

Machine Learning Identification of Protein Properties Useful for Specific Applications

Dissertation/Thesis by

Abdullah Mohammed Abdullah Khamis

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

King Abdullah University of Science and Technology

Thuwal, Kingdom of Saudi Arabia

2016

EXAMINATION COMMITTEE APPROVALS FORM

The dissertation of Abdullah Mohammed Abdullah Khamis is approved by the examination committee.

Committee Chairperson [Prof. Vladimir Bajic]

Committee Member [Prof. Takashi Gojobori]

Committee Member [Prof. Jesper Tegner]

Committee Member [Prof. Xin Gao]

© 2016

Abdullah Mohammed Abdullah Khamis

All Rights Reserved

ABSTRACT

Machine Learning Identification of Protein Properties Useful for Specific Applications

Abdullah Mohammed Abdullah Khamis

Proteins play critical roles in cellular processes of living organisms. It is therefore important to identify and characterize their key properties associated with their functions. Correlating protein's structural, sequence and physicochemical properties of its amino acids (aa) with protein functions could identify some of the critical factors governing the specific functionality. We point out that not all functions of even well studied proteins are known. This, complemented by the huge increase in the number of newly discovered and predicted proteins, makes challenging the experimental characterization of the whole spectrum of possible protein functions for all proteins of interest. Consequently, the use of computational methods has become more attractive.

Here we address two questions. The first one is how to use protein aa sequence and physicochemical properties to characterize a family of proteins. The second one focuses on how to use transcription factor (TF) protein's domains to enhance accuracy of predicting TF DNA binding sites (TFBSs).

To address the first question, we developed a novel method using computational representation of proteins based on characteristics of different protein regions (N-terminal, M-region and C-terminal) and combined these with the properties of protein aa

sequences. We show that this description provides important biological insight about characterization of the protein functional groups. Using feature selection techniques, we identified key properties of proteins that allow for very accurate characterization of different protein families. We demonstrated efficiency of our method in application to a number of antimicrobial peptide families.

To address the second question we developed another novel method that uses a combination of aa properties of DNA binding domains of TFs and their TFBS properties to develop machine learning models for predicting TFBSs. Feature selection is used to identify the most relevant characteristics of the aa for such modeling. In addition to reducing the number of required models to only 14 for several hundred TFs, the final prediction accuracy of our models appears dramatically better than with other methods.

Overall, we show how to efficiently utilize properties of proteins in deriving more accurate solutions for two important problems of computational biology and bioinformatics.

ACKNOWLEDGEMENTS

First and foremost, all my praise to "ALLAH", may He be glorified and exalted, because of the great favors and blessings that He has bestowed upon me during my life and for granting me the capability to complete my PhD study.

My great thanks go to my supervisor Prof. Vladimir Bajic for his support, time and guidance throughout my PhD research work. His invaluable advices and suggestions had high impact on my progress. It was a great privilege and honor to work with him. My great thanks are extended to my co-supervisor, Dr. Xin Gao, for his continuous support, follow-up and insightful ideas.

I am forever grateful for my beloved parents who have spared no effort to support me during all my life and who encouraged me to continue my education. My heartfelt gratitude is extended to my brother, sisters, my wife and my son for their support, understanding and encouragement during my PhD studies.

My appreciation also goes to my friends and colleagues in the Computational Bioscience Research Center (CBRC). Many thanks to Othman Soufan and Haitham Ashoor for the useful discussions that we had. Last but not least, I would like to thank King Abdullah University of Science and Technology (KAUST) for giving me the opportunity to pursue my PhD studies and for providing the stimulating research environment, outstanding computing resources and generous financial support and that helped me to do my research.

TABLE OF CONTENTS

EXAMINATION COMMITTEE APPROVALS FORM	2
ABSTRACT	4
ACKNOWLEDGEMENTS	6
TABLE OF CONTENTS.....	7
LIST OF ABBREVIATIONS	11
LIST OF SYMBOLS	13
LIST OF ILLUSTRATIONS.....	15
LIST OF TABLES.....	20
Chapter 1 Introduction	23
1.1 Background	23
1.2 Protein Functions.....	26
1.3 Inferring Protein Function.....	27
1.4 Contribution of the Dissertation.....	30
Chapter 2 Protein Functional Regions, Numerical Representation and Property Selection Methods.....	33
2.1 Amino Acid Properties	33
2.2 Protein Functional Regions	35
2.2.1 Dividing Protein Sequence into Regions	35
2.2.2 Protein DNA Binding Domains	36
2.3 Computational Representation of Proteins.....	37
2.3.1 Basic Representation Using Amino Acid Composition (AAC).....	37
2.3.2 Using Physicochemical Properties Weighted by Amino Acid Occurrences	38
2.3.3 Using Pseudo Amino Acid Composition Features.....	38
2.3.4 Numerical Representation of Protein Annotation.....	40
2.3.5 Using Physicochemical Properties to Represent Proteins	41
2.3.6 Using Physicochemical Properties and Neighboring Amino Acids Positions to Represent Proteins.....	42

2.4	Techniques of Feature Selection for Protein Properties Related to Problem in Question 42	
2.4.1	Unsupervised Feature Selection Techniques.....	43
2.4.2	Supervised Feature Selection Techniques	43
2.5	Machine Learning Classification Models.....	51
2.5.1	K-Nearest Neighbors	51
2.5.2	Artificial Neural Networks.....	52
2.5.3	Deep Learning	53
2.5.4	Ensemble Methods	53
Chapter 3 Distinct Profiling of Antimicrobial Peptides Using their Compositional and Physicochemical Properties		56
3.1	Summary	56
3.2	Introduction	57
3.3	Methods.....	60
3.3.1	Datasets	60
3.3.2	Peptide/Protein Sequence Models	62
3.3.3	Data Preparation.....	67
3.3.4	AMP Family-Specific Feature Selection.....	68
3.3.5	Clustering AMPs into Antimicrobial Families.....	69
3.3.6	Evaluation of Model Results	69
3.4	Results.....	72
3.4.1	Using Global Optimization of Unsupervised K-means Clustering for AMP Feature Selection 72	
3.4.2	Using Selected Feature Subsets to Cluster AMPs	75
3.4.3	Comparison between Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Differential Evolution (DE) optimization algorithms.....	83
3.4.4	Comparison of k-means clustering with Affinity Propagation (AP) clustering Algorithm	85
3.4.5	Comparison of clustering using different distance measures	86
3.4.6	Testing the Selected Properties on Non-AMPs and other AMP Databases.....	88
3.4.7	Selected Properties that Discriminate AMP Families	89
3.5	Discussion.....	91

