Y., Khan, A. M., Seah, S. H., Tan, T. W., Brusic, V. and Bajic, V. B. (2004). ANTIMIC: a database of antimicrobial sequences. *Nucleic Acids Res.*, **32**: 586-589.

Brahmachary, M., Schönbach, C., Yang, L., Huang, E., Tan, S. L., Chowdhary, R., Krishnan, S. P. T., Lin, C. Y., Hume, D. A., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. and Bajic V.B. (2006). Computational promoter analysis of mouse, rat and human antimicrobial peptide-coding genes. *BMC Bioinformatics*, **7**(Suppl 5):S8.

Breukink, E. and de Kruijff, B. (1999). The Lantibiotic nisin, a special case or not. *Biophys. Acta*, **1462**: 223-234.

Brogden, K. A. (2005). Antimicrobial peptides: Pore formers or metabolic inhibitors in bacteria? *Nat. Rev. Microbiol.*, **3**: 238-250.

Brogden, K. A., Ackermann, M. and Huttner, K. M. (1997). Small, anionic, and charge-neutralizing propeptide fragments of zymogens are antimicrobial. *Antimicrob. Agents Chem.*, **41**: 1615-1617.

Brotz, H., Bierbaum, G., Leopold, K., Reynolds, P. E. and Sahl, H. G. (1998). The lantibiotic mersacidin inhibits peptidoglycan synthesis by targeting lipid II. *Antimicrob Agents Chemother.*, **42**: 154-60.

Bulet, P., Dimarcq, J. L., Hetru, C., Lagueux, M., Charlet, M., Hegy, G., Van Dorsselaer, A. and Hoffmann, J. A. (1993). A novel inducible antibacterial peptide of *Drosophila* carries an O-glycosylated substitution. *J. Biol. Chem.*, **268**: 14893-14897.

Burgoyne, R. W. and Tan, D. H. (2008). Prolongation and quality of life for HIV-infected adults treated with highly active antiretroviral therapy (HAART): a balancing act. *J. Antimicrob. Chemother.* **61** (3): 469-73.

Buttò, S., Suligo, B., Fanales-Belasio, E., and Raimondo, M. (2010). Laboratory diagnostics for HIV infection. *Diagnostic tools for HIV infection*, **46** (1): 24-33.

Cadogan, M. and Dalgleish, A. G. (2008). HIV immunopathogenesis and strategies for intervention. *Lancet Infectious Diseases*, **8** (11): 675-684.

Carugo O. and Pongor S. (2001). A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Science*, **10**: 1470-1473.

Centers for Disease Control (CDC) (1982). Update on acquired immune deficiency syndrome (AIDS)-United States. *MMWR Morb. Mortal Wkly Rep.*, 31 (37): 507-508; 513-514.

Chang, T. L. and Klotman, M. E. (2004). Defensins: Natural Anti-HIV peptides. *AIDS Reviews*, **6**: 161-168.

Chang, T. L., François, F., Mosoian, A. and Klotman, M. E. (2003). CAF-mediated human immunodeficiency virus (HIV) type 1 transcriptional inhibition is distinct from alpha-defensin-1 HIV inhibition. *J. Virol.*, **77** (12): 6777-6784.

Chang, T. L., Klotman, M. E. (2004). Defensins: Natural anti-HIV Peptides. *AIDS reviews*, **6**: 161-168.

Chen, B., Colgrave, M. L., Daly, N. L., Rosengren, K. J., Gustafson, K. R. and Craik, D. J. (2005). Isolation and characterization of novel cyclotides from *Viola hederaceae*: solution structure and anti-HIV activity of vhl-1, a leaf-specific expressed cyclotide. *J. Biol. Chem.*, **280** (23): 22395-22405.

Chen, H. M., Wang, W., Smith, D. and Chan, S. C. (1997). Effects of the anti-bacterial peptide cecropin B and its analogs, cecropins B-1 and B-2, on liposomes, bacteria, and cancer cells. *Biochim Biophys Acta*, **1336** (2): 171-179.

Chen, L., Kwon, Y. D., Zhou, T., Wu, X., O'Dell, S., Cavacini, L., Hessell, A. J., Pancera, M., Tang, M., Xu, L., Yang, Z. Y., Zhang, M. Y., Arthos, J., Burton, D. R., Dimitrov, D. S.,

- Nabel, G. J., Posner, M. R., Sodroski, J., Wyatt, R., Mascola, J. R. and Kwong, P. D. (2009). Structural basis of immune evasion at the site of CD4 attachment on HIV-1 gp120. *Science*, **326** (5956): 1123-1127.
- Chen, R., Li, L. and Weng, Z. (2003). ZDOCK: an initial-stage protein-docking algorithm. *Proteins. Struct. Func. Bioinf.*, **52**: 80-87.
- Chen, Y., Cao, L., Zhong, M., Zhang, Y., Han, C., et al. (2012). Anti-HIV-1 Activity of a New Scorpion Venom Peptide Derivative Kn2-7. *PLoS ONE*, **7** (4): e34947. doi:10.1371/journal.pone.0034947.
- Chinen, J. and Shearer, W. T. (2002). Molecular virology and immunology of HIV infection. *Journal of Allergy and Clinical Immunology*, **110** (2): 189-198.
- Chitnis, A., Rawls, D. and Moore, J. (2000). Origin of HIV Type 1 in Colonial French Equatorial Africa? *AIDS Research and Human Retroviruses*, **16** (1): 5-8.
- Chokekijchai, S., Kojima, E., Anderson, S., Nomizu, M., Tanaka, M., Machida, M., Date, T., Toyota, K., Ishida, S. and Watanabe, K. (1995). NP-06: a novel anti-human immunodeficiency virus polypeptide produced by a *Streptomyces* species. *Antimicrob. Agents Chemother.*, **39** (10) 2345-2347.
- Chopra, I. J. (1993). The magainins: antimicrobial peptides with potential for topical application. *J. Antimicrob Chemother*, **32** (3): 351-353.
- Choudhary, A. and Raines, R. T. (2011). An evaluation of peptide-bond isosteres. *Chembiochem.*, **12**: 1801-1807.
- Christensen, D.J., Gottlin, E. B., Benson, R. E. and Hamilton, P. T. (2001). Phage display for target-based antibacterial drug discovery. *Drug Discov. Today*, **6**: 721-727.

Chu K. T. and Ng, T.B. (2003). Isolation of a large thaumatin-like antifungal protein from seeds of the Kweilin chestnut *Castanopsis chinensis*. *Biochem. and Biophys. Research Communications*, **301**: 364-370.

Chugh, J. K. and Wallace, B. A. (2001). Peptaibols: models for ion channels. *Biochemical Society Transactions*, **29** (4): 565-570.

Comeau, S. R., Gatchell, D. W., Vajda, S. and Camacho, C. J. (2004). ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Res.*, **32**: W96-W99.

Cornut, I., Thiaudiere, E. and Dufourcq, J. (1994). The amphipathic  $\alpha$ -helix concept: Application to the de novo design of ideally amphipathic Leu, Lys peptides with hemolytic activity higher than that of melittin. *FEBS Letters*, **349**: 29-33.

Coutsoudis, A., Kwaan, L. and Thomson, M. (2010). Prevention of vertical transmission of HIV-1 in resource-limited settings. *Expert review of anti-infective therapy*, **8** (10): 1163-1175.

Crosby, R. and Bounse, S. (2012). Condom effectiveness: where are we now? *Sexual health*, **9** (1): 10-17.

Daher, K. A., Selsted, M. E. and Lehrer, R. I. (1986). Direct inactivation of viruses by human granulocyte defensins. *J. Virol.*, **60** (3): 1068-1074.

Daly, N. L., Clark, R. J., Plan, M. R. and Craik, D. J. (2006). Kalata B8, a novel antiviral circular protein, exhibits conformational flexibility in the cystine knot motif. *Biochem. J.*, **393**: 619-626.

Daly, N. L., Gustafson, K. R. and Craik, D. J. (2004). The role of the cyclic peptide backbone in the anti-HIV activity of the cyclotide kalata B1. *FEBS Letters*, **574** (1-3): 69-72.



Daly, N. L., Koltay, A., Gustafson, K. R., Boyd, M. R., Casas-Finet, J. R. and Craik, D. J. (1999). Solution structure by NMR of circulin A: a macrocyclic peptide having anti-HIV activity. *J. Mol. Biol.*, **285**: 333-345.

Darnag, R., Mostapha Mazouz, E. L., Schmitzer, A., Villemin, D., Jarid, A. and Cherqaoui, D. (2010). Support vector machines: development of QSAR models for predicting anti-HIV-1 activity of TIBO derivatives. *Eur J Med Chem.*, **45** (4):1590-7.

Darveau, R. P., Cunningham, M. D., Seachord, C. L., Cassiano-Clough, L., Cosand, W. L., Blake, J. and Watkins, C. S. (1991).  $\beta$ -lactam antibiotics potentiate Magainin 2 antimicrobial activity in vitro and in vivo. *Antimicrob. Agents Chemother.*, **35** (6): 1153-1159.

Das, R., Qain, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., Khare, S., Tyka, M. D., Bhat, D., Chivian, D., Kim, D. E., Sheffler, W. H., Malmström, L., Wollacott, A. M., Wang, C., Andre, I. and Baker, D. (2007). Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins*, **69** (S8): 118-128.

De Jong, A., van Heel, A. J., Kok, J. and Kuipers, O. P. (2010). BAGEL2: mining for bacteriocins in genomic data. *Nucleic Acids Research*, **38**: 647-651.

De Jong, A., van Hijum, S. A. F. T., Bijlsma, J. J. E., Kok, J. and Kuipers, O. P. (2006). BAGEL: a web-based bacteriocin genome mining tool. *Nucleic Acids Research*, **34**: 273-279.

De Waal, A., Gomes, A. V., Mensink, A., Grootegoed, J. A., Westerhoff, H. V. (1991). Magainins affect respiratory control, membrane potential and motility of hamster spermatozoa. *FEBS Lettres*, **293**: 219-223.

Dearden, J. C. (2007). *In silico* prediction of ADMET properties: how far have we come? *Expert Opin. Drug Metab. Toxicol.*, **3**: 635-639.

deKruijj, B., Breukink, E., van-Kraaj, C., Demel, R. A., Siezen, R., Kuipers, O. P. (1999). *Biochim. Biophys. Acta*, (this issue).

Department of Health and Human Services (January 2005). A Pocket Guide to Adult HIV/AIDS Treatment January 2005 edition. Retrieved 2006-01-17.

Derua, R. Gustafson, K. R. and Pannell, L. K. (1996). Analysis of the Disulfide Linkage Pattern in Circulin A and B, HIV-Inhibitory Macrocyclic Peptides. *Biochemical and Biophysical Research Communications*, **228**: 632-638.

Detlefson, D., S. Hill, K. Volk, S. Klohr, M. Tsunakawa, T. Oki, P.-F. Lin, and M. Lee. Siamycins I and II, new anti-HIV-1 peptides: sequence analysis and structure determination of siamycin I. Submitted for publication.

Díaz, M. D., de la Torre, B. G., Fernández-Reyes, M., Rivas, L., Andreu, D. and Jiménez-Barbero, J. (2011). Structural framework for the modulation of the activity of the hybrid antibiotic peptide cecropin A-melittin [CA(1-7)M(2-9)] by N<sup>ε</sup>-lysine trimethylation. *Chem. biochem.*, **12** (14): 2177-2183.

Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, D., Voelz, V. A. (2007). The protein folding problem: when will it be solved? *Curr. Opin. Struct. Biol.*, **17** (3): 342-346.

Dimock, D., Thomas, V., Cushing, A., Purdy, J. B., Worrell, C., Kopp, J. B., Hazra, R. and Hadigan, C. (2011). Longitudinal assessment of metabolic abnormalities in adolescents and young adults with HIV-infection acquired perinatally or in early childhood. *Metabolism*, **60** (6): 874-80.

Dominguez, C., Boelens, R. and Bonvin, A. M. J. J. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, **125**: 1731-1737.

Dybul, M., Fauci, A. S., Bartlett, J. G., Kaplan, J. E. and Pau, A. K. (2002). Panel on Clinical Practices for Treatment of HIV “Guidelines for using antiretroviral agents among HIV-infected adults and adolescents”. *Ann. Intern. Med.*, **137** (5 Pt 2): 381-433.

Earley, K. W., Haag, J. R., Pontes, O., Opper, K., Juehne, T., Song, K. and Pikaard, C. S. (2006). Gateway-compatible vectors for plant functional genomics and proteomics. *Plant J.*, **45**: 616-629.

Eddy, S. (2003). HMMER User’s Guide: Biological sequence analysis using profile hidden Markov models. Washington University School of Medicine, 19-31, 65-85.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**: 755-763.

Eisenberg, D., Weiss, R. M. and Terwilliger, T. C. (1982). The helical hydrophobic moment of the amphipathicity of a helix. *Nature*, **299**: 371-374.

Ekins, S., Mestres, J., Testa, B. (2007). *In silico* pharmacology for drug discovery: applications to targets and beyond. *Br. J. Pharmacol.*, **152**: 21-37.

Epand, R. M. and Vogel H. J. (1999). Diversity of antimicrobial peptides and their mechanisms of action. *Biochimica et Biophysica Acta*, **1462**: 11-28.

Fears, D. (25th January 2005). Study: Many Blacks Cite AIDS Conspiracy. The Washington Post.

Fehlbaum, P., Bulet, P., Chernysh, S., Briand, J. P., Roussel, J. P., Letellier, L., Hetru, C. and Hoffmann, J. A. (1996). Structure-activity analysis of thanatin, a 21-residue inducible insect defense peptide with sequence homology to frog skin antimicrobial peptides. *Proc Natl Acad Sci USA*, **93** (3): 1221-5.

- Feyfant, E., Sali, A. and Fiser, A. (2007). Modelling mutations in protein structures. *Protein Sci.*, **16** (9): 2030-2041.
- Fjell, C. D., Hancock, R. E. W. and Cherkasov, A. (2007). AMPer: a database and an automated discovery tool for antimicrobial peptides. *Bioinformatics*, **23** (9): 1148-1155.
- Fjell, C. D., Hiss, J. A., Hancock, R. E. W. and Schneider, G. (2012). Designing antimicrobial peptides: form follows function. *Nature review*, **11**: 37-51.
- Fjell, C. D., Jenssen, H., Hilpert, K., Cheung, W. A. Panté, N., Hancock, R. E. W. and Cherkasov, A. (2009). Identification of novel antibacterial peptides by chemoinformatics and machine learning. *J. Med. Chem.*, **52** (7): 2006-15.
- Friedman-Kien, A. E. (1981). Disseminated Kaposi's sarcoma syndrome in young homosexual men. *J. Am. Acad. Dermatol.*, **5** (4): 468-471.
- Friedrich, C. L., Rozek, A., Patrzykat, A. and Hancock, R. E. W. (2001). Structure and mechanism of action of an indolicidin peptide derivative with improved activity against Gram-positive bacteria. *J. Biol Chem.*, **276**: 24015-24022.
- Frith, M. C., Saunders, N. F. W., Kobe, B. and Bailey, T. L. (2008). Discovering Sequence Motifs with Arbitrary Insertions and Deletions. *PLoS Comput Biol.*, **4** (5): e1000071. doi:10.1371/journal.pcbi.1000071.
- Fukumoto, K., Nagaoka, I., Yamataka, A., Kobayashi, H. Yanai, T., Kato, Y. and Miyano, T. (2005). Effect of antibacterial cathelicidin peptide CAP18/LL-37 on sepsis in neonatal rats. *Pediatr Surg Int.*, **21**: 20-24.
- Futaki, S., Suzuki, T., Ohashi, W., Yagami, T., Tanaka, S., Ueda, K. and Sugiura, Y. (2001). Arginine-rich peptides. An abundant source of membrane-permeable peptides having potential as carriers for intracellular protein delivery. *J. Biol Chem.*, **276** (8): 5836-5840.

- Gallo, R. C. (2006). A reflection on HIV/AIDS research after 25 years. *Retrovirology*, **3**: 72.
- Gallo, R. C., Sarin, P. S., Gelmann, E. P., Robert-Guroff, M., Richardson, E., Kalyanaraman, V. S., Mann, D., Sidhu, G. D., Stahl, R. E., Zolla-Pazner, S., Leibowitch, J. and Popovic, M. (1983). Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science*, **220** (4599): 865-867.
- Gallo, R. L. and Nizet, V. (2003). Endogenous production of antimicrobial peptides in innate immunity and human disease. *Curr. Allergy Asthma Rep.*, **3**: 402-409.
- Gallo, R. L., Ono, M., Povsic, T., Page, C., Eriksson, E., Klagsbrun, M. and Bernfield, M. (1994). Syndecans, cell surface heparan sulfate proteoglycans, are induced by a proline-rich antimicrobial peptide from wounds. *Proc Natl Acad Sci USA*, **91** (123): 11035-9.
- Ganz, T. (2003). Defensins: antimicrobial peptides of innate immunity. *Nat. Rev Immunol.*, **3**: 710-720.
- Ganz, T. and Lehrer, R. I. (1995). Defensins. *Pharmacol. Therapeutics*, **66**: 191-205.
- Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cumminsk, L. B., Arthur, L. O., Martine Peeters, M. and George M. Shaw, G. M., Sharp, P. M. and Hahn, B. H. (1999). Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature*, **397** (6718): 436-441.
- Gao, G. H., Liu, W., Dai, J. X., Wang, J. F., Hu, Z., Zhang, Y. and Wang, D. C. (2001). Solution structure of PAFP-S: a new knottin-type antifungal peptide from the seeds of *Phytolacca americana*. *Biochemistry*, **40** (37): 10973-10978.
- Gennaro, R. and Zanetti, M. (2000). Structural features and biological activities of the cathelicidin-derived antimicrobial peptides. *Biopolymers*, **55**: 31-49.

Giuliani, A., Pirri, G. and Nicoletto, S. F. (2007). Antimicrobial peptides: an overview of a promising class of therapeutics. *CEJB*, **2** (1): 1-33.

Goraya, J., Wang, Y., Li, Z., O'Flaherty, M., Knoop, F. C., Platz, J. E. and Conlon, J. M. (2000). Peptides with antimicrobial activity from four different families isolated from the skins of the North American frogs *Rana luteiventris*, *Rana berlandieri* and *Rana pipiens*. *European Journal of Biochemistry*, **267** (3): 894-900.

Gottlieb, M. S. (2006). Pneumocystis pneumonia-Los Angeles. 1981. *Am. J. Public Health*, **96** (6): 980-1; discussion 982-3. PMC 1470612. PMID 16714472. Archived from the original on April 22, 2009. Retrieved March 31, 2009.

Goudsmit, J., Lange, J. M., Paul, D. A. and Dawson, G. J. (1987). Antigenemia and antibody titers to core and envelope antigens in AIDS, AIDS-related complex, and subclinical human immunodeficiency virus infection. *J. Infect. Dis.*, **155**: 558-560.

Goulder, P. J., Brander, C., Tang, Y., Tremblay, C., Colbert, R. A., Addo, M. M., Rosenberg, E. S., Nguyen, T., Allen, R., Trocha, A., Altfeld, M., He, S., Bunce, M., Funkhouser, R., Pelton, S. I., Burchett, S. K., McIntosh, K., Korber, B. T. and Walker, B. D. (2001). Evolution and transmission of stable CTL escape mutations in HIV infection. *Nature*, **412** (6844): 334-338.

Groot, F., Sanders, R. W., ter Brake, O., Nazmi, K., Veerman, E. C. I., Bolscher, J. G. M. and Berkhout, B. (2006). Histatin 5-Derived Peptide with Improved Fungicidal Properties Enhances Human Immunodeficiency Virus Type 1 Replication by Promoting Viral Entry. *J. Virol.*, **80** (18) 9236-9243.

Gueguen, Y., Garnier, J., Robert, L., Lefranc, M., Mougnot, I., de Lorgeril, J., Janech, M., Gross, P. S., Warr, G. W., Cuthbertson, B., Barracco, M. A., Bulet, P., Aumelas, A., Yang, Y., Bo, D., Xiang, J., Tassanakajon, A., Piquemal, D. and Bachere, E. (2006). PenBase, the

shrimp antimicrobial peptide penaeidin database: Sequence-based classification and recommended nomenclature. *Developmental and Comparative Immunology*, **30**: 283-288.

Guo, C. J., Tan, N., Song, L., Douglas, S. D. and Ho, W. Z. (2004). Alpha-defensins inhibit HIV infection of macrophages through upregulation of CC-chemokines. *AIDS*, **18** (8): 1217-1218.

Gustafson, K. R., Sowder II, R. C., Henderson, L. E., Parsons, I. C., Kashman, Y., Cardellina II, J. H., James B. McMahon, J. B., Buckheit, Jr. R. W., Pannell, L. K. and Boyd, M. R. (1994). Circulins A and B: novel HIV-inhibitor macrocyclic peptide from tropical tree *Chassalia parvifolia*. *J. Am. Chem. SOC.*, **116**: 9337-9338.

Gustafson, K. R., Walton, L. K., Sowder II, R. C., Johnson, D. G., Pannell, L. K., Cardellina II, J. H. and Boyd, M. R. (2000). New Circulin Macrocyclic Polypeptides from *Chassalia parvifolia*. *J. Nat. Prod.*, **63**: 176-178.

Hallock, Y. F., Sowder II, R. C., Pannell, L. K., Hughes, C. B., Johnson, D. G., Gulakowski, R., Cardellina II, J. H. and Boyd, M. R. (2000). Cycloviolins A-D, Anti-HIV Macrocyclic Peptides from *Leonia cyrosa*1. *J. Org. Chem.*, **65**: 124-128.

Hammami, R., Hamida, J.B., Vergoten, G. and Fliss, I. (2009). PhytAMP: a database dedicated to antimicrobial plant peptides. *Nucleic Acids Research*, **37**: 963-968.

Hammami, R., Zouhir, A., Hamida, B. J. and Fliss, I. (2007). BACTIBASE: a new web-accessible database for bacteriocin characterization. *BMC Microbiology*, **7**: 89. doi:10.1186/1471-2180-7-89.

Hammami, R., Zouhir, A., Le Lay, C., Hamida, B. J. and Fliss, I. (2010). BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC Microbiology*, **10**: 22.

Hancock, R. E. W. and Chapple, D.S. (1999). Peptide antibiotics. *Antimicrob. Agents Chemother.*, **43**: 1317-1323.

Hancock, R. E. W. and Diamond, G. (2000). The role of cationic antimicrobial peptides in innate host defences. *Trends in Microbiology*, **8** (9): 402-410.

Hancock, R. E. W. and Lehrer, R. (1998). Cationic peptides: a new source of antibiotics. *Trends Biotechnol.*, **16**: 82-88.

Hancock, R. E. W. and Sahl, H. G. (2006). Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. *Nature Biotechnology*, **24** (12): 1551-1557.

Hancock, R.E.W. and Rozek, A. (2001). Role of membranes in the activities of antimicrobial cationic peptides. *FEMS Microbiology Letters*, **206**: 143-149.

Hauck, T. S., Giri, S., Gao, Y. and Chan, W. C. W. (2010). Nanotechnology diagnostics for infectious diseases prevalent in developing countries. *Advanced Drug Delivery Reviews*, **62**: 438-448.

Heath, K. V., Singer, J., O'Shaughnessy, M. V., Montaner, J. S. and Hogg, R. S. (2002). Intentional Nonadherence Due to Adverse Symptoms Associated With Antiretroviral Therapy. *J. Acquir. Immune Defic. Syndr.*, **31** (2): 211-217.

Heller, W. T., Waring, A. J., Lehrer, R. L. and Huang, H. W. (1998). Multiple states of beta-sheet peptide protegrin in lipid bilayers. *Biochemistry*, **37**: 17331-17338.

Hiemstra, P. S., Fernie-King, B. A., McMichael, J., Lachmann, P. J. and Sallenave, J. M. (2004). Antimicrobial peptides: mediators of innate immunity as templates for the development of novel anti-infective and immune therapeutics. *Curr Pharm Des.*, **10**: 2891-2905.



Hill, P., Yee, J., Selsted, M. E. and Eisenberg, D. (1991). Crystal structure of defensin HNP-3, an amphiphilic dimer: Mechanisms of membrane permeabilization. *Science, New Series*, **251** (5000): 1481-1485.

Hillisch, A., Pineda, L. F. and Hilgenfeld, R. (2004). Utility of homology models in the drug discovery process. *Drug Discv Today*, **9** (15): 659-669.

Hogeweg, P. (2011). The Roots of Bioinformatics in Theoretical Biology *PLoS Computational Biology*, **7** (3) DOI: 10.1371/journal.pcbi.1002021.

Hooper, E. (1999). *The River: A Journey to the Source of HIV and AIDS*. Little Brown and Company.

Huang, Y., Huang, J. and Chen, Y. (2010). Alpha-helical cationic antimicrobial peptides: relationships of structure and function. *Protein Cell*, **1**: 143-152.

Huber, D., Boyd, D., Xia, Y., Olma, M. H., Gerstein, M. and Beckwith, J. (2005). Use of thioredoxin as a reporter to identify a subset of Escherichia coli signal sequences that promote signal recognition particle-dependent translocation. *J. Bacteriol.*, **187**: 2983-2991.

Hwang P. M. and Vogel H. J. (1998). Structure-function relationship of antimicrobial peptides. *Biochem. Cell Biol.*, **76**: 235-246.

Hymes, K. B., Cheung, T., Greene, J. B., Prose, N. S., Marcus, A., Ballard, H., William, D. C. and Laubenstein, L. J. (1981). Kaposi's sarcoma in homosexual men-a report of eight cases. *Lancet*, **2** (8247): 598-600.

Ireland, D. C., Colgrave, M. L. and Craik, D. J. (2006). A novel suite of cyclotides from *Viola odorata*: sequence variation and the implications for structure, function and stability. *Biochem. J.*, **400**: 1-12.

Ireland, D. C., Wang, C. K. L., Wilson, J. A., Gustafson, K. R. and Craik, D. J. (2008). Cyclotides as Natural Anti-HIV Agents. *Biopolymers (Peptide Science)*, **90**: 51-60.

Jan, M. S., Huang, Y. H., Shieh, B., Teng, R. H., Yan, Y. P., Lee, Y. T., Liao, K. K. and Li, C. (2006). CC chemokines induce neutrophils to chemotaxis, degranulation, and alpha-defensin release. *J. Acquir. Immune Defic. Syndr.*, **41** (1): 6-16.

Janeway, C. A. Jr. (1998). Presidential address to the American Association of Immunologists. The road less traveled by: the role of innate immunity in the adaptive immune response. *J. Immunol.*, **161**: 539-544.

Javadpour, M. M., and Barkley, M. D. 1997. Self-assembly of designed antimicrobial peptides in solution and micelles. *Biochemistry*, **36**: 9540-9549.

Jenssen, H., Hamill, P. and Hancock, R. E. (2006). Peptide antimicrobial agents. *Clin Microbiol Rev.*, **19** (3): 491-511.

Jia, J., Yang, L and Z. Zhang (2006). EHPred: an SVM-based method for epoxide hydrolases recognition and classification. *J. Zhejiang Univ Sci B.*, **7** (1): 1-6.

Joint United Nations Programme on HIV/AIDS (2010). Overview of the global AIDS epidemic. *UN report on the global AIDS epidemic 2010*.

Juretić, D., Vukicević, D., Ilić, N., Antcheva, N. and Tossi, A. (2009). Computational design of highly selective antimicrobial peptides. *J. Chem. Inf. Model*, **49**, 2873-2882.

Kallings, L. O. (2008). The first postmodern pandemic: 25 years of HIV/AIDS. *J. Intern. Med.*, **263** (3): 218-243.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C. and Vakser, I. A. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl Acad. Sci. USA*, **89**: 2195-2199.

Keele, B. F., van Heuverswyn, F., Li, Y. Y., Bailes, E., Takehisa, J., Santiago, M. L., Bibollet-Ruche, F., Chen, Y., Wain, L. V., Liegois, F., Loul, S., Mpoudi-Ngole, E., Bienvenue, Y., Delaporte, E., Brookfield, J. F. Y., Sharp, P. M., Shaw, G. M., Peeters, M., and Hahn, B. H. (2006). Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1. *Science*, **313** (5786): 523-526.

Kendall, A. E. (2012). U.S. Response to the Global Threat of HIV/AIDS: Basic Facts. *Congressional Research Service*, 1-16.

Kitchen, D. B., Decornez, H., Furr, J. R. and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Rev. Drug Discov.*, **3**: 935-949.

Koltay, A., Daly, N. L. Gustafson, K. R. and Craik, D. J. (2005). Structure of Circulin B and Implications for Antimicrobial Activity of the Cyclotides. *Internat. J. of Peptide Research and Therapeutics*, **11** (1): 99-106.

Kondejewski, L. H., Jelokhani-Niaraki, M., Farmer, S. W., Lix, B., Kay, C. M., Sykes, B. D., Hancock, R. E. W. and Hodges, R. S. (1999). *J. Biol. Chem.*, **274**: 13181-13192.

Kowalski, M. L., Potz, J., Basiripour, L., Dorfman, T., Goh, W. C., Terwilliger, E., Dayton, A., Rosen, C., Haseltine, W. and Sodroski, J. (1987). Functional regions of the envelope glycoprotein of human immunodeficiency virus type 1. *Science*, **237**: 1351-1355.

Kragol, G., Lovas, S., Varadi, G., Condie, B. A., Hoffmann, R. and Otvos, L. Jr. (2001). The antibacterial peptide pyrrolicin inhibits the ATPase actions of DnaK and prevents chaperone-assisted protein folding. *Biochemistry*, **40**: 3016-3026.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. and Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, **161**: 269-288.

Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J. and Hendrickson, W. A. (1998). Structure of an HIVgp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*, **393**: 648-659.

Lai, R., Zheng, Y., Shen, J., Liu, G., Liu, H., Lee, W., Tang, S. and Zhang, Y. (2002). Antimicrobial peptides from skin secretions of Chinese red belly toad *Bombina maxima*. *Peptides*, **23**: 427-435.

Langer, T. and Hoffmann, R. D. (2001). Virtual screening: an effective tool for lead structure discovery. *Curr. Pharm. Design*, **7**: 509-527.

Lee, S. H., Kim, S. J., Lee, Y. S., Song, M. D., Kim, I. H. and Won, H. S. (2011). De novo generation of short antimicrobial peptides with simple amino acid composition. *Regulatory Peptides*, **166**: 36-41.

Lehrer R. I. and Ganz, T. (2002). Cathelicidins: a family of endogenous antimicrobial peptides. *Curr. Opin. Hematol.*, **9**: 18-22.

Lehrer, R. I., Barton, A., Daher, K. A., Harwig, S. S., Ganz, T. and Selsted, M. E. (1989). Interaction of human defensins with *Escherichia coli*. Mechanism of bactericidal activity. *J. Clin. Invest*, **84** (2): 553-561.

Lengauer, T. and Rarey, M. (1996). Computational methods for biomolecular docking. *Curr. Opin. Struct. Biol.*, **6** (3): 402-406.

Li, Y. and Chen, Z. (2008). RAPD: a database of recombinantly-produced antimicrobial peptides. *FEMS Microbiol Lett.*, **289**: 126-129.

Lin, A. L., Johnson, D. J., Stephan, K. T. and Yeh, C. K. (2004). Salivary secretory leukocyte protease inhibitor increases in HIV infection. *J. Oral Pathol. Med.*, **33**: 410-416.

Lin, P. F., Samanta, H., Bechtold, C. M., Deminie, C. A., Patick, A. K., Alam, M., Riccardi, K., Rose, R. E., White, R. J. and Colonno, R. J. (1996). Characterization of Siamycin I, a Human Immunodeficiency Virus Fusion Inhibitor. *Antimicrob. Agents Chemother.*, **40** (1): 133-138.

Lorina, C., Saidib, H., Belaidc, A., Zairic, A., Baleuxd, F., Hocinib, H., Bélecb, L., Hanic, K. and Tangya, F. (2005). The antimicrobial peptide Dermaseptin S4 inhibits HIV-1 infectivity in vitro. *Virology*, **334**: 264-275.

Lu, M., Blackow, S. and Kim, P. (1995). A trimeric structural domain of the HIV-1 transmembrane glycoprotein. *Nature Struct. Biol.*, **2**: 1075-1082.

Lyskov, S. and Gray, J. J. (2008). The RosettaDock server for local protein-protein docking. *Nucleic Acids Res.*, **36**: W233-W238.

Macindoe, G., Mavridis, L., Venkatraman, V., Devignes M. D. and Ritchie, D.W. (2010). HexServer: an FFT-based protein docking server powered by graphics processors. *Nucleic Acids Res.*, **38**: W445-W449.

Mandell, Bennett, and Dolan (2010). Chapter 121.

Marcos, J. F., Beachy, R. N., Houghten, R. A., Blondelle, S. E. and Perez-Payá, E. (1995). Inhibition of a plant virus infection by analogs of melittin. *Proc Natl. Acad. Sci. USA*, **92** (26): 12466-12469.

Markowitz, edited by William N. Rom; associate editor, Steven B. (2007). Environmental and occupational medicine (4th ed.). Philadelphia: Wolters Kluwer/Lippincott Williams and Wilkins. p. 745. ISBN 978-0-7817-6299-1.

Marr, A. K., Gooderham W. J. and Hancock R. E. W. (2006). Antibacterial peptides for therapeutic use: obstacles and realistic outlook. *Curr. Opin. Pharmacol.*, **6**: 468-472.

Marsden, M. D., Zack, J. A. (2009). Eradication of HIV: current challenges and new directions. *J Antimicrob Chemother.*, **63** (1): 7-10.

Matsuzaki, K. (2009). Control of cell selectivity of antimicrobial peptides. *Biochimica et Biophysica Acta*, **1788**: 1687-1692.

Matsuzaki, K., Nakayama, M., Fukui, M., Otaka, A., Funakoshi, S., Fujii, N., Bessho, K. and Miyajima, K. (1993). Role of disulfide linkages in tachyplesin–lipid interactions. *Biochemistry*, **32** (43): 11704-11710.

Matsuzaki, K., Sugishita, K., Fujii, N. and Miyajima, K. (1995). Molecular Basis for Membrane Selectivity of an Antimicrobial Peptide, Magainin 2. *Biochemistry*, **34** (10): 3423-3429.

Matsuzaki, K., Sugishita, K., Harada, M., Fujii, N. and Miyajima, K. (1997a). Interactions of an antimicrobial peptide, magainin 2, with outer and inner membranes of Gram-negative bacteria. *Biochim. Biophys. Acta*, **1327**: 119-130.

Matsuzaki, K., Yoneyama, S., Fujii, N., Miyajima, K., Yamada, K., Kirino, Y. and Anzai, K. (1997b). Membrane permeabilization mechanisms of a cyclic antimicrobial peptide, tachyplesin I, and its linear analog. *Biochemistry*, **36** (32): 9799-9806.

Mátyus, E., Kandt, C. and Tieleman, D. P. (2007). Computer simulation of antimicrobial peptides. *Curr. Med. Chem.*, **14** (26): 2789-2798.

McLellan, J. S., Pancera, M., Carrico, C., Gorman, J. et al., (2011). Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. *Nature*, **480**: 336-345.

McNeely, T. B., Shugars, D. C., Rosendahl, M., Tucker, C., Eisenberg, S. P. and Wahl, S. M. (1997). Inhibition of Human Immunodeficiency Virus Type 1 Infectivity by Secretory

Leukocyte Protease Inhibitor Occurs Prior to Viral Reverse Transcription. *Blood*, **90** (3): 1141-1149.

Miyata, T., Tokunaga, F., Yoneya, T., Yoshikawa, K., Iwanaga, S., Niwa, M., Takao, T. and Shimonishi, Y. (1989). Antimicrobial peptides, isolated from horseshoe crab hemocytes, tachyplesin II, and polyphemusins I and II: chemical structures and biological activity. *J. of Biochemistry*, **106** (4): 663-668.

Monaco, V., Locardi, E., Formaggio, F., Crisma, M., Mammi, S., Peggion, C., Toniolo, C., Rebuçat, S. and Bodo, B. (1998). Solution conformational analysis of amphiphilic helical, synthetic analogs of the lipopeptaibol trichogin GA IV. *J. Pept. Res.*, **52** (4): 261-272.

Monaco-Malbet, S., Berthet-Colominas, C., Novelli, A., Battai, N., Piga, N., Cheynet, V., Mallet, F. and Cusack, S. (2000). Mutual Conformational Adaptations in Antigen and Antibody upon Complex Formation between an Fab and HIV-1 Capsid Protein p24. *Structure*, **8**: 1069-1077.

Montessori, V., Press, N., Harris, M., Akagi, L. and Montaner, J. S. (2004). Adverse effects of antiretroviral therapy for HIV infection. *CMAJ*, **170** (2): 229-238.

Montville, T. J. and Chen, Y. (1998). Mechanistic action of pediocin and nisin: recent progress and unresolved questions. *Appl. Microbiol. Biotechnol.*, **50** (5): 511-519.

Moore, A. J., Devine, D. A. and Bibby, M. C. (1994). Preliminary experimental anticancer activity of cecropins. *Pept Res.*, **7** (5): 265-269.

Morimoto, M., Mori, H., Otake, T., Ueba, N., Kunita N., Niwa, M., Murakami, T. and Iwanaga, S. (1991). Inhibitory effect of Tachyplesin I on the proliferation of human immunodeficiency virus in vitro. *Chemotherapy*, **37**: 206-211.

Mulvenna, J. P., Wang, C. and Craik, D. J. (2006). CyBase: a database of cyclic protein sequence and structure. *Nucleic Acids Research*, **34**: 192-194.

Münk, C., Wei, G., Yang, O. O., Waring, A. J., Wang, W., Hong, T., Lehrer, R. I., Landau, N. R. and Cole, A. M. (2003). The theta-defensin, retrocyclin, inhibits HIV-1 entry. *AIDS Res Hum Retroviruses*, **19** (10): 875-881.

Murakami, T., Niwa, M., Tokunaga, F., Miyata, T. and Iwanaga, S. (1991). Direct Virus Inactivation of Tachyplesin I and Its Isopeptides from Horseshoe Crab Hemocytes. *Chemotherapy*, **37**: 327-334.

Mygind, P. H., Fischer, R. L., Schnorr, K. M., Hansen, M. T., Sönksen, C. P., Ludvigsen, S., Raventós, D., Buskov, S., Christensen, B., De Maria, L., Taboureau, O., Yaver, D., Elvig-Jørgensen, S. G., Sørensen, M. V., Christensen, B. E., Kjærulff, S., Frimodt-Møller, N., Lehrer, R. I., Zasloff, M. and Kristensen, H. H. (2005). Plectasin is a peptide antibiotic with therapeutic potential from a saprophytic fungus. *Nature*, **437**: 975-980.

Nakashima, H., Masuda, M., Murakami, T., Koyanagi, Y., Matsumoto, A., Fujii, N. and Yamamoto, N. (1992). Anti-human immunodeficiency virus activity of a novel synthetic peptide, T22 ([Tyr-5,12, Lys-7]polyphemusin II): a possible inhibitor of virus-cell fusion. *Antimicrob Agents Chemother.*, **36** (6): 1249-1255.

Nakashima, H., Yamamoto, N., Masuda, M. and Fujii, N. (1993). Defensins inhibit HIV replication in vitro. *AIDS*, **7** (8): 1129.

Navon-Venezia, S., Feder, R., Gaidukov, L., Carmeli, Y. and Mor, A. (2002). Antibacterial Properties of Dermaseptin S4 Derivatives with In Vivo Activity. *Antimicrob. Agents Chemother.*, **46** (3): 689-694.

Ngai, P. H. K. and Ng, T.B. (2004). Coccinin, an antifungal peptide with antiproliferative and HIV-1 reverse transcriptase inhibitory activities from large scarlet runner beans. *Peptides*, **25**: 2063-2068.



Ngai, P. H. K., Zhao, Z. and Ng, T.B. (2005). Agrocybin, an antifungal peptide from the edible mushroom *Agrocybe cylindracea*. *Peptides*, **26**: 191-196.

Nicolaou, C. A., Apostolakis, J. and Pattichis, C. S. (2009). *De novo* drug design using multiobjective evolutionary graphs. *J. Chem. Inf. Model.*, **49**: 295-307.

Nieuwkerk, P., Sprangers, M., Burger, D., Hoetelmans, R. M., Hugén, P. W., Danner, S. A., van Der Ende, M. E., Schneider, M. M., Schrey, G., Meenhorst, P. L., Sprenger, H. G., Kauffmann, R. H., Jambroes, M., Chesney, M. A., de Wolf, F., Lange, J. M. and the ATHENA Project (2001). Limited Patient Adherence to Highly Active Antiretroviral Therapy for HIV-1 Infection in an Observational Cohort Study. *Arch. Intern. Med.*, **161** (16): 1962-1968.

Nos-Barberá, S., Portolés, M., Morilla, A., Ubach, J., Andreu, D., Paterson, C. A. (1997). Effect of hybrid peptides of cecropin A and melittin in an experimental model of bacterial keratitis. *Cornea*, **16**: 101-106.

Oren, Z., Lerman, J. C., Gudmunsson, G. H., Birgitta Agerberth, B. and Shai, Y. (1999). Structure and organization of the human antimicrobial peptide LL-37 in phospholipid membranes: relevance to the molecular basis for its non-cell-selective activity. *Eur. J. Biochem.*, **341**: 501-513.

Ostresh, J. M., Blondelle, S. E., Dörner, B. and Houghten, R. A. (1996). Generation and use of nonsupported-bound peptide and peptidomimetic combinatorial libraries. *Methods Enzymol.*, **267**: 220-234.

Otvos, L. Jr., Rogers, M. E., Consolvo, P. J., Condie, B. A., Lovas, S., Bulet, P. and Blaszczyk-Thurin, M. (2000). Interaction between heat shock proteins and antimicrobial peptides. *Biochemistry*, **39** (46): 14150-14159.

Ouzounis, C. A. (2012). Rise and Demise of Bioinformatics? Promise and Progress. *PLoS Comput Biol.*, **8** (4): e1002487. doi:10.1371/journal.pcbi.1002487.

Pag, U., Oedenkoven, M., Sass, V., Shai, Y., Shamova, O., Antcheva, N., Tossi, A. and Sahl, H. G. (2008). Analysis of in vitro activities and modes of action of synthetic antimicrobial peptides derived from an  $\alpha$ -helical 'sequence template'. *J. Antimicrob. Chemother.*, **61** (2): 341-352.