3.6	Conclusion.....	100
Chapter 4 Novel ML Method of Prediction of Transcription Factor DNA Binding Sites. 101		
4.1	Summary	101
4.2	Introduction	102
4.3	Methods.....	106
4.3.1	Datasets	106
4.3.2	Modeling TF-TFBS sequence pairs	108
4.3.3	Data preparation.....	111
4.3.4	Positive (true) and negative (false) data	112
4.3.5	Random forests TFBS prediction model.....	114
4.3.6	Model evaluation metrics	114
4.3.7	Comparison of DRAF RF models with other model types.....	115
4.3.8	DRAF model validation on CHIP-seq data	116
4.3.9	Comparison between DRAF models, PWM models (HOCOMOCO, TRANSFAC) and DeepBind models on CHIP-seq datasets	119
4.4	Results.....	120
4.4.1	Selected properties of TFs	120
4.4.2	TF-TFBS predictions by DRAF models	122
4.4.3	Model evaluation on CHIP-seq data.....	127
4.4.4	Comparative performance of DRAF models	135
4.5	Discussion.....	137
4.6	Conclusion.....	139
Chapter 5 Conclusions and Future Work..... 140		
5.1	Introducing remarks.....	140
5.2	Comments on the Developed Methods.....	140
5.3	Contribution summary.....	142
5.4	Future Research	143
BIBLIOGRAPHY/REFERENCES		145
Appendix 1: Amino Acid Composition for 14 AMP Families.....		157
Appendix 2: Clustering Performance Using Different Values of Terminal Length Parameters.....		164

Appendix 3: Selected Properties Using GA to Discriminate 14 AMP Families 168
Appendix 4: DRAF Models Performance and Comparison of Results 183
Appendix 5: Sequence Logos for 321 CHIP-seq Datasets..... 193

LIST OF ABBREVIATIONS

AAC	Amino Acid Composition
AMP	Antimicrobial Peptide
AUC	Area under Curve
bp	Base-pairs
CADD	Computer-aided Drug Design
DE	Differential Evolution
EA	Evolution Algorithms
FN	False Negative
FP	False Positive
FSAP	Frog Skin Active Peptide
GA	Genetic Algorithm
GMR	Gaussian Mixture Regression
GO	Gene Ontology
HMM	Hidden Markov Models
MCC	Matthew's Correlation Coefficient
MDR	Multi-drug Resistant
MS	Mass Spectrometry
mRMR	Minimum Redundancy Maximum Relevance
NN	Neural Networks
PseAAC	Pseudo Amino Acid Composition

PSO	Particle Swarm Optimization
PWM	Position Weight Matrix
RF	Random Forests
ROC	Receiver Operating Characteristic Curve
SVM	Support Vector Machines
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TN	True Negative
TP	True Positive

LIST OF SYMBOLS

Symbol	Definition
d_c	Length of C-terminal
d_n	Length of one of four sub-regions within the N-terminal
e_i	Entropy of cluster i
$Freq_k$	Total occurrences of amino acid k in the protein sequence
H	Distance between two successive amino acids
$I(X, Y)$	Mutual information between X and Y
L	Length of a protein
m	Mutation rate
m_i	Number of objects in cluster i
m_{ij}	Number of objects of class j in cluster i
P	Initial population consists of n hypotheses
P_s	New population of n hypotheses
$Pr(h_i)$	Probability of hypothesis h_i
r	Fraction of the population to be replaced by crossover
$\Phi(R_i)$	Property value of amino acid R_i
S	Subset of features (properties)
$ S $	Total number of features in the subset S
T_k	k-th tier correlation between the k-th contiguous residues

σ	Standard deviation
u_i	Means of feature values

LIST OF ILLUSTRATIONS

Figure 1.1 Central Dogma of Molecular Biology. Genes on DNA are transcribed into RNA, which is processed into mRNA and later used in the process of protein synthesis.....	24
Figure 1.2 Hierarchy of protein sequence, structure, dynamics and function. Correlation between different levels of protein properties enable protein function prediction (Hensen, et al., 2012).....	28
Figure 1.3 Proposed framework in this study. The process starts by representing numerically functional regions of proteins using amino acid physicochemical properties. Then, ML feature selection is performed to determine properties to be used to build a model for characterizing proteins in relation to their function.	31
Figure 2.1. Different regions of a protein sequence. Definition of the N-terminal, M region and C-terminal regions (Matsuda, et al., 2005).....	36
Figure 2.2. Correlation between residues. The 1 st -tier, 2 nd -tier and 3 rd -tier correlation along the protein sequence (Chou, 2001).	40
Figure 2.3. Genetic Algorithm Flowchart. A flowchart shows different steps of the genetic algorithm.	49
Figure 2.4 An Example of a Neural Network Structure. A typical NN is composed of different layers, each of which consists of several nodes. Nodes within different layers are connected by links that have assigned weights.	52

Figure 2.5 Illustration of a Typical Structure of Random Forests. The Random Forest model is an ensemble of N trees, each of which learn from data using part of the features. All predictions from these sub-trees are combined to produce the Random Forest model prediction.	55
Figure 3.1. AMP length distribution. A histogram of the distribution of lengths of 478 AMPs from 14 target AMP families.	74
Figure 3.2. K-means clustering performance (sensitivity vs. precision). Plot of sensitivity vs. precision obtained from k-means clustering of 14 target AMP families using optimized set of selected features.	76
Figure 3.3. Comparison between AMP representation methods (precision). Bar plots of the precision of four AMP representation methods.	82
Figure 3.4. Comparison between AMP representation methods (sensitivity). Bar plots of the sensitivity of four AMP representation methods.	82
Figure 4.1. Distribution of the selected properties used by DRAF models. Distribution of 145 selected AAindex properties; 115 (79%) properties belong to six biological classes, while other 30 (21%) properties are 'unclassified' according to (Kawashima, et al., 2008).	122
Figure 4.2. The input data, training procedure and usage of DRAF models for prediction of TF-TFBS pairs. Sequences of TFs and their TFBS are represented in TF-TFBS pairs using physicochemical properties of TFs and binary representation of TFBSs. Then, DRAF models were constructed for each group of TFs depending on the TFBS length. Finally, DRAF models were tested using the test data and another set of independent ChIP-seq	

validation datasets. The DRAF models predict which TF-TFBS pair represent a valid target TFBS for a particular TF. 123

Figure 4.3. The prediction performance of DRAF models on the test data. DRAF models were applied on the test data using different settings for selecting thresholds on the models' prediction score that provided (A) the highest accuracy, (B) the highest sensitivity and (C) the highest specificity, on the training set. 124

Figure 4.4. The ROC curve for the evaluation of the DRAF models on the test set. The ROC curve (true positive rate vs. false positive rate) was performed for the prediction outputs obtained from all the 14 DRAF models on the test set. The AUC for the DRAF models is 0.9991. 125

Figure 4.5. Comparison between RF, NN, GMR and SVM models. All the four types of models were trained using the same training data and applied on the same test data. The models were compared using performance measures calculated on the test data. 126

Figure 4.6. Sequence logos for the predicted TFBS sequences on the human CHIP-seq datasets using DRAF models. The figure shows different sequence logos obtained from the DRAF predicted TFBS sequences from CHIP-seq datasets at different sensitivity levels. The complete set of sequence logos for the 321 ENCODE CHIP-seq datasets is provided in Appendix Table A5.1. 128

Figure 4.7. Comparison of DRAF, HOCOMOCO, TRANSFAC and DeepBind models using thresholds on the DRAF model output scores that provided the highest F-measure.

Comparison of performance of 14 DRAF models and 321 HOCOMOCO PWMs, 319

TRANSFAC PWMs and 302 DeepBind models on 321 ChIP-seq datasets obtained from ENCODE using thresholds on the DRAF model prediction scores that yielded highest F-measure scores on the training data. In A, the Y-axis represents the logarithm of distance (in nt) between false positive prediction occurrences on the background sequences (chromosome 21) averaged over all the tested ChIP-seq datasets. The labels on the top of the blue bars indicate the average sensitivity on the ChIP-seq datasets using the corresponding model. In B, the box plots show distance (in nt) between false positive prediction occurrences on the background sequences (chromosome 21) across all the tested ChIP-seq datasets. 129

Figure 4.8. Comparison of DRAF, HOCOMOCO, TRANSFAC and DeepBind models at different sensitivity levels (averaged over all ChIP-seq datasets). Comparison of 14 DRAF models and 321 HOCOMOCO PWMs, 319 TRANSFAC PWMs and 302 DeepBind models on 321 ChIP-seq datasets obtained from ENCODE. The X-axis represents different sensitivity levels (A: 10%, 20%...50% and B: 60%, 70%...90%) and the Y-axis represents the distance (in nt) between false positive prediction occurrences on the background sequences (chromosome 21) averaged over all the tested ChIP-seq datasets..... 132

Figure 4.9. Comparison of DRAF, HOCOMOCO, TRANSFAC and DeepBind models at different sensitivity levels. Comparison of 14 DRAF models and 321 HOCOMOCO PWMs, 319 TRANSFAC PWMs and 302 DeepBind models on 321 ChIP-seq datasets obtained from ENCODE at different sensitivity levels (10%, 20%...90%). The Y-axis represents the logarithm of the distance (in nt) between false positive prediction occurrences on the background sequences (chromosome 21) across all the tested ChIP-seq datasets. 133

Figure 4.10. Summary of Comparison between DRAF, HOCOMOCO, TRANSFAC and DeepBind models. A: boxplots show the average distance (in nt) between false positive prediction occurrences on the background sequences (chromosome 21) at different sensitivity levels obtained from testing corresponding models on the ChIP-seq datasets and averaged over all the datasets at each sensitivity level. B: boxplots show the folds-reduction in false positive predictions obtained in by DRAF models as compared to the corresponding HOCOMOCO, TRANSFAC and DeepBind models..... 134

LIST OF TABLES

Table 2.1. List of amino acids and values of two of their properties. A list of twenty amino acids, their 1-letter abbreviation and two of their physicochemical properties as in AAIndex database.	34
Table 2.2. Genetic Algorithm. Pseudo code for the GA algorithm as described in (Mitchell, 1997).....	47
Table 2.3. Crossover in genetic algorithm. Three common types of crossover operation during the genetic algorithm processing.....	50
Table 3.1. Number of peptides in each of the 14 AMP families/sub-families. Distribution of the obtained peptides from DAMPD database among 128 AMP families and sub-families.....	61
Table 3.2. The number of selected properties for AMP families. The number of features that characterize different regions of AMP sequence is selected using GA optimization of unsupervised k-means clustering. Annotations of columns are as follows: N-terminal length (dn), C-terminal length (dc), number of peptides (NP), original number of features (NF), number of selected features (NSF), number of clusters (NC), number of compositional features (NCF) and number of physicochemical features (NPF).	74
Table 3.3. K-means clustering performance using selected properties by GA algorithm. K-means clustering performance of 14 target AMP families using GA selected features...	77

Table 3.4. K-means clustering performance using amino acid composition properties. The k-means clustering performance of 14 target AMP families using 184 features that represent information about amino acid composition.	78
Table 3.5. Performance of k-means clustering using all properties. The k-means clustering performance of 14 target AMP families using the entire set of features to represent each family.	79
Table 3.6. K-means clustering performance using all physicochemical properties. The k-means clustering performance of 14 target AMP families using all the 294 physicochemical properties in each of the 6 regions (n1,n2,n3,n4,M and C).....	81
Table 3.7. K-means clustering performance using PSO algorithm. The performance of the k-means clustering of 14 target AMP families using features selected by the Particle Swarm Optimization (PSO) algorithm.....	83
Table 3.8. K-means clustering performance using DE algorithm. The performance of the k-means clustering of 14 target AMP families using features selected by the Differential Evolution (DE) optimization algorithm.	84
Table 3.9. Affinity Propagation clustering performance. The performance of the Affinity Propagation (AP) clustering of 14 target AMP families using features selected by the Genetic Algorithm.	85
Table 3.10. K-means clustering performance using city block distance. The performance of the k-means clustering (using city block distance measure) of 14 target AMP families using features selected by the Genetic Algorithm.	86

Table 3.11. K-means clustering performance using cosine distance. The performance of the k-means clustering (using cosine distance measure) of 14 target AMP families using features selected by the Genetic Algorithm.	87
Table 3.12. K-means clustering performance using correlation distance. The performance of the k-means clustering (using correlation distance measure) of 14 target AMP families using features selected by the Genetic Algorithm.	88
Table 4.1. TFBS model distribution according to TFBS length. Total of 250 TFBS models distributed in 14 models represent 14 distinct TFBS lengths.	107
Table 4.2. Amino acid classification. Distribution of 20 Amino acids according to their binding mode preference to DNA bases, adopted from Table 4 in (Luscombe and Thornton, 2002).	110
Table 4.3. Comparison between DRAF and other model types. The average prediction results on the test data using 4 model types, DRAF, Neural Networks (NN), Gaussian Mixture Regression (GMR) and Support Vector Machines (SVM).	127
Table 4.4. Average distance (in nt) between false prediction occurrences on the background sequences.	135
Table 4.5. Comparison of prediction results from different studies. This table shows the prediction accuracy of DRAF models and other models on different TF-TFBS test datasets.	136

Chapter 1

Introduction

1.1 Background

Cells are the basic building blocks of living organisms. A cell is the smallest unit of life that can replicate itself during the cell division process. Organisms differ in the type and number of cells they have. Unicellular organisms e.g. bacteria consist of a single cell. On the other hand, multicellular organisms consist of numerous cells. A human body, for example, is composed of trillions of cells.