Park, B. H., Huang, E. S. and Levitt, M. (1997). Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.*, **266**: 831-846.

Park, C. B., Kim, H. S. and Kim, S. C. (1998). Mechanism of action of the antimicrobial peptide buforin II: buforin II kills microorganisms by penetrating the cell membrane and inhibiting cellular functions. *Biochem. Biophys. Res. Commun.*, **244**: 253-257.

Patrzykat, A., Friedrich, C. L., Zhang, L., Mendoza, V. and Hancock, R. E. W. (2002). Sublethal concentrations of pleurocidin-derived antimicrobial human monocytes, macrophages and dendritic cells. *Immunology*, **106**: 517-525.

Peitch, M. C. (1997). Large scale protein modelling and model repository. *Proc Int Conf Intell Syst Mol Biol.*, **5**: 234-236.

Peitch, M. C., Schwede T. and Guex, N. (2000). Automated protein modelling- the proteome in 3D. *Pharmacogenomics*, **1** (3): 257-266.

Peschel, A. and Sahl, H. G. (2006). The co-evolution of host cationic antimicrobial peptides and microbial resistance. *Nat. Rev. Microbiol.*, **4**: 529-536.

Petsalaki, E., Stark, A., García-Urdiales, E., Russell, R.B. (2009). Accurate prediction of peptide binding sites on protein surfaces, *PLoS Comput Biol.*, **5** (3): e1000335.

Pirogova, E., Istivan, T., Gan, E. and Cosic, I. (2011). Advances in methods for therapeutic peptide discovery, design and development. *Curr. Pharm. Biotechnol.*, **12**: 1117-1127.

Plantier, J. C., Leoz, M., Dickerson, J. E., De Oliveira, F., Cordonnier, F., Lemée, V., Damond, F., Robertson, D. L. and Simon, F. (2009). A new human immunodeficiency virus derived from gorillas, *Nature Medicine*, **15** (8): 871-872.

Poole, A. M. And Ranganathan, R. (2006). Knowledge-based potentials in protein design. *Curr. Opin. Struct. Biol.*, **16** (4): 508-513.

Powderly, W. G. (2000). Cryptococcal Meningitis in HIV-Infected Patients. *Curr Infect Dis Rep.*, **2** (4): 352-357.

Powers, J. S. and Hancock, R. E. W. (2003). The relationship between peptide structure and antibacterial activity. *Peptides*, **24**: 1681-1691.

Prado-Prado, F. J., Martinez de la Vega, O., Uriarte, E., Ubeira, F. M., Chou, K. C. and González-Díaz, H. (2009). Unified QSAR approach to antimicrobials. 4. Multi-target QSAR modeling and comparative multi-distance study of the giant components of antiviral drug-drug complex networks. *Bioorg. Med. Chem.*, **17** (2): 569-575.

Pukala, T. L., John H. Bowie, J. H., Maselli, V. M., Musgrave, I. F. and Tyler, M. J. (2005). Host-defence peptides from the glandular secretions of amphibians: structure and activity. *Nat. Prod. Rep.*, **23**: 368-393.

Qu, X. D., Harwig, S. S., Oren, A. M., Shafer, W. M., Lehrer, R. I. (1996). Susceptibility of *Neisseria gonorrhoeae* to protegrins. *Infect. Immun.*, **64** (4): 1240-1245.

Quaranta, M. G., Mattioli, B. and Vella, S. (2012). Glances in Immunology of HIV and HCV Infection. *Advances in Virology*, doi: 10.1155/2012/434036.

Quiñones-Mateu, M. E., Lederman, M. M., Feng, Z., Chakraborty, B., Weber, J., Rangel, H. R., Marotta, M. L., Mirza, M., Jiang, B., Kiser, P., Medvik, K., Sieg, S. F. and Weinberg, A. (2003). Human epithelial beta-defensins 2 and 3 inhibit HIV-1 replication. *AIDS*, **17** (16): F39-48.

Rao, X. C., Li, S., Hu, J. C., Jin, X. L., Hu, X. M., Huang, J. J., Chen, Z. J., Zhu, J. M. and Hu, F. Q. (2004). A novel carrier molecule for high-level expression of peptide antibiotics in *Escherichia coli*. *Protein Expr. Purif.*, **36**: 11-18.

Rathinakumar, R., Walkenhorst, W. F. and Wimley, W. C. (2009). Broad-spectrum antimicrobial peptides by rational combinatorial design and high-throughput screening: the importance of interfacial activity. *J. Am. Chem. Soc.*, **131**: 7609-7617.

Reeves, J. D. and Doms, R. W. (2002). Human Immunodeficiency Virus Type 2. *J. Gen. Virol.*, **83** (Pt 6): 1253-1265.

Reynell, L. and Trkola, A. (2012). HIV vaccines: an attainable goal? *Swiss medical weekly*, **142**: w13535.

Rinaldi, A. C. (2002). Antimicrobial peptides from amphibian skin: an expanding scenario. *Current Opinion in Chemical Biology*, **6**: 799-804.

Robinson, J. A. (2011). Protein epitope mimetics as anti-infectives. *Curr. Opin. Chem. Biol.* **15**: 379-386.

Robinson, Jr. W. E., McDougall, B., Tran, D. and Selsted, M. E. (1998). Anti-HIV-1 activity of indolicidin, an antimicrobial peptide from neutrophils. *Journal of Leukocyte Biology*, **63**: 94-100.

Rohl, C. A., Strauss, C. E., Misura, K. M. and Baker, D. (2004). Protein structure prediction using Rosetta. *Meth Enzymol.*, **383**: 66-93.

Roy, A., Kucukural, A. and Zhang, Y. (2010). I-TASSER a unified platform for automated protein structure and function prediction. *Nature protocols*, **5** (4): 725-738.

Rozek, A., Friedrich, C. L. and Hancock, R. E. W. (2000). Structure of the bovine antimicrobial peptide indolicidin bound to dodecylphosphocholine and sodium dodecyl sulfate micelles. *Biochemistry*, **39**: 15765-15774.

Rubio, V., Shen, Y., Saijo, Y., Liu, Y., Gusmaroli, G., Dinesh-Kumar, S. P. and Deng, X. W. (2005). An alternative tandem affinity purification strategy applied to *Arabidopsis* protein complex isolation. *Plant J.*, **41**: 767-778.

Saether, O., Craik, D. J., Campbell, I. D., Sletten, K., Juul, J. and Norman, D. G. (1995). Elucidation of the primary and three-dimensional structure of the uterotonic polypeptide kalata B1. *Biochemistry*, **34**: 4147-4158.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.*, **33**: W363-W367.

Schuster, F. L. and Jacobs, L. S. (1992). Effects of magainins on ameba and cyst stages of *Acanthamoeba polyphaga*. *Antimicrob Agents Chemother.*, **36** (6): 1263-1271.

Schwab, I. R., Dries, D., Cullor, J., Smith, W., Mannis, M., Reid, T. and Murphy, C. J. (1992). Corneal storage medium preservation with defensins. *Cornea*, **11** (5): 370-375.

Schwede, T., Sali, A., Eswar, N. and Peitsch M. C. (2008). 'Protein Structure Modelling' in Schwede, T. and Peitsch M. C. (eds), Singapore 596224, world Scientific Publishing Co. Pte. Ltd.

Scott, M. G., Davidson, D. J., Gold, M. R., Bowdish, D. and Hancock, R. E. W. (2002). The human antimicrobial peptide LL-37 is a multifunctional modulator of innate immune responses. *J. Immunol*, **169**: 3883-3891.

Seebah, S., Suresh, A., Zhuo, S., Choong, Y.H., Chua, H., Chuon, D., Beuerman, R. and Verma, C. (2007). Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides. *Nucleic Acids Research*, **35**: 265-268.

Selsted, M. E., Novotny, M. J., Morris, W. L., Tang, Y. Q., Smith, W. and Cullor, J. S. (1992). Indolicidin, a novel bactericidal tridecapeptide amide from neutrophils. *J. Biol. Chem.*, **267**: 4292-4295.

Sepkowitz, K. A. (June 2001). AIDS-the first 20 years. *N. Engl. J. Med.*, **344** (23): 1764-72.

Shai, Y. (2002). From innate immunity to de-novo designed anti-microbial peptides. *Curr. Pharm. Des.*, **8**: 715-725.

Shi, J. Ross, C. R., Chengappa, M. M., Sylte, M. J., McVey, D. S. and Blecha, F. (1996). Antibacterial activity of a synthetic peptide (PR-26) derived from PR-39, a proline-arginine-rich neutrophil antimicrobial peptide. *Antimicrob. Agents Chemother.*, **40**: 115-121.

Shinnar, A. E. et al. (1996) in Peptides; *Chemistry and Biology. Proc. 14th Am. Peptide Symp.* (eds Kaumaya, P. and Hodges), 189-191 (Mayflower Scientific, Leiden, 1996).

Siegfried, N., Muller, M., Deeks, J. J. and Volmink, J. (2009). Siegfried, Nandi. ed. Male circumcision for prevention of heterosexual acquisition of HIV in men. *Cochrane database of systematic reviews (Online)* (2): CD003362.

Siegfried, N., van der Merwe, L., Brocklehurst, P. and Sint, T. T. (2011). Siegfried, Nandi. ed. Antiretrovirals for reducing the risk of mother-to-child transmission of HIV infection". *Cochrane database of systematic reviews (Online)* (7): CD003510.

Silva, O. N., Mulder, K. C. L., Barbosa, A. E. A. D., Otero-Gonzalez, A. J., Lopez-Abarrategui, C., Rezende, T. M. B., Dias, S. C. and Franco, O. L. (2011). Exploring the pharmacological potential of promiscuous host-defense peptides: from natural screenings to biotechnological applications. *Frontiers in Microbiology*, **2** (232): 1-14.

Simmaco, M., Mignogna, G. and Barra, D. (1998). Antimicrobial peptides from amphibian skin: what do they tell us? *Biopolymers*, **47**: 435-450.

Soballe, P. W., Maloy, W. L., Myrnga, M. L., Jacob, L. S. and Herlyn, M. (1995). Experimental local therapy of human melanoma with lytic magainin peptides. *Int J Cancer*, **60** (2): 280-284.

Sørensen, O., Bratt, T., Johnsen, A. H., Madsen, M. T. and Borregaard, N. (1999). The human antibacterial cathelicidin, hCAP-18, is bound to lipoproteins in plasma. *J. Biol. Chem.*, **274**: 22445-22451.

Sousa, L. B., Mannis, M. J., Schwab, I. R., Cullor, J., Hosotani, H., Smith, W. and Jaynes, J. (1996). The use of synthetic Cecropin (D5C) in disinfecting contact lens solutions. *CLAO J*, **22** (2): 114-117.

Stahl, H. G. (1994). Ciba Foundation Symposium 186, Antimicrobial Peptides, John Wiley and Sons, *Chichester*, 27-53.

Statistics, S. A. (2011). Mid-year population estimates. *Statistics South Africa*, 1-18.

Stolp, B. and Fackler, O. T. (2011). How HIV takes advantage of the cytoskeleton in entry and replication. *Viruses*, **3** (4): 293-311.

Subbalakshmi, C. and Sitaram, N. (1998). Mechanism of antimicrobial action of indolicidin. *FEMS Microbiol. Lett.*, **160**: 91-96.

Sun, L., Finnegan, C. M., Kish-Catalone, T., Blumenthal, R., Garzino-Demo, P., La Terra Maggiore, G. M., Berrone, S., Kleinman, C., Wu, Z., Abdelwahab, S., Lu, W. and Garzino-Demo, A. (2005). Human beta-defensins suppress human immunodeficiency virus infection: potential role in mucosal protection. *J. Virol.*, **79** (22): 14318-14329.

Sundararajan, V. S., Gabere, M. N., Pretorius, A., Adam, S., Christoffels, A., Lehvaslaiho, M., Archer, J. A. C., Bajic, V. B. (2011). DAMPD: a manually curated antimicrobial peptides databases. *Nucleic Acid Research*, doi: 10.1093/nar.gkr1063.

Sutthent, R., Gaudart, N., Chokpaibulkit, K., Tanliang, N., Kanoksinsombath, C. and Chaisilwatana, P. (2003). p24 Antigen Detection Assay Modified with a Booster Step for Diagnosis and Monitoring of Human Immunodeficiency Virus Type 1 Infection. *Journal of clinical microbiology*, **41** (3): 1016-1022.

Tam, J. P., Lu, Y. A., Yang, J. L., Chiu, K. W. (1999). An unusual structural motif of antimicrobial peptides containing end-to-end macrocycle and cystine-knot disulfides. *Proc. Natl. Acad. Sci. USA*, **96**: 8913-8918.

Tam, J. P., Wu, C. and Yang, J. L. (2000). Membranolytic selectivity of cystine-stabilized cyclic protegrins. *Eur. J. Biochem.*, **267**: 3289-3300.

Tamamura, H., Ikoma, R., Niwa, M., Funakoshi, S., Murakami, T. and Fujii, N. (1993). Antimicrobial activity and conformation of tachyplesin I and its analogs. *Chem Pharm Bull* (Tokyo). **41**: 978-980.

Tamamura, H., Murakami, T., Masuda, M., Otaka, A., Takada, W., Ibuka, T., Nakashima, H., Waki, M., Matsumoto, A., Yamamoto, N., et al. (1994). Structure-activity relationships of an anti-HIV peptides, T22. *Biochem Biophys Res Commun*, **205** (3): 1729-1735.



Tamamura, H., Murakami, T., Horiuchi, S., Sugihara, K., Otaka, A., Takada, W., Ibuka, T., Waki, M., Yamamoto, N. and Fujii, N. (1995). Synthesis of protegrin-related peptides and their antibacterial and anti-human immunodeficiency virus activity. *Chemical and Pharmaceutical Bulletin*, **43** (5): 853-858.

Thomas, S., Karnik, S., Barai, R. S., Jayaraman, V. K. and Idicula-Thomas, S. (2009). CAMP: a useful resource for research on antimicrobial peptides. *Nucleic Acids Research*, **38**: D774-D780.

Torrent, M., Andreu, D., Nogués, V. M. and Boix, E. (2011). Connecting Peptide Physicochemical and Antimicrobial Properties by a Rational Prediction Model. *PLoS ONE*, **6** (2): e16968.

Torrent, M., Di Tommaso, P., Pulido, D., Nogués, M. V., Notredame, C., Boix, E., Andreu, D. (2012). AMPA: an automated web server for prediction of protein antimicrobial regions. *Bioinformatics*, **28** (1): 130-131.

Torrent, M., Nogués, M. V. and Boix, E. (2012). Discovering New In Silico Tools for Antimicrobial Peptide Prediction. *Current Drug Targets*, **13**: 1148-1157.

Torrent, M., Nogués, V. M. and Boix, E. (2009). A theoretical approach to spot active regions in antimicrobial proteins. *BMC Bioinformatics*. **10**: 373.

Tossi, A. and Sandri, L. (2002). Molecular Diversity in Gene-Encoded, Cationic Antimicrobial Polypeptides. *Current Pharmaceutical Design*, **8**: 743-761.

Tovchigrechko, A. and Vakser, I. L. (2006). GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res.*, **34**: W310-W314.

UNAIDS (2011). World AIDS Day Report. pg. 1-10.

UNAIDS (May 18, 2012). The quest for an HIV vaccine.

Uniprot, C. (2009). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research* **38** (Database issue): D142-D148.

Valerio, Jr. L. G. (2009). *In silico* toxicology for the pharmaceutical sciences. *Toxicol. Appl. Pharmacol.*, **241**: 356-370.

van't Hof, W., Veerman, E. C. I., Helmerhorst, E. J. and Amerongen A. V. N. (2001). Antimicrobial peptides: properties and applicability. *Biol. Chem*, **382**: 597-619.

VanCompernelle, S. E., Taylor, R. J., Oswald-Richter, K., Jiang, J., Youree, B. E., Bowie, J. H., Tyler, M. J., Conlon, J. M., Wade, D., Aiken, C., Dermody, T. S., KewalRamani, V. N. Rollins-Smith L. A. and Unutmaz, D. (2005). Antimicrobial Peptides from Amphibian Skin Potently Inhibit Human Immunodeficiency Virus Infection and Transfer of Virus from Dendritic Cells to T cells. *J. Virol.*, **79** (18): 11598-11606.

Vizioli, J. and Salzet, M. (2002). Antimicrobial peptides from animals: focus on invertebrates. *Trends Pharmacol. Sci.*, **23**: 494-496.

Wachinger, M., Kleinschmidt, A., Winder, D., von Pechmann, N., Ludvigsen, A., Neumann, M., Holle, R., Salmons, B., Erfle, V. and Brack-Werner, R. (1998). Antimicrobial peptides melittin and cecropin inhibit replication of human immunodeficiency virus 1 by suppressing viral gene expression. *Journal of General Virology*, **79**: 731-740.

Wachinger, M., Saermark, T. and Erfle, V. (1992). Influence of amphipathic peptides on the HIV-1 production in persistently infected T lymphoma cells. *FEBS Lett.*, **309**: 235-241.

Wade, D. and Englund, J. (2002). Synthetic antibiotic peptides database. *Protein and Peptide Letters*, **9** (1): 53-57.

Wang, C. K. L., Colgrave, M. L., Gustafson, K. R., Ireland, D. C., Goransson, U. and Craik, D. J. (2008). Anti-HIV Cyclotides from the Chinese Medicinal Herb *Viola yedoensis*. *J. Nat. Prod.*, **71**: 47-52.

Wang, C. K., Kaas, Q., Chiche, L. and Craik, D. J. (2008). CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Research*, **36**: C206-C210.

Wang, G., Buckheit, K. W., Biswajit Mishra, B., Lushnikova, T. and Buckheit, Jr. R. W. (2011). *De Novo* Design of Antiviral and Antibacterial Peptides with Varying Loop Structures. *J. AIDS Clinic Res.*, S2:003. doi:10.4172/2155-6113.S2-003.

Wang, G., Li, X. and Wang, Z. (2009). APD2: the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Research*, **37**: D933-937.

Wang, G., Watson, K. M. and Buckheit, Jr. R. W. (2008). Anti-Human Immunodeficiency Virus Type 1 Activities of Antimicrobial Peptides Derived from Human and Bovine Cathelicidins. *Antimicrobial Agents and Chemotherapy*, **52** (9): 3438-3440.

Wang, G., Watson, K. M., Peterkofsky A. and Buckheit Jr. R. W. (2010). Identification of Novel Human Immunodeficiency Virus Type 1-Inhibitory Peptides Based on the Antimicrobial Peptide Database. *Antimicrob. Agents Chemother.*, **54** (3): 1343. DOI: 10.1128/AAC.01448-09.

Wang, H. and Ng, T. B. (2000). Ginkbilobin, a novel antifungal protein from Ginkgo biloba seeds with sequence similarity to embryo-abundant protein. *Biochemical and Biophysical Research Communications*, **279**: 407-411.

Wang, H. and Ng, T. B. (2001). Novel Antifungal Peptides from Ceylon Spinach Seeds. *Biochemical and Biophysical Research Communications*, **288**: 765-770.

Wang, H. and Ng, T. B. (2002). Ascalin, a new anti-fungal peptide with human immunodeficiency virus type 1 reverse transcriptase-inhibiting activity from shallot bulbs. *Peptides*, **23**: 1025-1029.

Wang, H. and Ng, T. B. (2002). Isolation of cicadin, a novel and potent antifungal peptide from dried juvenile cicadas. *Peptides*, **23**: 7-11.

Wang, H. and Ng, T.B. (2005). Isolation of an antifungal thaumatin-like protein from kiwi fruits. *Phytochemistry*, **61**: 1-6.

Wang, H. X. and Ng, T. B. (2005). An antifungal peptide from the coconut. *Peptides.*, **26**: 2392-2396.

Wang, S., Xu, F. and Demirci, U. (2010). Advances in developing HIV-1 viral load assays for resource-limited settings. *Biotechnol Adv.*, **28** (6): 770-781.

Wang, W., Owen, S. M., Rudolph, D. L., Cole, A. M., Hong, T., Waring, A. J., Lal, R. B. and Lehrer, R. I. (2004). Activity of  $\alpha$ - and  $\theta$ -defensins against primary isolates of HIV-1. *J Immunol.*, **173**: 515-520.

Wang, Z. and Wang, G. (2004). APD: the Antimicrobial Peptide Database. *Nucleic Acids Research*, **32**: D590-D592.

Warrilow, D., Stenzel, D. and Harrich, D. (2007). Isolated HIV-1 core is active for reverse transcription. *Retrovirology*, **4**: 77. doi:10.1186/1742-4690-4-77.

Wei, L., Huang, E. S. and Altman, R. B. (1999). Are predicted structures good enough to preserve functional sites? *Structure*, **7**: 643-650.

Weiss, R. A. (1993). How does HIV cause AIDS? *Science*, **260** (5112): 1273-1279.

White, S. H., Wimley, W. C. and Selsted, M. E. (1995). Structure, function and membrane integration of defensins. *Current Opinion in Structure Biology*, **5**: 521-527.

Whitmore, L. and Wallace, B. A. (2004). The Peptaibol Database: a database for sequences and structures of naturally occurring peptaibols. *Nucleic Acids Research*, **32**: 593-594.

WHO (2011). HIV/AIDS.