A typical cell in a eukaryotic organism contains Deoxyribonucleic Acid (DNA) packed into multiple chromosomes inside its nucleus. The DNA harbors the hereditary information of the living organism. According to the Central Dogma of Molecular Biology, genes on DNA are transcribed into RNA, which is processed into mRNA and later used in the process of protein synthesis (Figure 1.1). Today, however, we know that protein-coding genes cover only a few percentages of the human genome and that most of the human genome is transcribed generating transcripts of various functionality (Carninci, et al., 2005; Consortium, 2012; Consortium, et al., 2014; Gerstein, et al., 2012; Ravasi, et al., 2010).

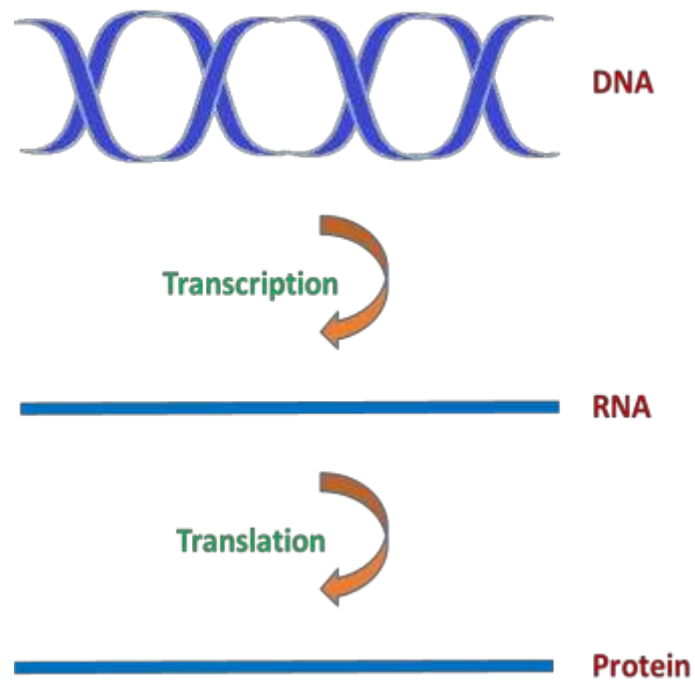


Figure 1.1 Central Dogma of Molecular Biology. Genes on DNA are transcribed into RNA, which is processed into mRNA and later used in the process of protein synthesis.

Proteins are the workhorse molecules of life as they play critical roles in cellular processes of living organisms. They exist in all types of cells, perform main activities that sustain life and they form a major structural component of muscles, skin, hair, organs and other tissues of the body. Currently, there are tens of millions of protein sequences available in numerous protein databases, such as, Uniprot (UniProt, 2015) and Protein Data Bank (wwPDB) (Berman, et al., 2003).

A typical protein is a large molecule composed of amino acids (also referred to as residues), which serve as building blocks for the protein (Raven and Johnson, 2002).

Amino acids are attached to each other through peptide bonds in a long chain to form the protein sequence. The primary structure of the protein is represented in terms of this linear chain of amino acid residues (Voet, et al., 1999). The arrangement of these amino acids in different orders within a long sequence can produce huge number of proteins. If the sequence of an amino acid chain consists of fewer than 50 amino acids, then it is called peptide sequence (McKee and McKee, 2012). There are 20 different amino acids that have a common internal structure of carbon, oxygen, hydrogen and nitrogen atoms. However, amino acids differ from each other based on “R” group attached to the fourth covalent bond of the central carbon.

Proteins are described in terms of their primary, secondary, tertiary and quaternary structures (Lieberman, et al., 2013). The primary structure of a protein is represented by its sequence of amino acids. The protein secondary structure is the three-dimensional (3D) folding structure of local segments inside the protein sequence. However, the tertiary structure of a protein refers to its overall 3D folding structure. Because many proteins contain multiple polypeptide chains, the quaternary structure of such proteins refers to the way these subunits interact with each other. Consequently, proteins differ between each other in terms of their amino acid sequence composition (i.e. primary structure) and their local and overall structures. This difference allows for and enables different functions performed by the proteins.

1.2 Protein Functions

Proteins perform wide variety of major functions in the cells. Because many proteins execute the same or similar functions, they can be classified into specific categories according to the function they perform. For example, the set of proteins that build and repair the structure of the cells are known as “Structural Components” of the cell. “Antibodies and antimicrobial” proteins are components of the immune system of the body and they are responsible specific defense against certain groups of pathogens. Proteins that are required for enabling chemical reactions inside the cells are called “enzymes”. “Transport” proteins carry molecules between different parts of the organism. Biological processes that of activation of transcripts and gene regulation involve proteins called “Transcription Factors (TFs)”.

A classification of protein functions using well-defined ontology of functions was provided by the Gene Ontology Consortium (Ashburner, et al., 2000). According to this classification, the protein functions are grouped into biological processes, molecular functions, and cellular components (Ashburner, et al., 2000), where each of these broad categories contains different granularity of descriptions of protein functions from a particular perspective (Lee, et al., 2007).

1.3 Inferring Protein Function

To understand what roles proteins play in cells and under which conditions, we need first to know what functions they could have. Many proteins do not have known functions and even for the well-studied proteins all their functions are not known. In addition, many novel proteins are discovered and need annotation of their function. As a result of the rapid increase in the number of discovered proteins, the experimental annotation of this huge number proteins has become more challenging and time consuming task (Bromberg, et al., 2009). Consequently, substantial amount of computational research has been performed to assist in annotating these proteins and studying their properties (Friedberg, 2006).

Explaining the protein function comes through a hierarchy that starts from the protein sequence through the protein structure up to the protein dynamics and exerted function, as shown in Figure 1.2 (Hensen, et al., 2012). Because the identification of protein function through its structure is more reliable (Pascual-Garcia, et al., 2010), there are many studies done on so called structure-function modeling. Predicting the protein function from its structure can be achieved by analyzing the global 3D structure of protein folds (Hegyí and Gerstein, 1999), or by using the local substructure of the protein (Hvidsten, et al., 2009). One direction to model protein structure-function relationship is by integrating multiple mass spectrometry (MS) data (Landreh, et al., 2011). Studying the ligand-binding sites on the protein surface to model their binding activity is another direction (Yuan, et al., 2013). Recent works utilize protein structure

information in computer-aided drug design (CADD) to improve the hit rate of drugs (Zheng, et al., 2013) and more accurate binding with receptors (Garcia, et al., 2012).