Wiest, A., Grzegorski, D., Xu, B. W., Goulard, C., Rebuffat, S., Ebbole, D. J., Bodo B. and Kenerley C. (2002). Identification of peptaibols from *Trichoderma virens* and cloning of a peptaibol synthetase. *J. Biol. Chem.*, **277**: 20862-20868.

Wolfe, N. D., Switzer, W. M., Carr, J. K., Bhullar, V. B., Shanmugam, V., Tamoufe, U., Prosser, A. T., Torimiro, J. N., Wright, A., Mpoudi-Ngole, E., McCutchan, F. E., Birx, D. L., Folks, T. M., Burke, D. S. and Heneine, W. (2004). Naturally acquired simian retrovirus infections in Central African Hunters. *The Lancet*, **363** (9413): 932-937.

Wong, J. H. and Ng, T. B. (2003). Gymnin, a potent defensin-like antifungal peptide from the Yunnan bean (*Gymnocladus chinensis* Baill). *Peptides*, **24**: 963-968.

Wong, J. H. and Ng, T. B. (2005). Lunatusin, a trypsin-stable antimicrobial peptide from lima beans (*Phaseolus lunatus* L.). *Peptides*, **26**: 2086-2092.

Wong, J. H. and Tzi Bun Ng T. B. (2005). Sesquin, a potent defensin-like antimicrobial peptide from ground beans with inhibitory activities toward tumor cells and HIV-1 reverse transcriptase. *Peptides*, **26**: 1120-1126.

World Health Organization, (2010). Antiretroviral therapy for HIV infection in adults and adolescents: recommendations for a public health approach. pp. 19-20.

Wu, S., Skolnick, J. and Zhang, Y. (2007). *Ab initio* modeling of small proteins by iterative TASSER simulations. *BMG Biol.*, **5**: 17.

Wu, Z., Cocchi, F., Gentles, D., Ericksen, B., Lubkowski, J., Devico, A., Lehrer, R. I. and Lu, W. (2005). Human neutrophil alpha-defensin 4 inhibits HIV-1 infection in vitro. *FEBS Lett.*, **579** (1): 162-166.

Wua, Z., Cocchia, F., Gentlesa, D., Ericksena, B., Lubkowskib, J., DeVicoa, A., Lehrerc, R. I. and Lua, W. (2005). Human neutrophil  $\alpha$ -defensin 4 inhibits HIV-1 infection in vitro. *FEBS Letters*, **579**: 162-166.

Yang, L., Harroun, T. A., Weiss, T. M., Ding, L. and Huang, H. W. (2001). Barrel-stave model or toroidal model? A case study on melittin pores. *Biophys. J.*, **81**: 1475-1485.

Yeaman, M. R. and Yount, N. Y. (2003). Mechanisms of antimicrobial peptide action and resistance. *Pharmacol. Rev.*, **55** (1): 27-55.

Zapata, W., Rodriguez, B., Weber, J., Estrada, H., Quiñones-Mateu, M. E., Zimmermann, P. A., Lederman, M. M. and Rugeles, M. T. (2008). Increased Levels of Human Beta-Defensins mRNA in Sexually HIV-1 Exposed But Uninfected Individuals. *Current HIV Research*, **6**: 531-538.

Zasloff, M. (1992). Antibiotic peptides as mediators of innate immunity. *Curr. Opin. Immunol.*, **4**: 3-7.

Zasloff, M. (2002). Antimicrobial peptides of multicellular organisms. *Nature*, **415** (6870): 389-395.

Zhang, L. and Falla, T. J. (2006). Antimicrobial peptides: therapeutic potential. *Expert Opin. Pharmacother.*, **7**: 653-663.

Zhang, L., Benz, R. and Hancock, R. E. W. (1999). Influence of proline residues on the antibacterial and synergistic activities of alpha-helical peptides. *Biochemistry*, **38**: 8102-8111.

Zhang, L., Rozek, A. and Hancock, R. E. W. (2001). Interaction of cationic antimicrobial peptides with model membranes. *J. Biol Chem.*, **276**: 35714-35722.

Zhang, L., Scott, M. G., Yan, H., Mayer, L. D. and Hancock, R. E. W. (2000). Interaction of polyphemusin I and structural analogs with bacterial membranes, lipopolysaccharide, and lipid monolayers. *Biochemistry*, **39**: 14504-14514.

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

Zhang, Y. J., Wang, J. H., Lee, W. H., Wang, Q., Heng Liu, H., Zheng, Y. T. and Zhanga, Y. (2003). Molecular characterization of *Trimeresurus stejnegeri* venom L-amino acid oxidase with potential anti-HIV activity. *Biochemical and Biophysical Research Communications*, **309**: 598-604.

Zhong, D., Yang, M., Zhang, Y., Wu, Q., Wang, B., Li, C. and Yu, R. (1998). In vitro sensitivity of oral gram-negative bacteria to the bactericidal activity of defensins. *Hua Xi Kou Qiang Yi Xue Za Zhi*, **16** (1): 26-28.

Zhou, T., Xu, L., Dey, B., Hessel, A. J., Ryk, D. V., Xiang, S. H., Yang, X., Zhang, M. Y., Zwick, M. B., Arthos, J., Burton, D. R., Dimitrov, D. S., Sodroski, J., Wyatt, R., Nabel, G. J. and Kwong, P. D. (2007). Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature*, **445**: 732-737.

# Appendix

## SUPPLEMENTARY MATERIAL A

### Supplementary Tables for Chapter 2

**Table A.1:** Results of all the experimentally validated anti-HIV peptides with their references

ID	Name of AMPs	Species	Keywords	References
APD ID AP00013	Aurein 1.2	Frog	antibiotic, anticancer and anti-HIV	>Wang et al., 2010. Identification of Novel Human Immunodeficiency Virus Type 1-Inhibitory Peptides Based on the Antimicrobial Peptide Database
APD ID AP00139	Cecropin A	Insects	Anti-HIV	>Wachinger et al., 1998. Antimicrobial peptides melittin and cecropin inhibit replication of human immunodeficiency virus 1 by suppressing viral gene expression
APD ID AP00144	Magainin 2	African clawed frog ( <i>Xenopus laevis</i> )	AntiHIV, Antifungal, antiviral, anticancer, anti-parasites and Antibacterial	>VanCompernelle et al., 2005. Antimicrobial Peptides from Amphibian Skin Potently Inhibit Human Immunodeficiency Virus Infection and Transfer of Virus from Dendritic Cells to T cells >Pukala et al., 2005. Host-defence peptides from the glandular secretions of amphibians: structure and activity
APD ID AP00146 SWISS-PROT ID P01501	Melittin	Insects ( <i>Apis mellifera</i> )	Antibacterial, Antiviral	>Wachinger et al., 1998. Antimicrobial peptides melittin and cecropin inhibit replication of human immunodeficiency virus 1 by suppressing viral gene expression
APD ID AP00150	Indolicidin	Cow ( <i>Bovine neutrophils</i> )	Antiviral	>Robinson, Jr. et al., 1998. Anti-HIV-1 activity of indolicidin, an antimicrobial peptide from neutrophils
APD ID AP00157	Dermaseptin-S1	South America Sauvage's leaf frog <i>Phyllomedusa sauvagii</i>	Antibacterial, Antifungal, Antiviral, antiHIV and anti-parasites	>Van Compernelle et al., 2005. Antimicrobial Peptides from Amphibian Skin Potently Inhibit Human Immunodeficiency Virus Infection and Transfer of Virus from Dendritic Cells to T Cells >Pukala et al., 2005. Host-defence peptides from the glandular secretions of amphibians: structure and activity
APD ID AP00160 SWISS-PROT	Dermaseptin-S4	South America Sauvage's leaf frog <i>Phyllomedusa</i>	Antibacterial, Antifungal, Antiviral, antiHIV and	>Lorin et al., 2005. The antimicrobial peptide Dermaseptin S4 inhibits HIV-1 infectivity in vitro



ID P80280		<i>sauvagii</i>	anti-parasites	
APD ID AP00176	HNP-1	alpha defensin, Lectin	AntiHIV, Antifungal and Antibacteria	>Zhang et al., 2002. Contribution of human alpha-defensin 1, 2 and 3 to the anti-HIV-1 activity of CD8 antiviral factor. >Chang and Klotman , 2004. Defensins: Natural Anti-HIV peptides. >Wang et al., 2004. Activity of alpha- and theta-defensins against primary isolates of HIV-1
AP00177	HNP-2	alpha defensin, Lectin	AntiHIV, Antifungal and Antibacteria	>Zhang et al., 2002. Contribution of human alpha-defensin 1, 2 and 3 to the anti-HIV-1 activity of CD8 antiviral factor. >Chang and Klotman , 2004. Defensins: Natural Anti-HIV peptides. >Wang et al., (2004). Activity of alpha- and theta-defensins against primary isolates of HIV-1
AP00178	HNP-3	alpha defensin, Lectin	AntiHIV, Antifungal and Antibacteria	>Zhang et al., 2002. Contribution of human alpha-defensin 1, 2 and 3 to the anti-HIV-1 activity of CD8 antiviral factor. >Chang and Klotman , 2004. Defensins: Natural Anti-HIV peptides. >Wang et al., (2004). Activity of alpha- and theta-defensins against primary isolates of HIV-1
APD ID AP00179	HNP-4	alpha defensin, Lectin	AntiHIV	>Chang and Klotman , 2004. Defensins: Natural Anti-HIV peptides. >Wu et al., 2005. Human neutrophil $\alpha$ -defensin 4 inhibits HIV-1 infection in vitro
SWISS-PROT ID P84868	Sesquin	Ground beans Seeds, ( <i>Vigna sesquipedalis</i> )	Antifungal, Anti HIV, antibacterial and anticancer	>Wong and Ng, 2005. Sesquin, a potent defensin-like antimicrobial peptide from ground beans with inhibitory activities toward tumor cells and HIV-1 reverse transcriptase
PUBMED ID 11835991 SWISS-PROT ID P83080	Maximin 1	large-webbed bell Toad ( <i>Bombina maxima</i> )	Antibacterial Antifungal and AntiHIV	>Lai et al., 2001. Antimicrobial peptides from skin secretions of Chinese red belly toad <i>Bombina maxima</i>
PUBMED ID 11835991 SWISS-PROT ID P83082	Maximin 3	large-webbed bell Toad ( <i>Bombina maxima</i> )	Antibiotic Antimicrobial and AntiHIV	>Lai et al., 2001. Antimicrobial peptides from skin secretions of Chinese red belly toad <i>Bombina maxima</i>
PUBMED ID 11835991 SWISS-PROT ID P83083	Maximin 4	large-webbed bell Toad ( <i>Bombina maxima</i> )	Antibiotic Antimicrobial and AntiHIV	>Lai et al., 2001. Antimicrobial peptides from skin secretions of Chinese red belly toad <i>Bombina maxima</i>
PUBMED ID 11835991 SWISS-PROT ID P83084	Maximin 5	large-webbed bell Toad ( <i>Bombina maxima</i> )	Antibiotic Antimicrobial and AntiHIV	>Lai et al., 2001. Antimicrobial peptides from skin secretions of Chinese red belly toad <i>Bombina maxima</i>

P56871	Circulin-A	Plant ( <i>Chassalia parvifolia</i> )	Antibiotic Antimicrobial	>Daly et al., 1999. Solution structure by NMR of circulin A: a macrocyclic knotted peptide having anti-HIV activity. >Derua et al., 1996. Analysis of the disulfide linkage pattern in circulin A and B, HIV-inhibitory macrocyclic peptides. >Gustafson et al., 1994. Circulins A and B: novel HIV-inhibitor macrocyclic peptide from tropical tree <i>Chassalia parvifolia</i>
P14213	Tachyplesin-I	Japanese horseshoe crab ( <i>Tachypleus tridentatus</i> )	Antibiotic Antimicrobial	>Morimoto et al., 1991. Inhibitory effect of Tachyplesin I on the proliferation of human immunodeficiency virus in vitro
Q6WP39	L-amino-acid oxidase	Chinese green tree viper ( <i>Trimeresurus stejnegeri</i> )	Antibiotic Antimicrobial Oxidoreductase Toxin	>Zhang et al., 2003. Molecular characterization of <i>Trimeresurus stejnegeri</i> venom L-amino acid oxidase with potential anti-HIV activity
SWISS-PROT ID P83171	Ginkbilobin (GNL)	maidenhair tree ( <i>Ginkgo biloba</i> )	Antifungal, antibacterial and antiviral	>Wang and Ng, 2000. Ginkbilobin, a novel antifungal protein from <i>Ginkgo biloba</i> seeds with sequence similarity to embryo-abundant protein.
SWISS-PROT ID P83186	Alpha-basrubrin	Plant ( <i>Basella alba</i> )	Antifungal, Antiviral	>Wang and Ng, 2001. Novel Antifungal Peptides from Ceylon Spinach Seeds
SWISS-PROT ID P83187	Beta-basrubrin	Plant ( <i>Basella alba</i> )	Antifungal, Antiviral	>Wang and Ng, 2001. Novel Antifungal Peptides from Ceylon Spinach Seeds
SWISS-PROT ID P83282	Cicadin	Insect ( <i>Cicada flammata</i> )	Antifungal, Antiviral	>Wang and Ng, 2002. Isolation of cicadin, a novel and potent antifungal peptide from dried juvenile cicadas
SWISS-PROT ID P84071.1	Ascalin	Plant: shallot. ( <i>Allium cepa</i> var. <i>aggregatum</i> )	Antifungal, Antiviral	>Wang and Ng, 2002. Ascalin, a new antifungal peptide with human immunodeficiency virus type 1 reverse transcriptase-inhibiting activity from shallot bulbs
SWISS-PROT ID P84200	Gymnin	Plant ( <i>Gymnocladus chinensis</i> )	Antifungal, Antiviral	>Wong and Ng, 2003. Gymnin, a potent defensin-like antifungal peptide from the Yunnan bean ( <i>Gymnocladus chinensis</i> Bail)
SWISS-PROT ID P84785.1	Coccinin	Plant ( <i>Phaseolus coccineus</i> )	Antifungal, Antiviral	>Ngai and Ng, 2004. Coccinin, an antifungal peptide with antiproliferative and HIV-1 reverse transcriptase inhibitory activities from large scarlet runner beans.
SWISS-PROT ID P84797	Agrocybin	Fungi ( <i>Agrocybe cylindracea</i> )	Antibacterial, Antifungal, Antiviral	>Ngai, Zhao and Ng, 2004. Agrocybin, an antifungal peptide from the edible mushroom <i>Agrocybe cylindracea</i>
SWISS-PROT ID P84637	Cycloviolin-A	Plant ( <i>Leonia cymosa</i> )	Antiviral and antiHIV	>Ireland et al., 2007. Cyclotides as Natural Anti-HIV Agents >Hallock et al., 2000. Cycloviolins A-D, Anti-HIV Macrocyclic Peptides from

				<i>Leonia cymosa</i>
SWISS-PROT ID P84638	Cycloviolin-B	Plant ( <i>Leonia cymosa</i> )	Antiviral and antiHIV	>Ireland et al., 2007. Cyclotides as Natural Anti-HIV Agents >Hallock et al., 2000. Cycloviolins A-D, Anti-HIV Macrocylic Peptides from <i>Leonia cymosa</i>
SWISS-PROT ID P84639	Cycloviolin-C	Plant ( <i>Leonia cymosa</i> )	Antiviral and antiHIV	>Ireland et al., 2007. Cyclotides as Natural Anti-HIV Agents >Hallock et al., 2000. Cycloviolins A-D, Anti-HIV Macrocylic Peptides from <i>Leonia cymosa</i>
SWISS-PROT ID P84640	Cycloviolin-D	Plant ( <i>Leonia cymosa</i> )	Antiviral and antiHIV	>Ireland et al., 2007. Cyclotides as Natural Anti-HIV Agents >Hallock et al., 2000. Cycloviolins A-D, Anti-HIV Macrocylic Peptides from <i>Leonia cymosa</i>
SWISS-PROT ID P84645.1	Palicourein	Plant ( <i>Palicourea condensate</i> )	Antiviral and antiHIV	>Ireland et al., 2007. Cyclotides as Natural Anti-HIV Agents >Bokesch et al., 2001. A Novel Anti-HIV Macrocylic Peptide from <i>Palicourea condensate</i>
GI 66361400	Vhl-1	Plant ( <i>Viola hederacea</i> )	Antiviral and antiHIV	>Chen et al., 2005. Isolation and characterization of novel cyclotides from <i>Viola hederaceae</i> : solution structure and anti-HIV activity of vhl-1, a leaf-specific expressed cyclotide
SWISS-PROT ID P84641	Circulin-C	tropical tree ( <i>Chassalia parviflora</i> )	Antiviral and antiHIV	>Ireland et al., 2007. Cyclotides as Natural Anti-HIV Agents >Gustafson et al., 2000. New Circulin Macrocylic Polypeptides from <i>Chassalia parvifolia</i>
SWISS-PROT ID P84642	Circulin-D	tropical tree ( <i>Chassalia parviflora</i> )	Antiviral and antiHIV	>Ireland et al., 2007. Cyclotides as Natural Anti-HIV Agents >Gustafson et al., 2000. New Circulin Macrocylic Polypeptides from <i>Chassalia parvifolia</i>
SWISS-PROT ID P84643	Circulin-E	tropical tree ( <i>Chassalia parviflora</i> )	Antiviral and antiHIV	>Ireland et al., 2007. Cyclotides as Natural Anti-HIV Agents >Gustafson et al., 2000. New Circulin Macrocylic Polypeptides from <i>Chassalia parvifolia</i>
SWISS-PROT ID P84644	Circulin-F	tropical tree ( <i>Chassalia parviflora</i> )	Antiviral and antiHIV	>Ireland et al., 2007. Cyclotides as Natural Anti-HIV Agents >Gustafson et al., 2000. New Circulin Macrocylic Polypeptides from <i>Chassalia parvifolia</i>
PUBMED ID 18081258	Cycloviolacin Y1	Plant ( <i>Viola yedoensis</i> )	Antiviral and antiHIV	>Ireland et al., 2007. Cyclotides as Natural Anti-HIV Agents >Wang et al., 2008. Anti-HIV Cyclotides from the Chinese Medicinal Herb <i>Viola</i>

				<i>yedoensis</i>
PUBMED ID 18081258	Cycloviolacin Y4	Plant ( <i>Viola yedoensis</i> )	Antiviral and antiHIV	>Ireland et al., 2007. Cyclotides as Natural Anti-HIV Agents >Wang et al., 2008. Anti-HIV Cyclotides from the Chinese Medicinal Herb <i>Viola yedoensis</i>
PUBMED ID 18081258	Cycloviolacin Y5	Plant ( <i>Viola yedoensis</i> )	Antiviral and antiHIV	>Ireland et al., 2007. Cyclotides as Natural Anti-HIV Agents >Wang et al., 2008. Anti-HIV Cyclotides from the Chinese Medicinal Herb <i>Viola yedoensis</i>
PUBMED ID 16308082	Coconut antifungal peptide	coconut	Antifungal, Antiviral	>Wang and Ng, 2005. An antifungal peptide from the coconut
SWISS-PROT ID P85175	Kalata-B8	Plant, African Herb. ( <i>Oldenlandia affinis</i> )	Antiviral and antiHIV	>Ireland et al., 2007. Cyclotides as Natural Anti-HIV Agents >Daly et al., 2006. Kalata B8, a novel antiviral circular protein, exhibits conformational flexibility in the cystine knot motif
SWISS-PROT ID P83958	Thaumat- like protein , Actc2	Kiwi “Yangtao” <i>Actinidia chinensis</i>	Antifungal, Antiviral	>Wang and Ng, 2002. Isolation of an antifungal thaumatin-like protein from kiwi fruits.
SWISS-PROT ID P83957	Thaumat- like protein	Chinese chinquapin “Kweilin chestnut” ( <i>Castanopsis chinensis</i> )	Antifungal, Antiviral	>Chu and Ng, 2003. Isolation of a large thaumatin-like antifungal protein from seeds of the Kweilin chestnut <i>Castanopsis chinensis</i> .
CBP00001  SWISS-PROT ID P56254	Kalata-B1	Plant, African Herb. ( <i>Oldenlandia affinis</i> )	Hemolytic, Anti-HIV	>Ireland et al., 2008. Cyclotides as natural anti-HIV agents >Daly NL et al., 2004. The role of the cyclic peptide backbone in the anti-HIV activity of the cyclotide kalata B1 > Wang et al., 2008. Anti-HIV Cyclotides from the Chinese Medicinal Herb <i>Viola yedoensis</i>
SWISS-PROT ID P56879	Circulin-B	Plant ( <i>Chassalia parviflora</i> )	Antibiotic Antimicrobial	>Derua R et al., 1996. Analysis of the disulfide linkage pattern in circulin-A and B, HIV-inhibitory macrocyclic peptides. >Gustafson et al., 1994. Circulins A and B: novel HIV-inhibitor macrocyclic peptide from tropical tree <i>Chassalia parvifolia</i> > Koltay et al., 2005. Structure of circulin- B and implications for antimicrobial activity of the cyclotides.
SWISS-PROT ID P03973	SLPI (secretory leukocyte protease	Human ( <i>Homo sapiens</i> )	Antibacterial, Antifungal, Antiviral	>McNeely et al., 1997. Inhibition of Human Immunodeficiency Virus Type 1 Infectivity by Secretory Leukocyte Protease Inhibitor Occurs Prior to Viral Reverse Transcription