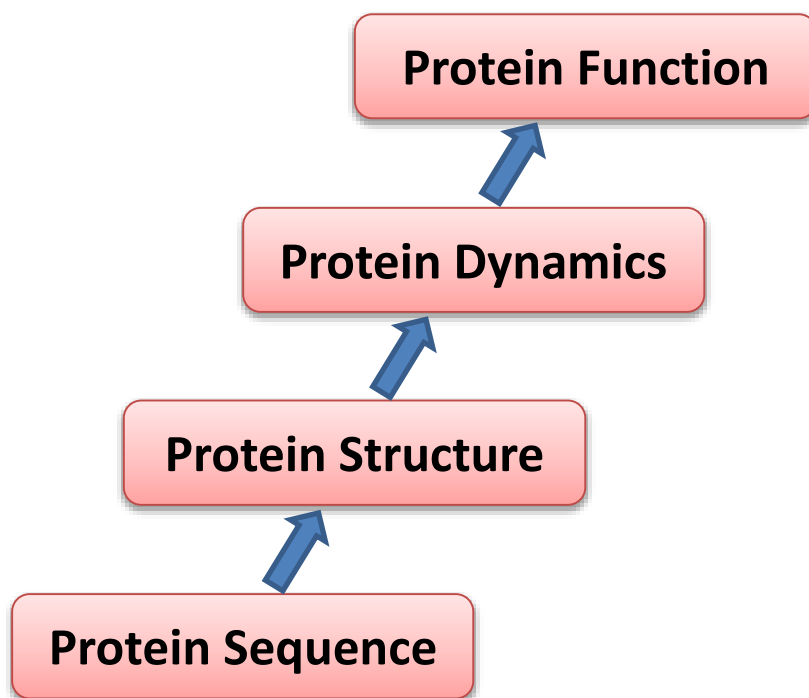


Figure 1.2 Hierarchy of protein sequence, structure, dynamics and function. Correlation between different levels of protein properties enable protein function prediction (Hensen, et al., 2012).

Despite the importance of protein structure in understanding protein function, very small percentage of proteins have their 3D structures experimentally confirmed (Berman, et al., 2000; Lee, et al., 2007). As of 2011, from over 80 million protein sequences, only 70,000 protein structures exist in protein data bank (Leszczynski, 2012). As a result, studying protein function through the analysis of protein sequence is remaining a common approach. This is supported by the following reasons. First, it is

widely accepted that the protein's functions depends on its 3D structure which is determined by its amino acids sequences (Berg, et al., 2002), so the primary sequence of a protein is the crucial contributor to its function. Second, it has been shown that the predicted functions for proteins with more than 50% sequence identity to the reference protein determined based on homology are valid in majority of cases (Sangar, et al., 2007). Third, protein function cannot be reliably predicted from its structure alone (i.e. in the absence of information from its sequence). It has been found that proteins with very different structure have similar function, and some proteins with very similar structure have different functions (Hensen, et al., 2012). Consequently, considerable efforts have been made to study different properties of amino acids in the protein sequence. One purpose of such research is to explain different functions of proteins based on properties extracted from their primary sequence (Clark and Radivojac, 2011).

Taking into account that each of the twenty amino acids is characterized by a wide variety of properties, some databases have been created containing quantitative parameters of individual amino acids. For example, AAIndex (Kawashima, et al., 2008) database contains numerical indices of 544 different physicochemical properties of amino acids.

There are 4,343 publications indexed in Thompson Reuters Web of Science Core Collection on "protein function from primary sequence" and 767 publications on "model protein function from primary sequence" as of February 26, 2016. This is a large volume of work in this field and it is difficult to address all relevant aspects of it. We will thus

only refer to few references that are most directly related to our work. To model protein functions based on characteristics of their amino acid sequences, some of the previous work used only amino acid composition features extracted from different segments of the primary sequence e.g. (Matsuda, et al., 2005), while some other work used physicochemical properties of amino acids e.g. (Cai, et al., 2009; Cai, et al., 2010). However, a study of all available properties of amino acids in different regions of the protein sequence is yet to be done. The utilization of this analysis of protein properties in different regions of protein amino acid sequences to study their characteristics and predict their function will definitely result in better models for predicting protein functions.

1.4 Contribution of the Dissertation

In this research we addressed two questions. The first one is how to use protein amino acid sequence and physicochemical properties to characterize a family of proteins. The second one focuses on how to use transcription factor (TF) protein's domains to enhance accuracy of predicting TF DNA binding sites (TFBSs). In dealing with both of these questions we used a simple framework (Figure 1.3) for identifying relevant characteristics of proteins based on their amino acid sequence that could serve as good protein representation in relation to their function.

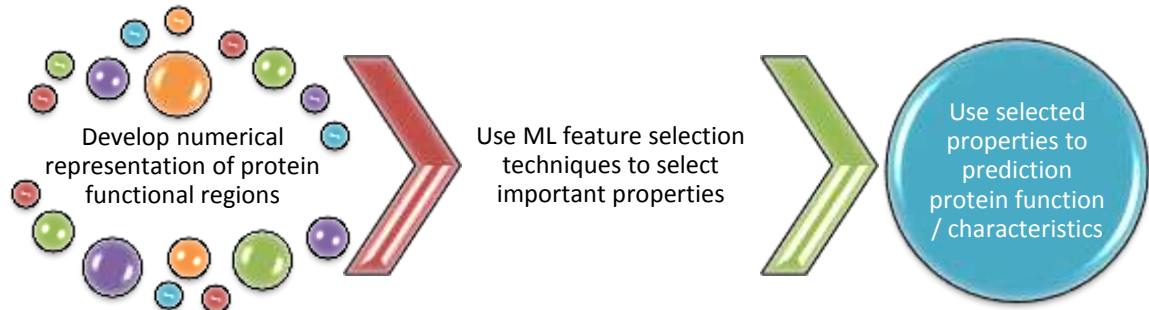


Figure 1.3 Proposed framework in this study. The process starts by representing numerically functional regions of proteins using amino acid physicochemical properties. Then, ML feature selection is performed to determine properties to be used to build a model for characterizing proteins in relation to their function.

To address the first question, we developed a novel method using computational representation of proteins based on characteristics of different protein regions (N-terminal, M-region and C-terminal) and combined these with the properties of protein amino acid sequences. We show that this description provides important biological insight about characterization of the protein functional groups. Using feature selection techniques, we identified key properties of proteins that allow for very accurate characterization of different protein families. We demonstrated efficiency of our

method in application to a number of antimicrobial peptide (AMP) families and achieved extraordinary accuracy in discrimination of these families. This is based on a very small set of distinguishing properties for different AMP families that our method identified as relevant for this purpose. Details are given in Chapter 3.

To address the second question we developed another novel method that uses a combination of amino acid properties of DNA binding domains of TFs and their TFBS properties to develop machine learning models for predicting TFBSs. Feature selection is used to identify the most relevant characteristics of the amino acid for such modeling. In addition to reducing the number of required models to only 14 for several hundred TFs, the final prediction accuracy of our models appears dramatically better than with other methods. Details are provided in Chapter 4.

Overall, we show how to efficiently utilize properties of proteins in deriving more accurate solutions for two important problems of computational biology and bioinformatics.

Chapter 2

Protein Functional Regions, Numerical Representation and Property Selection Methods

2.1 Amino Acid Properties

An amino acids chain forms the primary sequence of a protein. There are twenty amino acids used in protein sequences. Each amino acid can be suitably annotated by a single letter (Table 2.1). All amino acids have common structure of carbon, hydrogen, oxygen and nitrogen. Each amino acid differs from other amino acids by the side chain, known as an R group.

Amino acids are characterized by different set of physicochemical properties and can have values of these properties assigned to them. Because of the difference in the R group between amino acids, each amino acid differs from other amino acids in some properties. There are many attempts to measure and quantify the properties of amino acids (Gromiha, 2010). For example, (Kidera, et al., 1985) collected 188 physical properties of amino acids, while (Palliser and Parry, 2001) collected 127 hydrophobicity scales. Yet, AAIndex (Kawashima, et al., 2008) remains the most comprehensive database of physicochemical properties of amino acids. The latest update of the database contained 544 properties of amino acids (indices). Table 2.1 shows the values of two properties (hydrophobicity index and normalized frequency of alpha-helix) for each of the twenty amino acids.

Table 2.1. List of amino acids and values of two of their properties. A list of twenty amino acids, their 1-letter abbreviation and two of their physicochemical properties as in AAIndex database.

Amino Acid	Letter	Hydrophobicity index	Normalized frequency of alpha-helix
Alanine	A	0.61	1.42
Arginine	R	0.60	0.98
Asparagine	N	0.06	0.67
Aspartic acid	D	0.46	1.01
Cysteine	C	1.07	0.70
Glutamic acid	E	0.47	1.51
Glutamine	Q	0	1.11
Glycine	G	0.07	0.57
Histidine	H	0.61	1.00
Isoleucine	I	2.22	1.08
Leucine	L	1.53	1.21
Lysine	K	1.15	1.16
Methionine	M	1.18	1.45
Phenylalanine	F	2.02	1.13
Proline	P	1.95	0.57
Serine	S	0.05	0.77
Threonine	T	0.05	0.83
Tryptophan	W	2.65	1.08
Tyrosine	Y	1.88	0.69
Valine	V	1.32	1.06

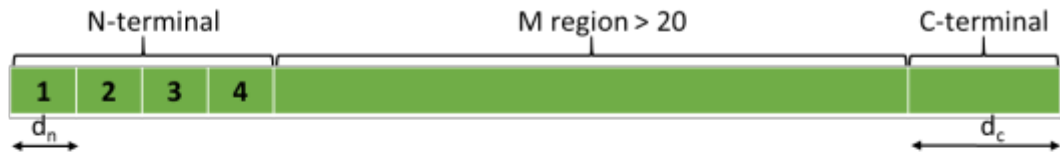
2.2 Protein Functional Regions

There are different regions in the protein sequence that contain various patterns and consequently may associate to different functions. For example, the N-terminal and C-terminal of the protein sequence are useful to identify protein cellular location (Emanuelsson, et al., 2000; Matsuda, et al., 2005; Reczko and Hatzigerorgiou, 2004). The binding characteristics of the TF proteins to DNA are highly determined by the DNA binding domain regions in the TF protein sequence.

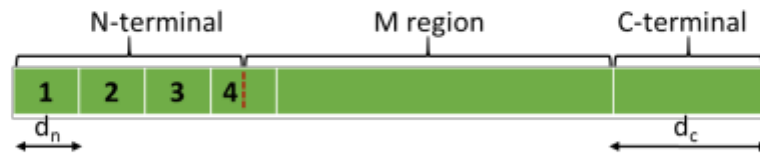
2.2.1 Dividing Protein Sequence into Regions

Proteins have different functions and different regions inside a protein sequence may have different characteristics. The motivation behind dividing a protein sequence into regions is to be able to study the local effects of region characteristics of protein sequence. For example, in a protein sequence one can distinguish three segments: N-terminal, M-region and C-terminal. While there is no confident method to identify precisely these regions, there are different computational methods that approximately identify these regions within the protein sequence. As an example, (Matsuda, et al., 2005) selected these regions based on the protein length L . In their method, N terminal was divided into sub sections (n_1 , n_2 , n_3 and n_4). Two parameters, d_n (which refers to the length of a sub region inside the N terminal) and d_c (refers to the length of C terminal) were used to control the selection of these regions. The definition of the protein regions based on the protein length is illustrated in Figure 2.1.

A) $L > 4*d_n + 20 + d_c$



B) $4*d_n + d_c < L < 4*d_n + 20 + d_c$



C) $L < 4*d_n + d_c$

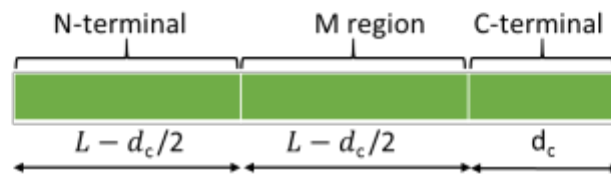


Figure 2.1. Different regions of a protein sequence. Definition of the N-terminal, M region and C-terminal regions (Matsuda, et al., 2005)

2.2.2 Protein DNA Binding Domains

A protein domain is a functional and conserved part of the protein associated to a particular function and contributes to the overall protein function. A protein may have more than one domain. The protein domain associated to the protein binding to the DNA is called 'DNA Binding Domain'.

Determining the DNA binding domains is useful for prediction of TF affinity to its binding sites on the DNA. There are several databases that contain lists of predicted domain sequences. For example, PFAM (Finn, et al., 2014) database contains a large collection of proteins families and their predicted domains. Other databases include, SMART (Letunic, et al., 2015), COG (Tatusov, et al., 2000) and CDD (Marchler-Bauer, et al., 2015).

2.3 Computational Representation of Proteins

There are different methods to represent proteins numerically using characteristics extracted from their primary sequence or general annotation of the protein functional properties. In this section, we summarize some of the common numerical representation methods for proteins.