	inhibitor)			>Lin et al., 2004. Salivary secretory leukocyte protease inhibitor increases in HIV infection
SWISS-PROT ID P14216	Polyphemusin II	Atlantic horseshoe crab ( <i>Limulus polyphemus</i> )	antibacterial, antiviral and antiHIV	>Miyata et al., 1989. Antimicrobial peptides, isolated from horseshoe crab hemocytes, tachyplesin II, and polyphemusins I and II: chemical structures and biological activity
SWISS-PROT ID P56226	Caerin 1.1	Green tree frog <i>Litoria splendida</i> , <i>Litoria rothii</i> and <i>Litoria caerulea</i>	Antibacterial, antiviral, anticancer and antiHIV	>VanCompernelle et al., 2005. Antimicrobial Peptides from Amphibian Skin Potently Inhibit Human Immunodeficiency Virus Infection and Transfer of Virus from Dendritic Cells to T cells
SWISS-PROT ID P81252	Caerin 1.9	Blue-thighed frog ( <i>Litoria chloris</i> )	Antibacterial, antifungal, anticancer and antiHIV	>VanCompernelle et al., 2005. Antimicrobial Peptides from Amphibian Skin Potently Inhibit Human Immunodeficiency Virus Infection and Transfer of Virus from Dendritic Cells to T cells
SWISS-PROT ID P56242	Caerin 4.1	Australian green tree frog ( <i>Litoria caerulea</i> )	Antibacterial, antiviral and antiHIV	>VanCompernelle et al., 2005. Antimicrobial Peptides from Amphibian Skin Potently Inhibit Human Immunodeficiency Virus Infection and Transfer of Virus from Dendritic Cells to T cells
SWISS-PROT ID P82066	Maculatin 1.1	Australian frog <i>Litoria genimaculata</i> and <i>Litoria eucnemis</i>	Antibacterial, antifungal, anticancer and antiHIV	>VanCompernelle et al., 2005. Antimicrobial Peptides from Amphibian Skin Potently Inhibit Human Immunodeficiency Virus Infection and Transfer of Virus from Dendritic Cells to T cells
SWISS-PROT ID P80711	Clavanin B	Sea squirt, tunicate ( <i>Styela clava</i> )	Antibacterial, antiviral and antiHIV	>Wang et al., 2010. Identification of Novel Human Immunodeficiency Virus Type 1-Inhibitory Peptides Based on the Antimicrobial Peptide Database
SWISS-PROT ID PDB ID: 1KJ6	Human beta defensin 3	Human defensin ( <i>Homo sapiens</i> )	Antibacterial, antifungal, anticancer, antiHIV and chemotactic	>Quiñones-Mateu et al., 2003. Human epithelial $\beta$ -defensins 2 and 3 inhibit HIV-1 replication >Zapata et al., 2008. Increased Levels of Human Beta-Defensins mRNA in Sexually HIV-1 Exposed But Uninfected Individuals >Chang and Klotman, 2004. Defensins: Natural anti-HIV peptides
SWISS-PROT ID P82042	Uperin 3.6	Australia Floodplain toadlet ( <i>Uperoleia inundata</i> and <i>Uperoleia mjobergii</i> )	Antibacterial, antiviral and antiHIV	>VanCompernelle et al., 2005. Antimicrobial Peptides from Amphibian Skin Potently Inhibit Human Immunodeficiency Virus Infection and Transfer of Virus from Dendritic Cells to T cells
SWISS-PROT	Uperin 7.1	Australia Brown tree frog	Antibacterial, antiviral and	>Wang et al., 2010. Identification of Novel Human Immunodeficiency Virus Type 1-

ID P82050		( <i>Litoria ewingi</i> )	antiHIV	Inhibitory Peptides Based on the Antimicrobial Peptide Database
SWISS-PROT ID PDB ID: 2KET	BMAP-27 (bovine myeloid antimicrobial peptide 27)	Cow ( <i>Bos Taurus</i> )	Antibacterial, antifungal, anticancer, antiviral and antiHIV	>Wang et al., 2008. Anti-Human Immunodeficiency Virus Type 1 Activities of Antimicrobial Peptides Derived from Human and Bovine Cathelicidins
SWISS-PROT ID P82422	Ponericin L2	Insects ( <i>Pachycondyla goeldii</i> )	Antibacterial, insects, antiviral and antiHIV	>Wang et al., 2010. Identification of Novel Human Immunodeficiency Virus Type 1-Inhibitory Peptides Based on the Antimicrobial Peptide Database
SWISS-PROT ID PDB ID: 1ZRV	Spinigerin	Insects ( <i>Pseudacanthoermes spiniger</i> )	Antibacterial, antifungal, antiviral and antiHIV	>Wang et al., 2010. Identification of Novel Human Immunodeficiency Virus Type 1-Inhibitory Peptides Based on the Antimicrobial Peptide Database
SWISS-PROT ID P82821	Ranatueringin-6	North America frog ( <i>Rana catesbeiana</i> )	Antibacterial, antiviral and antiHIV	>VanCompernelle (2005). Antimicrobial Peptides from Amphibian Skin Potentially Inhibit Human Immunodeficiency Virus Infection and Transfer of Virus from Dendritic Cells to T cells
MSWISS-PROT ID P82824	Ranatueringin-9	North America frog ( <i>Rana catesbeiana</i> )	Antibacterial, antiviral and antiHIV	>Wang et al., 2010. Identification of Novel Human Immunodeficiency Virus Type 1-Inhibitory Peptides Based on the Antimicrobial Peptide Database
SWISS-PROT ID PDB ID: 1HVZ	RTD-1 (rhesus theta defensin-1)	leukocytes, Rhesus Macaque ( <i>Macaca mulatta</i> )	Antibacterial, antifungal, antiviral and antiHIV	>Wang et al., 2004. Activity of $\alpha$ - and $\theta$ -Defensins against Primary Isolates of HIV-1
SWISS-PROT ID PDB ID: 1PGI	Protegrin 1	Pig	Antibacterial, antifungal, antiviral and antiHIV	>Tam, Wu and Yang, 2000. Membranolytic selectivity of cysteine-stabilized cyclic protegrins >Tamamura et al., 1995. Synthesis of protegrin-related peptides and their antibacterial and anti-human immunodeficiency virus activity
SWISS-PROT ID PDB ID: 1RKK	Polyphemusin I	Atlantic horseshoe crab ( <i>Limulus polyphemus</i> )	Antibacterial, antiviral and antiHIV	>Miyata et al., 1989. Antimicrobial peptides, isolated from horseshoe crab hemocytes, tachyplesin II, and polyphemusins I and II: chemical structures and biological activity
SWISS-PROT ID PDB ID: 1E4S	HBD-1 (Human beta-defensin 1)	keratinocytes; platelets ( <i>Homo sapiens</i> )	Antibacterial, anticancer, antiviral and antiHIV	>Zapata et al., 2008. Increased Levels of Human Beta-Defensins mRNA in Sexually HIV-1 Exposed But Uninfected Individuals
SWISS-PROT ID PDB ID: 2jos	Piscidin 3	Fish: hybrid striped bass ( <i>Morone saxatilis M. chrysops</i> )	Antibacterial, anticancer, antiviral and antiHIV	>Wang et al., 2010. Identification of Novel Human Immunodeficiency Virus Type 1-Inhibitory Peptides Based on the Antimicrobial Peptide Database

SWISS-PROT IDPDB ID: IMAG	Gramicidin A	soil bacterium ( <i>Bacillus brevis</i> )	Antibacterial, antivirus and antiHIV	>Bourinbaier and Coleman, 1997. The effect of gramicidin, a topical contraceptive and antimicrobial agent with anti-HIV activity, against herpes simplex viruses type 1 and 2 in vitro >Bourinbaier and Lee-Huang, 1994. Comparative in vitro study of contraceptive agents with anti-HIV activity: gramicidin, nonoxynol-9, and gossypol. >Bourinbaier, Krasinski and Borkowsky, 1994. Anti-HIV effect of gramicidin in vitro: potential for spermicide use
SWISS-PROT ID P15516	human Histatin 5	Human ( <i>Homo sapiens</i> )	Antibacterial, antifungal, antivirus and antiHIV	>Groot et al., 2006. Histatin 5-Derived Peptide with Improved Fungicidal Properties Enhances Human Immunodeficiency Virus Type 1 Replication by Promoting Viral Entry
SWISS-PROT ID PDB ID: 1FD3	Human beta defensin 2	skin, lung, trachea epithelia, and uterus, Homo sapiens	Antibacterial, antifungal, Chemotactic antivirus and antiHIV	> Quiñones-Mateu et al., 2003. Human epithelial $\beta$ -defensins 2 and 3 inhibit HIV-1 replication >Zapata et al., 2008. Increased Levels of Human Beta-Defensins mRNA in Sexually HIV-1 Exposed But Uninfected Individuals >Chang and Klotman, 2004. Defensins: Natural anti-HIV peptides
SWISS-PROT ID (APD ID: AP00532)	Lunatusin	Plant: Lima bean ( <i>Phaseolus lunatus L.</i> )	Antibacterial, antifungal, anticancer antivirus and antiHIV	>Wong and Ng, 2005. Lunatusin, a trypsin-stable antimicrobial peptide from lima beans ( <i>Phaseolus lunatus L.</i> )
SWISS-PROT ID (APD ID: AP00599)	Brevinin-2- related peptide	mink frog, ( <i>Rana septentrionalis</i> )	Antibacterial, antifungal, antivirus and antiHIV	>Wang et al., 2010. Identification of Novel Human Immunodeficiency Virus Type 1-Inhibitory Peptides Based on the Antimicrobial Peptide Database
SWISS-PROT ID (APD ID: AP00640)	Maculatin 1.3	Frog ( <i>Litoria eucnemis</i> )	Antibacterial, anticancer, antivirus and antiHIV	>Wang et al., 2010. Identification of Novel Human Immunodeficiency Virus Type 1-Inhibitory Peptides Based on the Antimicrobial Peptide Database
SWISS-PROT ID (APD ID: AP00663)	Esculentin-2P	Frog ( <i>Rana pipiens</i> )	Antibacterial, antivirus and antiHIV	>VanCompernelle (2005). Antimicrobial Peptides from Amphibian Skin Potently Inhibit Human Immunodeficiency Virus Infection and Transfer of Virus from Dendritic Cells to T cells >Goraya et al., 2000. Peptides with antimicrobial activity from four different families isolated from the skins of the North American frogs <i>Rana luteiventris</i> , <i>Rana berlandieri</i> and <i>Rana pipiens</i>
SWISS-PROT ID (APD ID: AP00706)	Dahlein 5.6	Australia frog ( <i>Litoria dahlia</i> )	Antibacterial, antivirus and antiHIV	>VanCompernelle (2005). Antimicrobial Peptides from Amphibian Skin Potently Inhibit Human Immunodeficiency Virus Infection and Transfer of Virus from Dendritic Cells to T cells
SWISS-PROT	RTD-2	Bone marrow,	Antibacterial,	>Wang et al., 2004. Activity of $\alpha$ - and $\theta$ -

ID (APD ID:AP00724)	(rhesus theta-defensin 2)	or blood leukocytes, rhesus monkey ( <i>Macaca mulatta</i> )	antifungal, antiviral and antiHIV	Defensins against Primary Isolates of HIV-1
SWISS-PROT ID (APD ID:AP00725)	RTD-3 (rhesus theta-defensin 3)	Bone marrow, or blood leukocytes, rhesus monkey ( <i>Macaca mulatta</i> )	Antibacterial, antifungal, antiviral and antiHIV	>Wang et al., 2004. Activity of $\alpha$ - and $\theta$ -Defensins against Primary Isolates of HIV-1
SWISS-PROT ID (APD ID:AP00764)	Dermaseptin-S9	South America, hylid frog: <i>Phyllomedusa sauvagei</i>	Antibacterial, chemotactic, antiviral and antiHIV	>Wang et al., 2010. Identification of Novel Human Immunodeficiency Virus Type 1-Inhibitory Peptides Based on the Antimicrobial Peptide Database
SWISS-PROT ID (APD ID:AP01137)	Siamycin I	Streptomyces strain AA6532	antiviral and antiHIV	>Lin et al., 1996. Characterization of Siamycin I, a Human Immunodeficiency Virus Fusion Inhibitor >Detlefsen et al., 1995. Siamycins I and II, new anti-HIV-1 peptides: II. Sequence analysis and structure determination of siamycin I.
SWISS-PROT ID (APD ID:AP01138)	NP-06	Streptomyces strain AA6532	antiviral and antiHIV	>Chokekijchai et al., 1995. NP-06: a novel anti-human immunodeficiency virus polypeptide produced by a Streptomyces species
SWISS-PROT ID (APD ID:AP01223)	Ascaphin-8	North America, costal frog ( <i>Ascaphus truei</i> )	Antibacterial, antifungal, anticancer, antiviral and antiHIV	>Wang et al., 2010. Identification of Novel Human Immunodeficiency Virus Type 1-Inhibitory Peptides Based on the Antimicrobial Peptide Database
SWISS-PROT ID (APD ID:AP01269)	Melectin	the Cleptoparasitic bee ( <i>Melecta albifrons</i> )	Antibacterial, antiviral and antiHIV	>Wang et al., 2010. Identification of Novel Human Immunodeficiency Virus Type 1-Inhibitory Peptides Based on the Antimicrobial Peptide Database
SWISS-PROT ID (APD ID:AP01271)	Palustrin-3AR	North American crawfish frog ( <i>Rana areolata</i> )	Antibacterial, antiviral and antiHIV	>VanCompernelle (2005). Antimicrobial Peptides from Amphibian Skin Potently Inhibit Human Immunodeficiency Virus Infection and Transfer of Virus from Dendritic Cells to T cells
SWISS-PROT ID (APD ID:AP01273)	Esculentin-1ARb	North American crawfish frog ( <i>Rana areolata</i> )	Antibacterial, antiviral and antiHIV	>VanCompernelle (2005). Antimicrobial Peptides from Amphibian Skin Potently Inhibit Human Immunodeficiency Virus Infection and Transfer of Virus from Dendritic Cells to T cells
SWISS-PROT ID (APD ID:AP01348)	Temporin-LTc	Anura: Ranidae, Chinese broad-folded frog ( <i>Hylarana latouchii</i> )	Antibacterial, antiviral and antiHIV	>Wang et al., 2010. Identification of Novel Human Immunodeficiency Virus Type 1-Inhibitory Peptides Based on the Antimicrobial Peptide Database
SWISS-PROT	Temporin-	Asia frog,	antiviral and	>Wang et al., 2010. Identification of Novel



ID (APD ID:AP01434)	PTa	( <i>Hylarana picturata</i> )	antiHIV	Human Immunodeficiency Virus Type 1-Inhibitory Peptides Based on the Antimicrobial Peptide Database
SWISS- PROT ID (PDB ID: 2K6O)	Cathelicidin: LL-37 (leucine leucine-37)	neutrophils, skin, sweat, <i>Homo sapiens</i> ; Also <i>Pan troglodytes</i>	Antibacterial, chemotactic, antiviral and antiHIV	>Wang et al., 2008. Anti-Human Immunodeficiency Virus Type 1 Activities of Antimicrobial Peptides Derived from Human and Bovine Cathelicidins
SWISS- PROT ID (APD ID:AP01064)	Cycloviolacin O13	Plant ( <i>Viola odorata</i> )	Antiparasites, antiviral and antiHIV	>Ireland et al., 2008. Cyclotides as natural anti-HIV agents > Ireland et al., 2006. A novel suite of cyclotides from <i>Viola odorata</i> : sequence variation and the implications for structure, function and stability
SWISS- PROT ID (APD ID:AP01065)	Cycloviolacin O14	Plant ( <i>Viola odorata</i> )	Antiparasites, antiviral and antiHIV	>Ireland et al., 2008. Cyclotides as natural anti-HIV agents > Ireland et al., 2006. A novel suite of cyclotides from <i>Viola odorata</i> : sequence variation and the implications for structure, function and stability
SWISS- PROT ID	Cycloviolacin O24	Plant ( <i>Viola odorata</i> )	Antiparasites, antiviral and antiHIV	>Ireland et al., 2008. Cyclotides as natural anti-HIV agents > Ireland et al., 2006. A novel suite of cyclotides from <i>Viola odorata</i> : sequence variation and the implications for structure, function and stability
SWISS- PROT ID P83836	cycloviolacin O12 Alternative name(s): varv peptide E	Plant ( <i>Viola tricolor</i> , <i>Viola arvensis</i> , <i>Viola baoshanensis</i> , <i>Viola yedoensis</i> and <i>Viola abyssinica</i> )	Anticancer, antiviral and antiHIV	>Wang et al., 2008. Anti-HIV Cyclotides from the Chinese Medicinal Herb <i>Viola yedoensis</i>
SWISS- PROT ID P01190	Alpha-MSH (alpha-melanocyte-stimulating hormone)	<i>Homo sapiens</i>	Antibacterial, antifungal, antiviral and antiHIV	>Barcellini et al., 2000. $\alpha$ -Melanocyte-stimulating hormone peptides inhibit HIV-1 expression in chronically infected promonocytic U1 cells and in acutely infected monocytes

**Table A.2:** List of all the single domains identified during the genome sequences scan of database by HMMER profiles and their classification according to their E-values

Name of the Species	Gene name	Sequence of putative AMPs	E-values
<i>Homo sapiens</i>	ENSP0000342082	CLRYKKPECQSDWQCPGKKRCCPDTCGIKCLDPVDTNPTRRRKPGKCPVITYGQCLMLNPPNFCEMDGQCKRDLKCCMGM	1.40E-54
<i>Danaus plexippus</i>	EHJ64257	KWKIFKKIEKVGRNVRDGIKAGPAVQVVQATSIAK	3.10E-15
<i>Bombyx mori</i>	BGIBMGA006280-TA	RWKLFFKIEKVGRNVRDGLIKAGPAIAVIGQAKSLGK	1.10E-13
<i>Bombyx mori</i>	BGIBMGA014285-TA	RWKLFFKIEKVGRNVRDGLIKAGPAIAVIGQAKSLGK	1.10E-13
<i>Danaus plexippus</i>	EHJ64256	KWKFFKIEKVGRNIRDGIKAGPAVQVVLGEAKAIGK	7.30E-13
<i>Danaus plexippus</i>	EHJ71827	RWKFLKIEKVGRKVRDGVKAGPAVGVVGGQATSIYK	3.00E-12
<i>Danaus plexippus</i>	EHJ68082	KWKPFKLEKIQRVRDGIKAGPAVQVVGEAAAILK	2.90E-11
<i>Danaus plexippus</i>	EHJ72632	KWKPFKLEKIQRVRDGIKAGPAVQVVGEAAAILK	2.90E-11
<i>Bombyx mori</i>	BGIBMGA000024-TA	RWKIFKKIEKMGRNIRDGIVKAGPAIEVLGSAKAIGK	2.70E-10
<i>Bombyx mori</i>	BGIBMGA000023-TA	RWKIFKKIEKMGRNIRDGIVKAGPAIEVLGSAKAIGK	2.70E-10
<i>Bombyx mori</i>	BGIBMGA000021-TA	RWKIFKKIEKMGRNIRDGIVKAGPAIEVLGSAKAIGK	2.70E-10
<i>Bombyx mori</i>	BGIBMGA000036-TA	RWKIFKKIEKMGRNIRDGIVKAGPAIEVLGSAKAIGK	2.70E-10
<i>Bombyx mori</i>	BGIBMGA000037-TA	RWKIFKKIEKMGRNIRDGIVKAGPAIEVLGSAKAIGK	2.70E-10
<i>Bombyx mori</i>	BGIBMGA000038-TA	RWKIFKKIEKMGRNIRDGIVKAGPAIEVLGSAKAIGK	2.70E-10
<i>Homo sapiens</i>	ENSP0000303532	CLKSGAICHVPFCPRRYKQIGTCGLPGTKCCKKP	3.50E-10
<i>Homo sapiens</i>	ENSP0000424598	CLKSGAICHVPFCPRRYKQIGTCGLPGTKCCKKP	3.50E-10
<i>Setaria italica</i>	Si023829m	TPGESCILIPCTAAIGCSCKDKVCY	3.30E-08
<i>Heliconius melpomene</i>	HMELO10650-PA	WNPFKELEKAGQVRDAIISAKPAVDVVGGQATAIK	8.60E-08
<i>Setaria italica</i>	Si024202m	IPGESCFIIPCVTAAIGCSQCQDRVCY	9.80E-08