2.3.1 Basic Representation Using Amino Acid Composition (AAC)

Here, a protein P is represented by the following:

$$Feature_i(P) = \frac{Freq_i}{Length(P)} \quad (2.1)$$

where $Freq_i$ is the total count of an amino acid i in the sequence of protein P . The length of the primary sequence of protein P is denoted as $Length(P)$.

2.3.2 Using Physicochemical Properties Weighted by Amino Acid Occurrences

In this representation, a protein P is represented by a set of features such that each feature i corresponds to a physicochemical property i . That is, a feature i of P is the average value of the physicochemical characteristics i weighted by the relative occurrences of individual amino acids in the sequence, according to the following:

$$Feature_i(P) = \sum_{\text{Amino acid } k=1}^{20} \frac{Freq_k}{Length(P)} * Property Value_i(k) \quad (2.2)$$

where $Freq_k$ is the total count of amino acid k in the primary sequence of the protein P ; $Property Value_i(k)$ is the value of physicochemical property i for amino acid k ; $Length(P)$ is the length of P . This score is calculated for all features $i = 1, 2, \dots, m$, generating a feature vector of length m for each protein P .

2.3.3 Using Pseudo Amino Acid Composition Features

Pseudo amino acid composition (PseAAC) was first introduced by (Chou, 2001) to predict protein subcellular localization. We follow here his annotation. In this representation, a protein P of length L is represented as:

$$P = [p_1, p_2, \dots, p_{20}, p_{20+1}, \dots, p_{20+\lambda}] \quad , \quad (\lambda < L) \quad (2.3)$$

where the $20 + \lambda$ elements are given by

$$p_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{k=1}^{\lambda} T_k} & , \quad (1 \leq u \leq 20) \\ \frac{w T_{u-20}}{\sum_{i=1}^{20} f_i + w \sum_{k=1}^{\lambda} T_k} & , \quad (20 + 1 \leq u \leq 20 + \lambda) \end{cases} \quad (2.4)$$

and where f_u is the frequency of amino acid at position u . The weight factor is represented by w . Also, T_k represents the k -th tier correlation between the k -th most contiguous residues. This correlation is calculated as follows:

$$T_k = \frac{1}{L-k} \sum_{i=1}^{L-k} J_{i,i+k} \quad (k < L) \quad (2.5)$$

The factor $J_{i,i+k}$ is shown in Figure 2.2 and is calculated as follows:

$$J_{i,i+k} = \frac{1}{\Gamma} \sum_{q=1}^{\Gamma} [\Phi_q(R_{i+k}) - \Phi_q(R_i)]^2 \quad (2.6)$$

where $\Phi_q(R_i)$ is the q -th property value (e.g. hydrophobicity value) of amino acid R_i at position i . The total number of properties considered is represented by Γ .

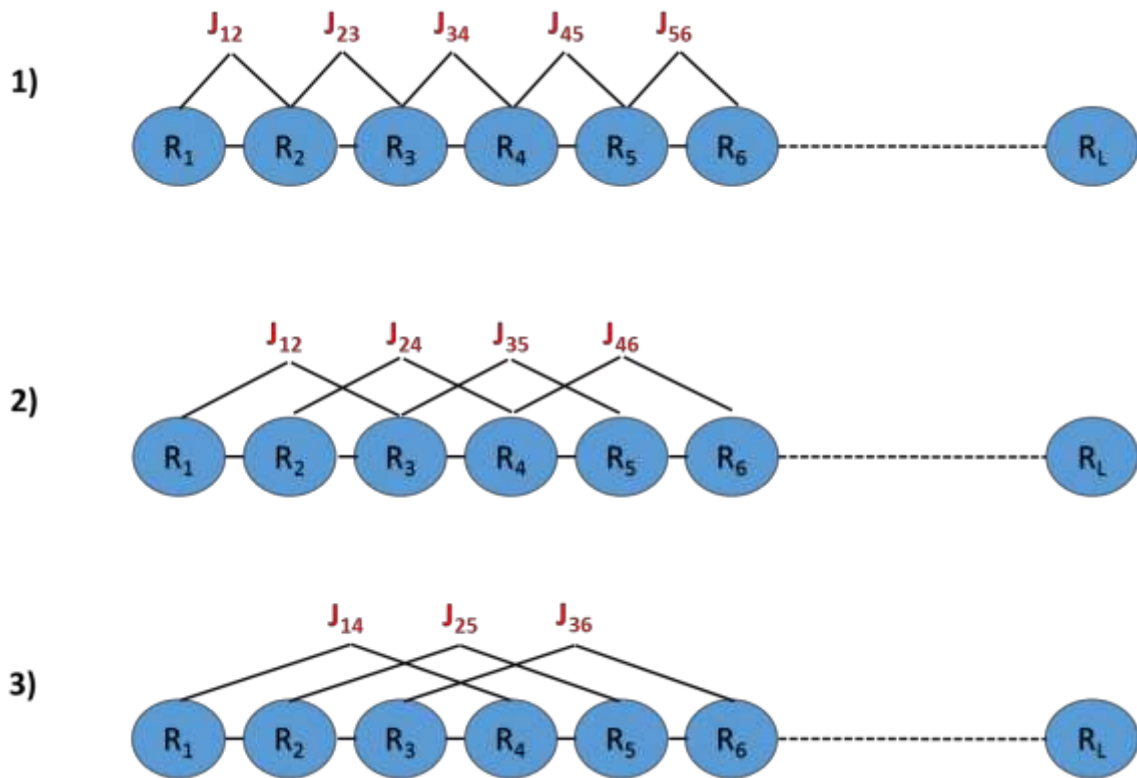


Figure 2.2. Correlation between residues. The 1st-tier, 2nd-tier and 3rd-tier correlation along the protein sequence (Chou, 2001).

2.3.4 Numerical Representation of Protein Annotation

This representation has been proposed in (Qian, et al., 2007) to model TFs in order to predict their DNA binding sites. Each TF-target gene-binding site (TF-TFT-TFBS) triplet is represented by a feature vector with the structure:

TF Properties (8151 Values)	TFT Properties (8151 Values)	TFBS Properties (125 values)
-----------------------------	------------------------------	------------------------------

where TF properties is a vector with 8151 values of all possible annotations in Interpro (Apweiler, et al., 2001). Each value in this vector is set to either 0 or 1 to mark the presence or absence, respectively, of protein existence in the corresponding Interpro entry. A similar vector of 8151 values is used to represent the TF target gene. The TFBS sequence is represented in 125 binary values as follows. The TFBS sequence is limited to the length 25 bp, such that if the sequence is shorter than 25 bp, it is extended to the length 25 by adding suffix of Ns. However, if the sequence is larger than 25 bp, the first 25 bp are only encoded. Then, each nucleotide is represented by 5 binary values as A=00001, C=00010, G=00100, T=01000, N=10000.