<i>Litoria caerulea</i>	tr Q800R8 Q800R8_LITCE	GLFGILGSVAKHVLPVVPVIAEH	2.40E-07
<i>Setaria italica</i>	SI023827m	ISCGETCLMLPCFIQAIGCRCKNKICY	3.40E-06
<i>Heliconius melpomene</i>	HMEI008612-PA	RWKFWKVEHAGQNRDGIKAGPAVAVKCGKVSFVRLIR	4.40E-06
<i>Litoria caerulea</i>	tr Q800R9 Q800R9_LITCE	GLFSVLGSVAKHVPRVVPVIAEH	5.00E-06
<i>Danaus plexippus</i>	EH168081	WNPFELEKAGQRVDAISAAPAVEVVQGASSILK	6.30E-06
<i>Homo sapiens</i>	ENSP00000296435	CDKDNKRFBALLGDFFRKSKEKIGKFKRIVQRIKDFLRNL	6.80E-06
<i>Homo sapiens</i>	ENSP00000458149	CDKDNKRFBALLGDFFRKSKEKIGKFKRIVQRIKDFLRNL	6.80E-06
<i>Homo sapiens</i>	ENSP00000297439	CVSSGGQCLYSACPFTKIQTGYRGKAKCCK	4.20E-05
<i>Setaria italica</i>	SI023843m	IQCGQTCFWIPCLDLGCSCKDNICY	4.40E-05
<i>Bombyx mori</i>	BGIBMGA000020-TA	KRKVKIIEKIGRNVGGVITAGPAVVVVVQAAASVGM	6.80E-05
<i>Zea mays</i>	GRMZM2G032198_P01	ISCGESCVIIPCVSTLLGCRCENKLCV	0.0002
<i>Zea mays</i>	GRMZM2G374405_P01	VPCFESCVFVPCISSVVGCRCENNVCV	0.00065
<i>Amolops jingdongensis</i>	tr G3ETP8 G3ETP8_AMOJI	GLFSIFKTAAKFVGNLLKQAGK	1.30E-003
<i>Amolops jingdongensis</i>	tr G3ETP9 G3ETP9_AMOJI	GLFSIFKTAAKFVGNLLKQAGK	1.30E-003
<i>Amolops jingdongensis</i>	tr G3ETQ0 G3ETQ0_AMOJI	GLFSIFKTAAKFVGNLLKQAGK	1.30E-003
<i>Amolops jingdongensis</i>	tr G3ETP4 G3ETP4_AMOJI	GIFSLFKTAAKFVGNLLKEAGK	1.60E-003
<i>Amolops jingdongensis</i>	tr G3ETP6 G3ETP6_AMOJI	GIFSLFKTAAKFVGNLLKEAGK	1.60E-003
<i>Amolops jingdongensis</i>	tr G3ETP5 G3ETP5_AMOJI	GIFSLFKTAAKFVGNLLKEAGK	1.60E-003
<i>Amolops jingdongensis</i>	tr G3ETP7 G3ETP7_AMOJI	GIFSLFKTAAKFVGNLLKEAGK	1.60E-003
<i>Amolops mantzorum</i>	tr E1B242 E1B242_9NEOB	GIFSLIKTAAKFVGNLLKQAGK	2.40E-003
<i>Amolops lifanensis</i>	tr E1AZ79 E1AZ79_9NEOB	GIFSLIKTAAKFVGNLLKQAGK	2.40E-003

<i>Amolops loloensis</i>	tr C5H0D5 C5H0D5_9NEOB	GIFSLIKTAAKFVGNLLKQAGK	2.40E-003
<i>Rana ornativentris</i>	tr D1MYB8 D1MYB8_9NEOB	GLFNVPK GALKTAGKHVAGSLLNQ	5.00E-003
<i>Rana chensinensis</i>	tr F1AEM1 F1AEM1_9NEOB	GLLSVPKGVLTAGKNVAKNVAGS	6.30E-003
<i>Rana chensinensis</i>	tr F1AEM3 F1AEM3_9NEOB	GLLSVPKGVLTAGKNVAKNVAGS	6.30E-003
<i>Rana chensinensis</i>	tr F1AEM2 F1AEM2_9NEOB	GLLSVPKGVLTAGKNVAKNVAGS	6.30E-003
<i>Phyllomedusa sauvagei</i>	tr Q1EN15 Q1EN15_PHYSA	GLRSKIWLWVLLMIWQESNKFKKM	7.30E-003
<i>Rana amurensis</i>	tr A0AAR9 A0AAR9_RANAM	GLLSVPKGVLTAGKNVAGSLLDQ	9.30E-003
<i>Rana amurensis</i>	tr A0AAS0 A0AAS0_RANAM	GLFSVVKGVLTAGKNVAGSLLDQ	9.40E-003



## Appendix A

### Supplementary Figures for Chapter 2

```

Query HMM:  amphibians
Accession:  [none]
Description: [none]
           [HMM has been calibrated; E-values are empirical estimates]

Scores for complete sequences (score includes all domains):
Sequence      Description                      Score    E-value  N
-----
Maximin_4     12.0    0.0009  1
Maximin_3     10.3    0.0018  1
RANATUERIN2P  5.3     0.012   1

Parsed for domains:
Sequence      Domain  seq-f  seq-t    hmm-f  hmm-t    score  E-value
-----
Maximin_4     1/1     1     24 [.   1     24 []   12.0   0.0009
Maximin_3     1/1     5     27 .]   1     24 []   10.3   0.0018
RANATUERIN2P  1/1     1     24 [.   1     24 []   5.3    0.012

Alignments of top-scoring domains:
Maximin_4: domain 1 of 1, from 1 to 24: score 12.0, E = 0.0009
      *->Glldfikslakvvgkalvkaianh<-*
      G+ +++ s++k ++k+l+k+ a++
Maximin_4     1     GIGGVILSAGKAALKGLAKVLAEK     24

Maximin_3: domain 1 of 1, from 5 to 27: score 10.3, E = 0.0018
      *->Glldfikslakvvgkalvkaianh<-*
      +l+++k+++k+++k l+++ h
Maximin_3     5     KILSGLKTALKGAAKELAST-YLH     27

RANATUERIN2P: domain 1 of 1, from 1 to 24: score 5.3, E = 0.012
      *->Glldfikslakvvgkalvkaianh<-*
      Gl d++k +ak ++++ + + + +
RANATUERIN    1     GLMDTVKNVAKNLAGHMLDKLKCK     24
  
```

Figure A.1: HMMER Amphibians profile query results using the Amphibians testing set

```

Query HMM:  microorganisms
Accession:  [none]
Description: [none]
  [HMM has been calibrated; E-values are empirical estimates]


Scores for complete sequences (score includes all domains):
Sequence Description                               Score   E-value  N
-----
SiamycinI                                           34.7    3.5e-11  1

Parsed for domains:
Sequence Domain  seq-f seq-t      hmm-f hmm-t      score  E-value
-----
SiamycinI    1/1      4   21 .]      1   19 []      34.7  3.5e-11

Alignments of top-scoring domains:
SiamycinI: domain 1 of 1, from 4 to 21: score 34.7, E = 3.5e-11
          *->vgsCndlAqclYGavvalw<-*
          vgsCnd+A+c+Y avv++w
SiamycinI    4   VGSCNDFAGCGY-AVVCFW      21

```

Figure A.2: HMMER Microorganisms profile query results using the Microorganisms testing set



```

Query HMM:  vertebrates
Accession:  [none]
Description: [none]
  [HMM has been calibrated; E-values are empirical estimates]

Scores for complete sequences (score includes all domains):
Sequence Description                               Score   E-value  N
-----
RTD-2                                               11.2    0.00027  1

Parsed for domains:
Sequence Domain  seq-f seq-t      hmm-f hmm-t      score  E-value
-----
RTD-2    1/1      1   15 [.      1   16 []      11.2  0.00027

Alignments of top-scoring domains:
RTD-2: domain 1 of 1, from 1 to 15: score 11.2, E = 0.00027
          *->icacrRgPfcpcirtr<-*
          +c+crRg +c+c + r
RTD-2    1   RCLCRRG-VCRCLCRR      15

```

Figure A.3: HMMER Vertebrates profile query results using the Vertebrates testing set

Query HMM: defensins  
 Accession: [none]  
 Description: [none]  
 [HMM has been calibrated; E-values are empirical estimates]

Scores for complete sequences (score includes all domains):

Sequence	Description	Score	E-value	N
hBD-3		-1.5	0.0014	1
hNP-4		-9.1	0.012	1

Parsed for domains:

Sequence	Domain	seq-f	seq-t	hmm-f	hmm-t	score	E-value
hBD-3	1/1	11	44 ..	1	79 []	-1.5	0.0014
hNP-4	1/1	2	33 .]	1	79 []	-9.1	0.012

Alignments of top-scoring domains:

hBD-3: domain 1 of 1, from 11 to 44: score -1.5, E = 0.0014

```

*->ClksgkPECqsDWqCPGkkrClpdacGikCLDPvDtPnpkrkkPGkC
      C  g+                rC  +c                ++
hBD-3  11  CRVRGG-----RCAVLSC-----LPKEE----- 28
  
```

```

PvkyGtCLmLnPPnFCefegrvkRiakccrgl<-*
  
```

```

++G+                C  g                kccr
hBD-3  29  --QIGK-----CSTRG-----RKCCRRK      44
  
```

humanneutrophilpeptide-4 (HNP-4): domain 1 of 1, from 2 to 33: score -9.1, E = 0.012

```

*->ClksgkPECqsDWqCPGkkrClpdacGikCLDPvDtPnpkrkkPGkC
      C                C+ ++c                ++ +
humanneutr  2  CS-----CRLVFC-----RRTEL----- 14
  
```

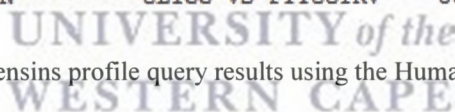
```

PvkyGtCLmLnPPnFCefegrvkRiakccrgl<-*
  
```

```

+ G                C  +g v  + cc
humanneutr  15  --RVGN-----CLIGG-VS-FTYCCTRV      33
  
```

Figure A.4: HMMER Human Defensins profile query results using the Human Defensins testing set



```

Query HMM:   fishs
Accession:   [none]
Description: [none]
[HMM has been calibrated; E-values are empirical estimates]

```

```

Scores for complete sequences (score includes all domains):
Sequence      Description                               Score    E-value  N
-----
PolyphemusinI 20.2      1.7e-06  1
TachyplesinI  9.0      0.00047  1

```

```

Parsed for domains:
Sequence      Domain  seq-f seq-t   hmm-f hmm-t   score  E-value
-----
PolyphemusinI 1/1      1   18 []    1   21 []    20.2  1.7e-06
TachyplesinI 1/1      1   17 []    1   21 []    9.0   0.00047

```

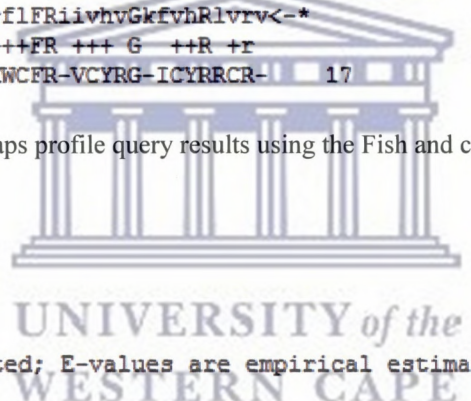
```

Alignments of top-scoring domains:
PolyphemusinI: domain 1 of 1, from 1 to 18: score 20.2, E = 1.7e-06
      *->frflFRiivhvGkfvhRlvrv<-*
      +r++FR +++ G f++R++r
Polyphemus      1      RRWCFR-VCYRG-FCYRKCR-      18

TachyplesinI: domain 1 of 1, from 1 to 17: score 9.0, E = 0.00047
      *->frflFRiivhvGkfvhRlvrv<-*
      +++FR +++ G ++R +r
Tachyplesi     1      -KWCFR-VCYRG-ICYRRCR-      17

```

Figure A.5: HMMER Fish and craps profile query results using the Fish and craps testing set



```

Query HMM:   insects
Accession:   [none]
Description: [none]
[HMM has been calibrated; E-values are empirical estimates]

Scores for complete sequences (score includes all domains):
Sequence      Description                               Score    E-value  N
-----
[no hits above thresholds]

Parsed for domains:
Sequence      Domain  seq-f seq-t   hmm-f hmm-t   score  E-value
-----
[no hits above thresholds]

Alignments of top-scoring domains:
[no hits above thresholds]

```

Figure A.6: HMMER Insects profile query results using the Insects testing set



Query HMM: plants  
 Accession: [none]  
 Description: [none]  
 [HMM has been calibrated; E-values are empirical estimates]

Scores for complete sequences (score includes all domains):

Sequence	Description	Score	E-value	N
circulinC		56.4	1.2e-16	1
cycloviolinD		55.7	1.9e-16	1
circulinB		54.3	4.9e-16	1
cycloviolacinY5		44.4	4.8e-13	1
cycloviolacinO14		30.1	9.7e-09	1
kalataB8		26.5	1.2e-07	1

Parsed for domains:

Sequence	Domain	seq-f	seq-t	hmm-f	hmm-t	score	E-value
circulinC	1/1	2	30 .]	1	35 []	56.4	1.2e-16
cycloviolinD	1/1	2	30 .]	1	35 []	55.7	1.9e-16
circulinB	1/1	3	31 .]	1	35 []	54.3	4.9e-16
cycloviolacinY5	1/1	2	30 .]	1	35 []	44.4	4.8e-13
cycloviolacinO14	1/1	4	31 .]	1	35 []	30.1	9.7e-09
kalataB8	1/1	4	31 .]	1	35 []	26.5	1.2e-07

Alignments of top-scoring domains:

```

circulinC: domain 1 of 1, from 2 to 30: score 56.4, E = 1.2e-16
      *->ipCgesCvfipaaCitpvlGCsCkdlrnkvCyarN<-*
      ipCgesCvfip  Cit+v GCsCk  +kvCy rN
circulinC  2  IPCGESCVFIP--CITSVAGCSCK---SKVCY-RN  30

cycloviolinD: domain 1 of 1, from 2 to 30: score 55.7, E = 1.9e-16
      *->ipCgesCvfipaaCitpvlGCsCkdlrnkvCyarN<-*
      +pCgesCvfip  Ci++++GCsCk  nkvcy rN
cycloviolin  2  EPCGESCVFIP--CISAAIGCSCK---NKVCY-RN  30

circulinB: domain 1 of 1, from 3 to 31: score 54.3, E = 4.9e-16
      *->ipCgesCvfipaaCitpvlGCsCkdlrnkvCyarN<-*
      ipCgesCvfip  Ci+  lGCsCk  nkvcy rN
circulinB  3  IPCGESCVFIP--CISTLLGCsCK---NKVCY-RN  31

cycloviolacinY5: domain 1 of 1, from 2 to 30: score 44.4, E = 4.8e-13
      *->ipCgesCvfipaaCitpvlGCsCkdlrnkvCyarN<-*
      ipC+esCv+ip  C++ + +GCsC+  +kvCy N
cycloviolacin  2  IPCAESCVWIP--CTVIALVGCSCS---DKVCY--N  30

cycloviolacinO14: domain 1 of 1, from 4 to 31: score 30.1, E = 9.7e-09
      *->ipCgesCvfipaaCitpvlGCsCkdlrnkvCyarN<-*
      + CgesC+++  C+tp  GCsC+  +++C +N
cycloviolacin  4  PACGESCFKGGK--CYTP--GCSCS--KYPLCA-KN  31

kalataB8: domain 1 of 1, from 4 to 31: score 26.5, E = 1.2e-07
      *->ipCgesCvfipaaCitpvlGCsCkdlrnkvCyarN<-*
      +Cge+C ++  C+t  GC+C  ++ vC+ ++
kalataB8  4  LNCGETCLLGT--CYTT--GCTCN--KYRVCT-KD  31
  
```

Figure A.7: HMMER Plants profile query results using the Plants testing set

GLAM2scan  
Version 1064

```
glam2scan -o plants_result.glam2 p plants.glam2 plants_query.fasta
```

```
*****  
circulinC          1 GIPCGESC VFIPCIITSVAGCSCSKSKVCYRN 30 + 95.7  
*****  
cycloviolinD      1 GFPCGESC VFIPCISAAIGCSCKNKVCYRN 30 + 92.7  
*****  
circulinB         1 GVIPCGESC VFIPCISTLLGCSCKNKVCYRN 31 + 85.1  
*****  
cycloviolacinY5   1 GIPCAESC VWIPCTVTALVGCSCSDKVCY.N 30 + 69.7  
*****  
cycloviolacin014  2 SIPACGESC FKGKCYT.P.GCSCSKYPLCAKN 31 + 60.9  
*****  
kalataB8         2 SVLNCGET CLLGTCYT.T.GCTCNKYRVCTKD 31 + 53.2
```

Figure A.8: GLAM2 Plants profile query results using the Plants testing set

GLAM2scan  
Version 1064

```
glam2scan -o vertebrates_result.glam2 p vertebrates.glam2 vertebrates_query.fasta
```

```
*****  
RTD-2             1 RCLCRRGVC RCLCRR 15 + 47.2  
UNIVERSITY of the  
WESTERN CAPE
```

Figure A.9: GLAM2 Vertebrates profile query results using the Vertebrates testing set

GLAM2scan  
Version 1064

```
glam2scan -o microorganisms_result.glam2 p microorganisms.glam2  
microorganisms_query.fasta
```

```
*****  
RP71955          1 CLGIGSCNDFAGCGYAVVCFW 21 + 75.8
```

Figure A.10: GLAM2 Microorganisms profile query results using the Microorganisms testing set

```

GLAM2scan
Version 1064

glam2scan -o fishs_result.glam2 p fishs.glam2 fishs_query.fasta

*****
TachyplestinI      4 FRVCYRGICY..RRCR 17 + -2.82

*****
PolyphemusinI     1 RRWCFR.VCYRG.FCY 14 + -4.68

```

Figure A.11: GLAM2 Fish and crabs profile query results using the Fish and crabs testing set

```

GLAM2scan
Version 1064

glam2scan -o insects_result.glam2 p insects.glam2 insects_query.fasta

*****
Cicadin           26 VVKPN 30 + -0.509

```

Figure A.12: GLAM2 Insects profile query results using the Insects testing set

```

GLAM2scan
Version 1064

glam2scan -o amphibians_result.glam2 p amphibians.glam2 amphibians_query.fasta

*****
Maximin_3        4 GKILSGLKIALKGAAKELA 22 + 26.0

*****
RANATUERIN2P    1 GL.MDTVKNVAKNLAGHML 18 + 21.5

*****
Maximin_5        4 AKILGGVKTFFKGALKELA 22 + 20.6

*****
Maximin_4        1 GI.GGVLLSAGKAALKGLA 18 + 18.1

```

Figure A.13: GLAM2 Amphibians profile query results using the Amphibians testing set

```

GLAM2scan
Version 1064

glam2scan -o Human_defensin_result.glam2 p defensins.glam2 defensins_query.fasta

*****
hBD-3          16 GRCAVLSCLPKKEEQIGKCSTRGRKCC 41 + 36.8
*****
*****...****
humanneutrophilpeptide-4 2 CSCRLVFCRRIELRVGNCLIGGVSTFYCC 30 + 32.5

```

Figure A.14: GLAM2 Human Defensins profile query results using the Human Defensins testing set

```

HMM file:          insects.hmm [insects]
Sequence database: Heliconius_melpomene.fasta
per-sequence score cutoff: [none]
per-domain score cutoff: [none]
per-sequence Eval cutoff: <= 0.01
per-domain Eval cutoff: [none]
-----

Query HMM:  insects
Accession:  [none]
Description: [none]
  [HMM has been calibrated; E-values are empirical estimates]

Scores for complete sequences (score includes all domains):
Sequence      Description                               Score      E-value    N
-----
HMEL010650-PA pep:novel scaffold:Hmel1:HE671115:15532  37.1      8.6e-08   1
HMEL008612-PA pep:novel scaffold:Hmel1:HE671894:39461  31.4      4.4e-06   1

Parsed for domains:
Sequence      Domain  seq-f seq-t  hmm-f hmm-t  score  E-value
-----
HMEL010650-PA  1/1     25   60  ..    1   37 []  37.1  8.6e-08
HMEL008612-PA  1/1     26   65  ..    1   37 []  31.4  4.4e-06

Alignments of top-scoring domains:
HMEL010650-PA: domain 1 of 1, from 25 to 60: score 37.1, E = 8.6e-08
      *->KWKLFKKiEkvdkgiadkikllkkavAvlgklltllK<-*
      W  FK +Ek ++ ++d i+ +ktav vtg+++++K
      HMEL010650  25  -WNPFKELEKAGQRVRDAIISAKPAVDVVGQATAIIK  60

HMEL008612-PA: domain 1 of 1, from 26 to 65: score 31.4, E = 4.4e-06
      *->KWKLFKKiEkvdkgiadkikllkkavAv.lgkl..ltllK<-*
      +WK +KK+E+ +++i+d+i+++++tavAv+ gk++  1+
      HMEL008612  26  RWKFWKKVEHAGQNIRDGIIKAGPAVAVkCGKVsfVRLIR  65

```

Figure A.15: Results of the *Heliconius melpomene* genome query with the Insects training model

```

HMM file:                insects.hmm [insects]
Sequence database:       Danaus_plexippus.fasta
per-sequence score cutoff: [none]
per-domain score cutoff:  [none]
per-sequence Eval cutoff: <= 0.01
per-domain Eval cutoff:  [none]

```

```

-----
Query HMM:  insects
Accession:  [none]
Description: [none]
[HMM has been calibrated; E-values are empirical estimates]

```

Scores for complete sequences (score includes all domains):			
Sequence	Description	Score	E-value N
EHJ64257	pep:novel supercontig:DanPle_1.0:JH391103:80	62.2	3.1e-15 1
EHJ64256	pep:novel supercontig:DanPle_1.0:JH391103:40	54.3	7.3e-13 1
EHJ71827	pep:novel supercontig:DanPle_1.0:JH384507:85	52.3	3e-12 1
EHJ68082	pep:novel supercontig:DanPle_1.0:JH387124:69	49.0	2.9e-11 1
EHJ72632	pep:novel supercontig:DanPle_1.0:JH384044:71	49.0	2.9e-11 1
EHJ68081	pep:novel supercontig:DanPle_1.0:JH387124:20	31.3	6.3e-06 1

Parsed for domains:								
Sequence	Domain	seq-f	seq-t	hmm-f	hmm-t	score	E-value	
EHJ64257	1/1	27	63 ..	1	37 []	62.2	3.1e-15	
EHJ64256	1/1	27	63 .]	1	37 []	54.3	7.3e-13	
EHJ71827	1/1	27	63 ..	1	37 []	52.3	3e-12	
EHJ68082	1/1	26	62 ..	1	37 []	49.0	2.9e-11	
EHJ72632	1/1	26	62 ..	1	37 []	49.0	2.9e-11	
EHJ68081	1/1	25	60 ..	1	37 []	31.3	6.3e-06	

Alignments of top-scoring domains:

```

EHJ64257: domain 1 of 1, from 27 to 63: score 62.2, E = 3.1e-15
      *->KWKLFKKiEkvdkgiadkikllkkavAvlgklltllK<-*
      KWK+FKKiEkv+++++d+i+++++av v+g+++++K
      EHJ64257 27  KWKLFKKiEkVGRNVRDGIKAGPAVQVVGQATSIK 63

EHJ64256: domain 1 of 1, from 27 to 63: score 54.3, E = 7.3e-13
      *->KWKLFKKiEkvdkgiadkikllkkavAvlgklltllK<-*
      KWK FKKiEkv+++i+d+i+++++av vlg ++++ K
      EHJ64256 27  KWKFFKKiEkVGRNIRDGIKAGPAVQVVGGEAKAIGK 63

EHJ71827: domain 1 of 1, from 27 to 63: score 52.3, E = 3e-12
      *->KWKLFKKiEkvdkgiadkikllkkavAvlgklltllK<-*
      +WK KKiEkv+++++d+++++av v+g+++++ K
      EHJ71827 27  RWKFLKKiEkVGRKVRDGVKAGPAVGVVGQATSIYK 63

EHJ68082: domain 1 of 1, from 26 to 62: score 49.0, E = 2.9e-11
      *->KWKLFKKiEkvdkgiadkikllkkavAvlgklltllK<-*
      KWK FKK+Ek+++ ++d+i+++++av v+g + ++lK
      EHJ68082 26  KWKPFKKLEKIGQRVRDGIKAGPAVQVVGEEAAAILK 62

EHJ72632: domain 1 of 1, from 26 to 62: score 49.0, E = 2.9e-11
      *->KWKLFKKiEkvdkgiadkikllkkavAvlgklltllK<-*
      KWK FKK+Ek+++ ++d+i+++++av v+g + ++lK
      EHJ72632 26  KWKPFKKLEKIGQRVRDGIKAGPAVQVVGEEAAAILK 62

EHJ68081: domain 1 of 1, from 25 to 60: score 31.3, E = 6.3e-06
      *->KWKLFKKiEkvdkgiadkikllkkavAvlgklltllK<-*
      W FK +Ek ++ ++d i+ + +av v+g++ ++lK
      EHJ68081 25  -WNPFKLEKAGQRVDAIISAAPAVEVVGQASSILK 60

```

Figure A.16: Results of the *Danaus plexippus* genome query with the Insects training model

HMM file: defensins.hmm [defensins]  
 Sequence database: Homo\_sapiens.fasta  
 per-sequence score cutoff: [none]  
 per-domain score cutoff: [none]  
 per-sequence Eval cutoff: <= 0.01  
 per-domain Eval cutoff: [none]

Query HMM: defensins  
 Accession: [none]  
 Description: [none]  
 [HMM has been calibrated; E-values are empirical estimates]

Scores for complete sequences (score includes all domains):

Sequence	Description	Score	E-value	N
ENSP00000342082	pep:known chromosome:GRCh37:20:438808	195.5	1.4e-54	1
ENSP00000303532	pep:known chromosome:GRCh37:8:7752151	48.1	3.5e-10	1
ENSP00000424598	pep:known chromosome:GRCh37:8:7272382	48.1	3.5e-10	1
ENSP00000296435	pep:known chromosome:GRCh37:3:4826483	33.8	6.8e-06	1
ENSP00000458149	pep:known chromosome:GRCh37:3:4826486	33.8	6.8e-06	1
ENSP00000297439	pep:known chromosome:GRCh37:8:6728097	31.2	4.2e-05	1

Parsed for domains:

Sequence	Domain	seq-f	seq-t	hmm-f	hmm-t	score	E-value
ENSP00000342082	1/1	43	121 ..	1	79 []	195.5	1.4e-54
ENSP00000303532	1/1	31	64 .]	1	79 []	48.1	3.5e-10
ENSP00000424598	1/1	31	64 .]	1	79 []	48.1	3.5e-10
ENSP00000296435	1/1	128	167 ..	1	79 []	33.8	6.8e-06
ENSP00000458149	1/1	125	164 ..	1	79 []	33.8	6.8e-06
ENSP00000297439	1/1	37	68 .]	1	79 []	31.2	4.2e-05

Alignments of top-scoring domains:

ENSP00000342082: domain 1 of 1, from 43 to 121: score 195.5, E = 1.4e-54  
 \*->ClksgkPECqsDWqCPGkkrClpdacGikCLDPvDtPnpkrkkPGkC  
 Cl+++kPECqsDWqCPGkkrC+pd+cGikCLDPvDtPnp+r+kPGkC  
 ENSP000003 43 CLRYKKECQSDWQCPCGKRCPPDTCGKICLDPVDTNPTRRRKPGKC 89  
 PvkyGtCLmLnPPnFCefegrvkRiakccrgl<-\*  
 Pv+yG+CLmLnPPnFCe++g++kR++kcc++g+  
 ENSP000003 90 PVTYGOCLMLNPPNFCFEMDQCKRDLKCCMGM 121

ENSP00000303532: domain 1 of 1, from 31 to 64: score 48.1, E = 3.5e-10  
 \*->ClksgkPECqsDWqCPGkkrClpdacGikCLDPvDtPnpkrkkPGkC  
 Clksg+ +C+p++c p+r+k  
 ENSP000003 31 CLKSGA-----ICHPVFC-----PRRYK----- 48  
 PvkyGtCLmLnPPnFCefegrvkRiakccrgl<-\*  
 ++Gt C+++g +kcc+++  
 ENSP000003 49 --QIGT-----CGLPG-----TKCCKKP 64

ENSP00000424598: domain 1 of 1, from 31 to 64: score 48.1, E = 3.5e-10  
 \*->ClksgkPECqsDWqCPGkkrClpdacGikCLDPvDtPnpkrkkPGkC  
 Clksg+ +C+p++c p+r+k  
 ENSP000004 31 CLKSGA-----ICHPVFC-----PRRYK----- 48  
 PvkyGtCLmLnPPnFCefegrvkRiakccrgl<-\*  
 ++Gt C+++g +kcc+++  
 ENSP000004 49 --QIGT-----CGLPG-----TKCCKKP 64

ENSP00000296435: domain 1 of 1, from 128 to 167: score 33.8, E = 6.8e-06  
 \*->ClksgkPECqsDWqCPGkkrClpdacGikCLDPvDtPnpkrkkPGkC  
 C k k + +l+d++ +k+k+  
 ENSP000002 128 CDKDNK-----RFALLGDF-----RKSKE----- 147

```

PvkyGtCLmLnPPnFCefegrvkRiakccrgl<-*
k+G+          ef+++v+Ri++++r+l
ENSP000002    148 --KIGK-----EFKRIVQRIKDFLRNL    167

ENSP00000458149: domain 1 of 1, from 125 to 164: score 33.8, E = 6.8e-06
*->ClksgkPECqsDWqCPGkkrClpdacGikCLDPvDtPnpkrkkPGkC
C k k          + +l+d++          +k+k+
ENSP000004    125 CDKDNK-----RFALLGDF-----RKSKE----- 144

PvkyGtCLmLnPPnFCefegrvkRiakccrgl<-*
k+G+          ef+++v+Ri++++r+l
ENSP000004    145 --KIGK-----EFKRIVQRIKDFLRNL    164

ENSP00000297439: domain 1 of 1, from 37 to 68: score 31.2, E = 4.2e-05
*->ClksgkPECqsDWqCPGkkrClpdacGikCLDPvDtPnpkrkkPGkC
C++sg+        +Cl++ac          p+++k
ENSP000002    37  CVSSGG-----QCLYSAC-----PIFTK----- 54

PvkyGtCLmLnPPnFCefegrvkRiakccrgl<-*
++Gt          C++++          akcc+
ENSP000002    55 --IQGI-----CYRGK-----AKCCK--    68

```

Figure A.17: Results of the *Homo sapiens* genome query with the Humans Defensins training model

```

HMM file:                plants.hmm [plants]
Sequence database:       Zea mays.fasta
per-sequence score cutoff: [none]
per-domain score cutoff: [none]
per-sequence Eval cutoff: <= 0.01
per-domain Eval cutoff: [none]
-----
Query HMM:  plants
Accession:  [none]
Description: [none]
[HMM has been calibrated; E-values are empirical estimates]

Scores for complete sequences (score includes all domains):
Sequence      Description      Score    E-value    N
-----
GRMZM2G032198_P01 pep:known chromosome:AGPv3:3:180544 28.8    0.0002    1
GRMZM2G374405_P01 pep:novel chromosome:AGPv3:3:180853 27.0    0.00065   1

Parsed for domains:
Sequence      Domain  seq-f  seq-t    hmm-f  hmm-t    score  E-value
-----
GRMZM2G032198_P01  1/1    56    82 ..    1    35 []    28.8  0.0002
GRMZM2G374405_P01  1/1    57    83 ..    1    35 []    27.0  0.00065

Alignments of top-scoring domains:
GRMZM2G032198_P01: domain 1 of 1, from 56 to 82: score 28.8, E = 0.0002
*->ipCgesCvfipaaCitpvlGCsCkdlrnkvCyarN<-*
i+CgesCv ip C++ 1GC C+ nk+C
GRMZM2G032 56  ISGESCVIIP--CVSTLLGCRCE---NKLCV--- 82

GRMZM2G374405_P01: domain 1 of 1, from 57 to 83: score 27.0, E = 0.00065
*->ipCgesCvfipaaCitpvlGCsCkdlrnkvCyarN<-*
+pC esCvf p Ci++v+GC C+ n+vC
GRMZM2G374 57  VPCFESC VFVP--CISSVVGCRCE---NNVCV--- 83

```

Figure A.18: Results of the *Zea mays* genome query with the Plants training model

```

HMM file:          plants.hmm [plants]
Sequence database: Setaria_italica.fasta
per-sequence score cutoff: [none]
per-domain score cutoff: [none]
per-sequence Eval cutoff: <= 0.01
per-domain Eval cutoff: [none]

```

```

Query HMM:  plants
Accession:  [none]
Description: [none]
[HMM has been calibrated; E-values are empirical estimates]

```

Scores for complete sequences (score includes all domains):

Sequence	Description	Score	E-value	N
Si023829m	pep:novel scaffold:JGIv2.0:scaffold_3:43583	40.2	3.3e-08	1
Si024202m	pep:novel scaffold:JGIv2.0:scaffold_3:43490	38.6	9.8e-08	1
Si023827m	pep:novel scaffold:JGIv2.0:scaffold_3:43483	33.5	3.4e-06	1
Si023843m	pep:novel scaffold:JGIv2.0:scaffold_3:43604	29.8	4.4e-05	1

Parsed for domains:

Sequence	Domain	seq-f	seq-t	hmm-f	hmm-t	score	E-value
Si023829m	1/1	55	81	1	35	40.2	3.3e-08
Si024202m	1/1	53	79	1	35	38.6	9.8e-08
Si023827m	1/1	56	82	1	35	33.5	3.4e-06
Si023843m	1/1	55	79	1	35	29.8	4.4e-05

Alignments of top-scoring domains:

```

Si023829m: domain 1 of 1, from 55 to 81: score 40.2, E = 3.3e-08
      *->ipCgesCvfipaaCitpvlGCsCkdlrnkvCyarN<-*
      +pCgesC ip Cit+++GCsCk +kvCy
      Si023829m 55 TPCGESCILIP--CITAAIGCSCK---DKVCY--- 81

Si024202m: domain 1 of 1, from 53 to 79: score 38.6, E = 9.8e-08
      *->ipCgesCvfipaaCitpvlGCsCkdlrnkvCyarN<-*
      ipCgesC+ ip C+t+++GCsC+ + vCy
      Si024202m 53 IPCGESCFILIP--CVTAAIGCSQ---DRVCY--- 79

Si023827m: domain 1 of 1, from 56 to 82: score 33.5, E = 3.4e-06
      *->ipCgesCvfipaaCitpvlGCsCkdlrnkvCyarN<-*
      i+Cge+C + p C+ +++GC Ck nk Cy
      Si023827m 56 ISCGETCLMLP--CFIQAIGCRCK---NKICY--- 82

Si023843m: domain 1 of 1, from 55 to 79: score 29.8, E = 4.4e-05
      *->ipCgesCvfipaaCitpvlGCsCkdlrnkvCyarN<-*
      i+Cg++C++ip C+ GCsCk ++ Cy
      Si023843m 55 IQCGQTCFWIP--CLDL--GCSCCK---DNICY--- 79

```

Figure A.19: Results of the *Setaria italica* genome query with the Plants training model



HMM file: insects.hmm [insects]  
 Sequence database: Bombyx\_mori.fasta  
 per-sequence score cutoff: [none]  
 per-domain score cutoff: [none]  
 per-sequence Eval cutoff: <= 0.01  
 per-domain Eval cutoff: [none]

-----  
 Query HMM: insects  
 Accession: [none]  
 Description: [none]  
 [HMM has been calibrated; E-values are empirical estimates]

Scores for complete sequences (score includes all domains):

Sequence	Description	Score	E-value	N
BGIBMGA006280-TA	pep:known scaffold:Bmor1:nscaf2852:1	56.9	1.1e-13	1
BGIBMGA014285-TA	pep:known scaffold:Bmor1:scaffold108	56.9	1.1e-13	1
BGIBMGA000024-TA	pep:known scaffold:Bmor1:nscaf1071:3	45.6	2.7e-10	1
BGIBMGA000023-TA	pep:known scaffold:Bmor1:nscaf1071:3	45.6	2.7e-10	1
BGIBMGA000021-TA	pep:known scaffold:Bmor1:nscaf1071:3	45.6	2.7e-10	1
BGIBMGA000036-TA	pep:known scaffold:Bmor1:nscaf1071:3	45.6	2.7e-10	1
BGIBMGA000037-TA	pep:known scaffold:Bmor1:nscaf1071:3	45.6	2.7e-10	1
BGIBMGA000038-TA	pep:known scaffold:Bmor1:nscaf1071:3	45.6	2.7e-10	1
BGIBMGA000020-TA	pep:novel scaffold:Bmor1:nscaf1071:3	27.7	6.8e-05	1

Parsed for domains:

Sequence	Domain	seq-f	seq-t	hmm-f	hmm-t	score	E-value
BGIBMGA006280-TA	1/1	27	63	1	37	56.9	1.1e-13
BGIBMGA014285-TA	1/1	27	63	1	37	56.9	1.1e-13
BGIBMGA000024-TA	1/1	27	63	1	37	45.6	2.7e-10
BGIBMGA000023-TA	1/1	27	63	1	37	45.6	2.7e-10
BGIBMGA000021-TA	1/1	27	63	1	37	45.6	2.7e-10
BGIBMGA000036-TA	1/1	27	63	1	37	45.6	2.7e-10
BGIBMGA000037-TA	1/1	27	63	1	37	45.6	2.7e-10
BGIBMGA000038-TA	1/1	27	63	1	37	45.6	2.7e-10
BGIBMGA000020-TA	1/1	27	63	1	37	27.7	6.8e-05

Alignments of top-scoring domains:

BGIBMGA006280-TA: domain 1 of 1, from 27 to 63: score 56.9, E = 1.1e-13  
 \*->KWKLFKKiEkvdkgiadkikllkkavAvlgklltllK<-\*  
 +WKLFKKiEkv+++++d+++++a Av+g+++1 K  
 BGIBMGA006 27 RWKLFKKIEKVGGRNVRDGLIKAGPAIAVIGQAKSLGK 63

BGIBMGA014285-TA: domain 1 of 1, from 27 to 63: score 56.9, E = 1.1e-13  
 \*->KWKLFKKiEkvdkgiadkikllkkavAvlgklltllK<-\*  
 +WKLFKKiEkv+++++d+++++a Av+g+++1 K  
 BGIBMGA014 27 RWKLFKKIEKVGGRNVRDGLIKAGPAIAVIGQAKSLGK 63

BGIBMGA000024-TA: domain 1 of 1, from 27 to 63: score 45.6, E = 2.7e-10  
 \*->KWKLFKKiEkvdkgiadkikllkkavAvlgklltllK<-\*  
 +WK+FKKiEk+++++i+d+i +++++ vlg +++ K  
 BGIBMGA000 27 RWKIFKKIEKMGRNIRDGIVKAGPAIEVLGSAKAIGK 63

BGIBMGA000023-TA: domain 1 of 1, from 27 to 63: score 45.6, E = 2.7e-10  
 \*->KWKLFKKiEkvdkgiadkikllkkavAvlgklltllK<-\*  
 +WK+FKKiEk+++++i+d+i +++++ vlg +++ K  
 BGIBMGA000 27 RWKIFKKIEKMGRNIRDGIVKAGPAIEVLGSAKAIGK 63

BGIBMGA000021-TA: domain 1 of 1, from 27 to 63: score 45.6, E = 2.7e-10  
 \*->KWKLFKKiEkvdkgiadkikllkkavAvlgklltllK<-\*  
 +WK+FKKiEk+++++i+d+i +++++ vlg +++ K  
 BGIBMGA000 27 RWKIFKKIEKMGRNIRDGIVKAGPAIEVLGSAKAIGK 63



Parsed for domains:

Sequence	Domain	seq-f	seq-t	hmm-f	hmm-t	score	E-value
tr Q800R8 Q800R8_LITCE	1/1	50	73	..	1 24 []	32.4	2.4e-07
tr Q800R9 Q800R9_LITCE	1/1	50	73	..	1 24 []	28.0	5e-06
tr G3ETP8 G3ETP8_AMOJI	1/1	40	62	..	1 24 []	19.9	0.0013
tr G3ETP9 G3ETP9_AMOJI	1/1	40	62	..	1 24 []	19.9	0.0013
tr G3ETQ0 G3ETQ0_AMOJI	1/1	40	62	..	1 24 []	19.9	0.0013
tr G3ETP4 G3ETP4_AMOJI	1/1	44	66	..	1 24 []	19.7	0.0016
tr G3ETP6 G3ETP6_AMOJI	1/1	44	66	..	1 24 []	19.7	0.0016
tr G3ETP5 G3ETP5_AMOJI	1/1	44	66	..	1 24 []	19.7	0.0016
tr G3ETP7 G3ETP7_AMOJI	1/1	44	66	..	1 24 []	19.7	0.0016
tr E1B242 E1B242_9NEOB	1/1	40	62	..	1 24 []	19.1	0.0024
tr E1AZ79 E1AZ79_9NEOB	1/1	40	62	..	1 24 []	19.1	0.0024
tr C5H0D5 C5H0D5_9NEOB	1/1	40	62	..	1 24 []	19.1	0.0024
tr D1MYB8 D1MYB8_9NEOB	1/1	41	64	..	1 24 []	18.0	0.005
tr F1AEM1 F1AEM1_9NEOB	1/1	42	65	..	1 24 []	17.7	0.0063
tr F1AEM3 F1AEM3_9NEOB	1/1	42	65	..	1 24 []	17.7	0.0063
tr F1AEM2 F1AEM2_9NEOB	1/1	42	65	..	1 24 []	17.7	0.0063
tr Q1EN15 Q1EN15_PHYSA	1/1	46	69	.]	1 24 []	17.5	0.0073
tr AOAAR9 AOAAR9_RANAM	1/1	39	62	..	1 24 []	17.1	0.0093
tr AOAAS0 AOAAS0_RANAM	1/1	41	64	..	1 24 []	17.1	0.0094

Alignments of top-scoring domains:

tr Q800R8 Q800R8_LITCE: domain 1 of 1, from 50 to 73: score 32.4, E = 2.4e-07	*->G1ldfikslakvvgkalvkaianh<-*	
	GL++++s+ak+v++++v+tiath	
tr Q800R8  50	GLFGILGSVAKHVLPHVVPVIAEH	73
tr Q800R9 Q800R9_LITCE: domain 1 of 1, from 50 to 73: score 28.0, E = 5e-06	*->G1ldfikslakvvgkalvkaianh<-*	
	GL++++s+ak+v++ +v+tiath	
tr Q800R9  50	GLFSVLGSVAKHVVPVIAEH	73
tr G3ETP8 G3ETP8_AMOJI: domain 1 of 1, from 40 to 62: score 19.9, E = 0.0013	*->G1ldfikslakvvgkalvkaianh<-*	
	GL++++k++ak+vgk+l+k a +	
tr G3ETP8  40	GLFSIFKTAAKFVGKDLLKQ-AGK	62
tr G3ETP9 G3ETP9_AMOJI: domain 1 of 1, from 40 to 62: score 19.9, E = 0.0013	*->G1ldfikslakvvgkalvkaianh<-*	
	GL++++k++ak+vgk+l+k a +	
tr G3ETP9  40	GLFSIFKTAAKFVGKDLLKQ-AGK	62
tr G3ETQ0 G3ETQ0_AMOJI: domain 1 of 1, from 40 to 62: score 19.9, E = 0.0013	*->G1ldfikslakvvgkalvkaianh<-*	
	GL++++k++ak+vgk+l+k a +	
tr G3ETQ0  40	GLFSIFKTAAKFVGKDLLKQ-AGK	62
tr G3ETP4 G3ETP4_AMOJI: domain 1 of 1, from 44 to 66: score 19.7, E = 0.0016	*->G1ldfikslakvvgkalvkaianh<-*	
	G+++ +k++ak+vgk+l+k+ a +	
tr G3ETP4  44	GIFSLFKTAAKFVGKDLLKE-AGK	66
tr G3ETP6 G3ETP6_AMOJI: domain 1 of 1, from 44 to 66: score 19.7, E = 0.0016	*->G1ldfikslakvvgkalvkaianh<-*	
	G+++ +k++ak+vgk+l+k+ a +	
tr G3ETP6  44	GIFSLFKTAAKFVGKDLLKE-AGK	66
tr G3ETP5 G3ETP5_AMOJI: domain 1 of 1, from 44 to 66: score 19.7, E = 0.0016	*->G1ldfikslakvvgkalvkaianh<-*	
	G+++ +k++ak+vgk+l+k+ a +	
tr G3ETP5  44	GIFSLFKTAAKFVGKDLLKE-AGK	66
tr G3ETP7 G3ETP7_AMOJI: domain 1 of 1, from 44 to 66: score 19.7, E = 0.0016	*->G1ldfikslakvvgkalvkaianh<-*	
	G+++ +k++ak+vgk+l+k+ a +	
tr G3ETP7  44	GIFSLFKTAAKFVGKDLLKE-AGK	66

```

tr|E1B242|E1B242_9NEOB: domain 1 of 1, from 40 to 62: score 19.1, E = 0.0024
      *->G1ldfikslakvvgkalvkaianh<-*
      G+++ ik++ak+vgk+l+k a +
tr|E1B242| 40 G1FSLIKTAAKFVGNLLKQ-AGK 62

tr|E1AZ79|E1AZ79_9NEOB: domain 1 of 1, from 40 to 62: score 19.1, E = 0.0024
      *->G1ldfikslakvvgkalvkaianh<-*
      G+++ ik++ak+vgk+l+k a +
tr|E1AZ79| 40 G1FSLIKTAAKFVGNLLKQ-AGK 62

tr|C5H0D5|C5H0D5_9NEOB: domain 1 of 1, from 40 to 62: score 19.1, E = 0.0024
      *->G1ldfikslakvvgkalvkaianh<-*
      G+++ ik++ak+vgk+l+k a +
tr|C5H0D5| 40 G1FSLIKTAAKFVGNLLKQ-AGK 62

tr|D1MYB8|D1MYB8_9NEOB: domain 1 of 1, from 41 to 64: score 18.0, E = 0.005
      *->G1ldfikslakvvgkalvkaianh<-*
      G1++++k ++k +gk++++ +n
tr|D1MYB8| 41 GLFNVPFGALKTAGKHHVAGSLLNQ 64

tr|F1AEM1|F1AEM1_9NEOB: domain 1 of 1, from 42 to 65: score 17.7, E = 0.0063
      *->G1ldfikslakvvgkalvkaianh<-*
      G1l+++k ++k +gk+++k++a +
tr|F1AEM1| 42 GLLSVFKGVLTAGKNVAKNVAGS 65

tr|F1AEM3|F1AEM3_9NEOB: domain 1 of 1, from 42 to 65: score 17.7, E = 0.0063
      *->G1ldfikslakvvgkalvkaianh<-*
      G1l+++k ++k +gk+++k++a +
tr|F1AEM3| 42 GLLSVFKGVLTAGKNVAKNVAGS 65

tr|F1AEM2|F1AEM2_9NEOB: domain 1 of 1, from 42 to 65: score 17.7, E = 0.0063
      *->G1ldfikslakvvgkalvkaianh<-*
      G1l+++k ++k +gk+++k++a +
tr|F1AEM2| 42 GLLSVFKGVLTAGKNVAKNVAGS 65

tr|Q1EN15|Q1EN15_PHYSA: domain 1 of 1, from 46 to 69: score 17.5, E = 0.0073
      *->G1ldfikslakvvgkalvkaianh<-*
      G1++++i++++v++++t++++t++++t
tr|Q1EN15| 46 GLRSKIWLWVLLMIWQESNKFKKM 69

tr|A0AAR9|A0AAR9_RANAM: domain 1 of 1, from 39 to 62: score 17.1, E = 0.0093
      *->G1ldfikslakvvgkalvkaianh<-*
      G1l+++k ++k+vgk++++ ++
tr|A0AAR9| 39 GLLSVFKGVLTGKGVGNVAGSLLDQ 62

tr|A0AAS0|A0AAS0_RANAM: domain 1 of 1, from 41 to 64: score 17.1, E = 0.0094
      *->G1ldfikslakvvgkalvkaianh<-*
      G1++++k ++k+vgk++++ ++
tr|A0AAS0| 41 GLFSVVKGVLTGKGVGNVAGSLLDQ 64

```

Figure A.21: Results of the various amphibians' sequences from UniprotKB query with the Amphibians training model

```

HMM file: vertebrates.hmm [vertebrates]
Sequence database: Pelodiscus_sinensis.fasta
per-sequence score cutoff: [none]
per-domain score cutoff: [none]
per-sequence Eval cutoff: <= 0.01
per-domain Eval cutoff: [none]
-----

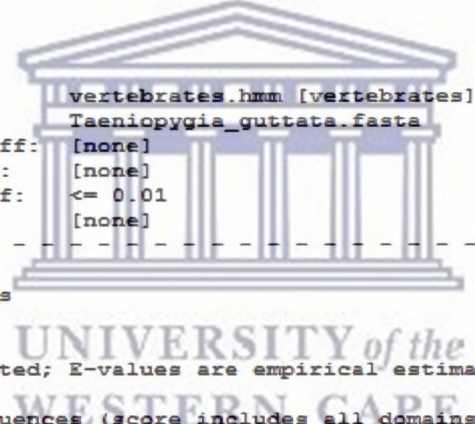
Query HMM: vertebrates
Accession: [none]
Description: [none]
[HMM has been calibrated; E-values are empirical estimates]

Scores for complete sequences (score includes all domains):
Sequence Description Score E-value N
-----
ENSPSIP00000012700 pep:novel scaffold:Pe1Sin_1.0:JH20 22.1 0.0046 4

Parsed for domains:
Sequence Domain seq-f seq-t hmm-f hmm-t score E-value
-----
ENSPSIP00000012700 4/4 139 153 .. 1 16 [] 6.9 20
ENSPSIP00000012700 1/4 10 25 .. 1 16 [] 5.1 48
ENSPSIP00000012700 2/4 64 79 .. 1 16 [] 5.1 48
ENSPSIP00000012700 3/4 118 133 .. 1 16 [] 5.1 48

```

Figure A.22: Results of the *Pelodiscus sinensis* genome query with the Vertebrates training model



```

HMM file: vertebrates.hmm [vertebrates]
Sequence database: Taeniopygia_guttata.fasta
per-sequence score cutoff: [none]
per-domain score cutoff: [none]
per-sequence Eval cutoff: <= 0.01
per-domain Eval cutoff: [none]
-----

Query HMM: vertebrates
Accession: [none]
Description: [none]
[HMM has been calibrated; E-values are empirical estimates]

Scores for complete sequences (score includes all domains):
Sequence Description Score E-value N
-----
ENSTGUP00000001760 pep:novel chromosome:taeGut3.2.4:2 42.3 3.2e-09 9

Parsed for domains:
Sequence Domain seq-f seq-t hmm-f hmm-t score E-value
-----
ENSTGUP00000001760 9/9 158 173 .. 1 16 [] 8.4 9.2
ENSTGUP00000001760 7/9 126 141 .. 1 16 [] 6.9 18
ENSTGUP00000001760 8/9 142 157 .. 1 16 [] 6.9 18
ENSTGUP00000001760 5/9 78 93 .. 1 16 [] 5.2 40
ENSTGUP00000001760 6/9 110 125 .. 1 16 [] 5.2 40
ENSTGUP00000001760 2/9 17 32 .. 1 16 [] 3.2 1e+02
ENSTGUP00000001760 4/9 62 77 .. 1 16 [] 3.2 1e+02
ENSTGUP00000001760 1/9 1 16 [ 1 16 [] 3.0 1.1e+02
ENSTGUP00000001760 3/9 33 48 .. 1 16 [] 0.4 3.7e+02

```

Figure A.23: Results of the *Taeniopygia guttata* genome query with the Vertebrates training model

```

HMM file:                vertebrates.hmm [vertebrates]
Sequence database:       Ictidomys_tridecemlineatus.fasta
per-sequence score cutoff: [none]
per-domain score cutoff: [none]
per-sequence Eval cutoff: <= 0.01
per-domain Eval cutoff: [none]
-----
Query HMM:   vertebrates
Accession:   [none]
Description: [none]
[HMM has been calibrated; E-values are empirical estimates]
Scores for complete sequences (score includes all domains):
Sequence          Description          Score    E-value    N
-----
ENSSTOP00000021787 pep:known_by_projection scaffold:s    23.4    0.0018    7
ENSSTOP00000018545 pep:known_by_projection scaffold:s    23.4    0.0018    7
ENSSTOP00000020387 pep:known_by_projection scaffold:s    23.4    0.0018    7

Parsed for domains:
Sequence          Domain  seq-f  seq-t    hmm-f  hmm-t    score  E-value
-----
ENSSTOP00000021787  6/7    4319  4334 ..     1    16 []    10.2    4.3
ENSSTOP00000018545  6/7    4314  4329 ..     1    16 []    10.2    4.3
ENSSTOP00000020387  6/7    4163  4178 ..     1    16 []    10.2    4.3
ENSSTOP00000021787  7/7    5020  5032 ..     1    16 []     4.9    50
ENSSTOP00000018545  7/7    5015  5027 ..     1    16 []     4.9    50
ENSSTOP00000020387  7/7    4864  4876 ..     1    16 []     4.9    50
ENSSTOP00000021787  1/7     456   469 ..     1    16 []     3.9    81
ENSSTOP00000018545  1/7     456   469 ..     1    16 []     3.9    81
ENSSTOP00000020387  1/7     333   346 ..     1    16 []     3.9    81
ENSSTOP00000021787  5/7    3956  3972 ..     1    16 []     3.1    1.1e+02
ENSSTOP00000018545  5/7    3949  3965 ..     1    16 []     3.1    1.1e+02
ENSSTOP00000020387  5/7    3800  3816 ..     1    16 []     3.1    1.1e+02
ENSSTOP00000021787  4/7    2694  2706 ..     1    16 []     0.7    3.6e+02
ENSSTOP00000018545  4/7    2687  2699 ..     1    16 []     0.7    3.6e+02
ENSSTOP00000020387  4/7    2538  2550 ..     1    16 []     0.7    3.6e+02
ENSSTOP00000021787  2/7     908   923 ..     1    16 []     0.5    3.8e+02
ENSSTOP00000018545  2/7     902   917 ..     1    16 []     0.5    3.8e+02
ENSSTOP00000020387  2/7     779   794 ..     1    16 []     0.5    3.8e+02
ENSSTOP00000021787  3/7    1874  1887 ..     1    16 []     0.1    4.7e+02
ENSSTOP00000018545  3/7    1868  1881 ..     1    16 []     0.1    4.7e+02
ENSSTOP00000020387  3/7    1745  1758 ..     1    16 []     0.1    4.7e+02

```

Figure A.24: Results of the *Ictidomys tridecemlineatus* genome query with the Vertebrates training mode

## SUPPLEMENTARY MATERIAL B

### Supplementary Tables and Figures for Chapter 3

**Table B.1:** The 10 best putative anti-HIV AMPs with respect of their E-values, the positive control and negative control used in the 3-D structure prediction by the I-TASSER server

Gene name	New appellation in the Chapter3	Number of amino acids residue	Sequences of the putative antimicrobial peptides
ENSP00000342082	Molecule1	79	CLR YKKPECOSDWQCPCGKKRCCPDTCGHKCLDPVDTNPTRRKPCKPVTYGGCLMLNPPNFCEMDGQCKRDLKCCMGM
EHJ64257	Molecule2	37	KWKIFKKIEKVGRNVRDGIHKAGPAVQVVGGQATSIK
BGIBMGA006280-TA	Molecule3	37	RWKLFKKIEKVGRNVRDGLIKAGPAIAVIGQAKSLGK
EHJ64256	Molecule4	37	KWKFFKKIEKVGRNIRDGIHKAGPAVQVLGEAKAIGK
EHJ71827	Molecule5	37	RWKELKKIEKVGRKVRDGVIIKAGPAVGVVGGQATSIYK
EHJ68082	Molecule6	37	KWKPFKKLEKIGQRVDRDGIHKAGPAVQVVGEAAAILK
BGIBMGA000024-TA	Molecule7	37	RWKIFKKIEKMGRIIRDGIHKAGPAIEVLGSAKAIGK
ENSP00000303532	Molecule8	34	CLKSGAICHVPFCPRRYKQIGTCGLPGTKCKCKP
Si023829m	Molecule9	29	TPCGESCILIPCITAAIGCCKDKVCY
HMEL010650-PA	Molecule10	36	WNPFKELEKAGQRVDAISAKPAVDVVGGQATAIK
	Kn2-7	13	FIKRIARLLRKKIF
	MucroporinS1	11	SLIGGLVSAFK

**Table B.2:** Sequences of the HIV Proteins used for the 3-D structures prediction by the I-TASSER server

	No amino acids residue	PDB ID	Sequences of the HIV proteins
<b>gp120</b>	317	2NXZ:A	EVVLVNVTENFNMMWKNDMVEQMHEDIISLWDQSLKPCVKLTLPLCVGAGSCNTSVITQACPKVSFEPIPIHYCAPAG FAILKCNNKTFNGTGPCTNVSTVQCTHGIRPVVSSQLLLNGSLAEEVVIRSVNFTDNAKTIIVQLNTSVEINCTGAG HCNIARAKWNNTLKQIASKLREQFGNNKTIFKQSSGGDPEIVTHWFCGGEFFYCNSTQLFNSTWFNSTWSTEGS NNTEGSDTITLPCRKIQIINMWQVKGAMYPPISSQIRCSSNITGLLLTRDGGNSNNESEIFRPGGGDMRDNRWSE LYK YKVVVKIE
<b>gp41</b>	70	1AIK:C	WMEWDREINNYTSLIHSLEESQNQQEKNEQELL
		1AIK:N	SGIVQQNNLLRAIEAQQLLQLTVWGIIKQLQARIL
<b>p24</b>	210	1E6I:P	VHQAISPRTLNAWVKVVEEKAFSPEVIPMFSALSEGATPQDLNMLNTVGGHQAAQMMLKETINEEAAEWDRVH PVHAGPIAPGQMRPRGSDHAGTITSLQEIQWMTNNPIPVEIYKRWILGLNKIVRMYSPSILDIRQGPKEPFRD YVDRFYKTLRAEQASQEVKNWMTETLLVQNAHPDCKTILKALGPAATLEEMMTACQG
<b>p17</b>	133	2HMX:A	HMGARASVLSGGELDKWEKIRLRPGKKQYKLLKHIVWASRELERFAVNPGLLETSEGCRCRQILGQLQPSLQGTSEE LRSLYNTIAVLYCVHQRIDVKTKEALDKIEEQNKSKKKAQQAADTGNNSQVSNQY



**Table B.3:** Results of the docking parameters calculations used for the HIV proteins-anti-HIV AMP interaction studies. The HIV proteins targeted as therapeutics tools interacting with the 10 best putative anti-HIV AMPs using the PatchDock docking algorithm

	gp120			gp41		
	Area (Å <sup>2</sup> )	ACE	Transformation coordinates	Area (Å <sup>2</sup> )	ACE	Transformation coordinates
Molecule1	2433.90	281.93	-0.94 0.61 -2.68 14.27 19.05 2.11	1949.30	-501.08	2.28 0.60 -0.05 25.81 10.97 12.71
Molecule2	1846.10	-479.45	0.12 -1.32 -1.74 -30.81 -6.64 6.38	1473.70	-126.11	-2.66 -0.62 -0.03 22.53 19.78 -3.82
Molecule3	1926.00	410.28	-1.31 0.66 1.01 -5.19 -9.03 -16.06	1407.90	-147.52	0.37 0.34 -2.87 24.22 17.06 13.64
Molecule4	1562.20	151.51	-0.72 0.32 0.91 -4.69 -29.60 -16.82	1291.10	243.94	2.52 0.29 3.02 11.31 8.84 -7.66
Molecule5	1766.40	-312.22	-2.27 0.54 -0.16 -23.88 -3.70 2.07	1486.40	-274.94	-0.73 -0.35 1.90 26.87 12.89 13.25
Molecule6	1890.50	239.95	0.50 -1.01 -1.90 12.99 -10.55 9.48	1437.00	-415.76	-0.36 0.70 0.69 26.09 12.31 13.61
Molecule7	1906.00	295.00	2.31 1.12 2.99 8.08 -27.26 11.81	1344.50	-163.99	-2.22 0.16 0.01 28.54 -0.00 5.27
Molecule8	1564.90	71.02	-1.23 0.07 1.76 9.50 5.77 -10.61	997.80	19.11	-2.49 -0.50 1.50 25.93 7.40 5.48
Molecule9	1324.00	-119.83	-1.56 0.20 -1.48 -12.14 -9.92 17.51	1013.60	-171.45	2.43 0.13 -0.73 32.82 15.47 -8.81
Molecule10	1916.20	-34.06	-2.33 0.66 -2.29 9.00 14.84 -0.33	1416.50	-35.04	-2.17 -0.35 -0.59 25.87 -0.37 14.59
+ve control (Kn2-7)	958.60	130.49	-1.58 -0.59 -1.11 21.86 -8.84 19.68	695.30	-201.59	0.35 -0.01 2.42 35.22 14.73 -0.12
-ve control (MucroporinS1)	903.40	58.53	-2.75 -0.33 -2.00 -7.20 -7.96 14.27	598.00	-155.49	1.24 0.01 -0.69 25.99 22.66 -9.28

ACE: Atomic Contact Energy

**Table B.4:** Results of the docking parameters calculations used for the HIV proteins-anti-HIV AMP interaction studies. The HIV proteins targeted as diagnostics tools interacting with the 10 best putative anti-HIV AMPs using the PatchDock docking algorithm

	p24				p17			
	Area (Å <sup>2</sup> )	ACE	Transformation coordinates	Area (Å <sup>2</sup> )	ACE	Transformation coordinates		
Molecule1	2330.80	89.53	1.13 0.19 -1.79 -32.59 -2.80 6.80	2329.10	311.56	-0.18 0.44 0.09 7.13 -15.92 -12.46		
Molecule2	1825.00	312.32	-0.64 -0.98 -2.71 -29.88 7.38 -1.06	1456.10	366.62	1.60 0.79 2.95 -2.91 -9.15 -6.66		
Molecule3	1534.10	119.26	2.86 -0.58 -0.56 -30.13 -4.02 2.50	1800.90	236.37	2.96 0.17 2.35 -2.02 -28.45 -3.40		
Molecule4	1790.90	-116.13	-2.21 0.64 2.35 -35.15 15.46 10.89	1486.90	484.40	2.40 0.99 -1.88 -4.65 -11.19 -1.01		
Molecule5	1681.50	230.81	2.57 -0.73 2.95 -22.02 24.23 1.75	1688.10	44.98	-2.02 0.66 0.78 -0.22 12.63 -11.26		
Molecule6	1859.20	-174.20	2.82 0.29 -0.84 -34.55 32.55 14.08	1792.50	-19.14	2.49 -0.58 -1.28 -1.91 -16.84 1.90		
Molecule7	1677.80	263.40	1.65 -0.53 0.17 -28.92 6.63 4.35	1765.70	158.86	2.97 1.26 0.58 1.89 1.72 -5.27		
Molecule8	1253.40	152.90	-0.07 0.91 -0.85 -2.56 4.20 15.29	1710.30	481.71	-0.04 -0.45 -2.23 -0.32 -8.12 -4.83		
Molecule9	1318.10	-552.96	0.67 -0.94 2.85 -20.23 7.83 -15.49	1230.10	156.21	-2.90 -0.47 2.60 -8.52 -18.56 5.61		
Molecule10	1659.20	165.71	0.64 -0.40 0.84 -39.41 15.33 1.88	1463.10	473.86	0.10 0.33 2.15 1.85 -3.50 -7.31		
+ve control (Kn2-7)	1309.40	-416.08	0.92 -0.49 -2.73 -35.56 40.09 4.47	1154.50	310.53	2.99 0.61 -0.42 2.94 -10.16 -6.89		
-ve control (MucroporinS1)	840.30	-8.38	-3.11 0.75 1.00 -28.80 -9.87 13.56	791.50	38.18	0.45 -0.16 2.64 7.32 -5.18 -15.97		

ACE: Atomic Contact Energy



UNIVERSITY *of the*  
WESTERN CAPE