2.3.5 Using Physicochemical Properties to Represent Proteins

The amino acid properties of TFs were included in the TF-TFBS model in a study by (Cai, et al., 2009). In addition to the 20 amino acid composition features, only six physicochemical properties of amino acids were used to generate a 132-dimensional feature vector for a TF. These properties are hydrophobicity, polarizability, predicted solvent accessibility, predicted secondary structure, normalized van der Waals volume, and polarity. The TFBS of length 25 bp is represented in a 100 element vector where each nucleotide is represented by 4 binary digits. This representation is illustrated as follows:

TF Properties (132 Values)	TFBS Properties (100 Values)
----------------------------	------------------------------

2.3.6 Using Physicochemical Properties and Neighboring Amino Acids Positions to Represent Proteins

This representation was suggested in (Cai, et al., 2010) to predict protein DNA-binding residues on the protein sequences. 506 AAIndex properties were used to represent each amino acid and its eight surrounding amino acids (four from upstream and four from downstream). As a result, 4554 features are used to represent a particular amino acid. In addition, nine conservation scores are used to represent that amino acid as follows. PSI-BLAST (Altschul, et al., 1997) is used to find homologous proteins to a particular protein. Then, ClustalW (Larkin, et al., 2007) is used to make multiple alignment of these sequences. Finally, the conservation score is calculated using CONSCORE (Valdar, 2002) approach.

TF Properties (506 x 9) Values	TF Conserved Properties (9 Values)
--------------------------------	------------------------------------

2.4 Techniques of Feature Selection for Protein Properties Related to Problem in Question

Here we discuss some methods that can be used to select subset of protein properties that are more relevant to problem in question. It is assumed that proteins are represented by feature vectors suitable for the problem analyzed.

2.4.1 Unsupervised Feature Selection Techniques

Unsupervised feature selection techniques aim to select features using unsupervised methods such as clustering or matrix factorization techniques. The unsupervised methods are less likely to suffer from overfitting and are useful when there are more unlabeled data than labeled data (Guyon and Elisseeff, 2003). In the clustering approach, the features are grouped into clusters using some of the clustering algorithms (e.g. K-means or hierarchical clustering). Then, similar features within a particular cluster are replaced by a representative feature from the cluster (e.g. the centroid point). The matrix factorization methods include different methods that reduce the dimensionality of the data by mapping the data from the original space to a different space. Principal Component Analysis (PCA) (Jolliffe, 2002), Singular Value Decomposition (SVD) (Golub and Van Loan, 1983) and Non-negative Matrix Factorization (NMF) (Sra and Inderjit, 2006) are examples of the linear matrix factorization methods. Isomap (Tenenbaum, et al., 2000) is an example of nonlinear methods for dimensionality reduction.

2.4.2 Supervised Feature Selection Techniques

The supervised feature selection techniques use class labels of the data to guide the process of feature selection from the training data. These methods are classified into two broad categories, individual feature ranking and feature subset selection methods.

2.4.2.1 'Univariate' Individual Feature Selection Methods

Individual feature ranking methods evaluate each feature independently, one feature at a time, to select the most relevant features to the class label. These methods are fast as they evaluate features one at a time and each feature is evaluated independently of other features. Individual feature ranking methods differ in the metrics used to evaluate and rank individual features, such as, Area Under Curve (AUC), mutual information, Pearson correlation coefficient, T-test and Fisher scores.

2.4.2.2 Subset 'Multivariate' Feature Selection Methods

While individual feature ranking methods are fast and easy to understand, the ranked features may not necessarily represent the best subset of features. A feature that is ranked very low in the list of relevant features for a class label, may be very useful when combined with other features. For this reason, subset feature selection methods aim to evaluate different combination of features to select the subset of features that can improved the prediction accuracy or other performance metrics of the model.

The subset feature selection methods are classified into two main categories according to the explicit usage of the learning model in the feature selection step. These two categories are the filter and the wrapper methods. If the learning model is used as a part of the evaluation function for the subset of features then it is called the wrapper method, otherwise it is the filter method.

One category of filter methods is the correlation based feature selection (CFS) where the selected subset of features represent features which are highly correlated with the

prediction class and minimally correlated with each other. An example of such methods is the Minimum Redundancy Maximum Relevance (mRMR) (Peng, et al., 2005) method. Wrapper methods can work as the 'forward' or 'backward' feature selection mode. In the forward selection, the feature subset selection process starts with an empty set of features and the features are evaluated and the best is added to the subset. The process iteratively progresses until the satisfactory set is determined or some other constraints are reached. On the hand, in the backward selection, the initial subset consists of all features. Then, the worst feature is removed and this process goes iteratively until the satisfactory feature set is obtained or some other constraints are reached.

The wrapper methods are powerful and demonstrated their usefulness in different applications (Saeys, et al., 2007). The population-based heuristics methods, e.g. Particle Swarm Optimization (PSO) (Kennedy, 2010; Kennedy and Eberhart, 1995) and Genetic Algorithms (GA) (Holland, 1992) are examples of these methods. While the predictive power of the wrapper methods comes from the usage of the learning model during the feature selection process, they are time consuming and are more prone to over-fitting.

In the next two subsections we explain two of the feature subset selection methods. Firstly, we will discuss the mRMR method. Secondly, we will discuss a global optimization method, Genetic Algorithm, which can be used to select the set of properties that yield the highest score for the objective function. In the case of selecting protein properties, the objective function can be designed to evaluate properties based on their relevance to the problem in question.

2.4.2.2.1 Minimum Redundancy Maximum Relevance (mRMR)

mRMR method was proposed in (Peng, et al., 2005) to select subset of properties that are non-redundant (i.e. dissimilar to each other) and at the same time have maximum relevance to the prediction class of the sample. Following (Peng, et al., 2005) the minimum redundancy is defined as:

$$\min R(S) = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j) \quad (2.7)$$

Here $I(x_i, x_j)$ is the mutual information between features i and j , and S is the subset of features (properties) that we want to select, and $|S|$ represents the total number of features in the subset S . The maximum relevance aims to maximize the total relevance of the selected features to the sample class label. It is defined by:

$$\max D(S, c) = \frac{1}{|S|} \sum_{x_i \in S} I(x_i, c) \quad (2.8)$$

where $c = (c_1, c_2, \dots, c_i)$ is the class label for the sample. Then, the mRMR is defined as the maximum difference between (2.8) and (2.7) as follows:

$$\text{mRMR} = \max(D - R) \quad (2.9)$$

2.4.2.2.2 Genetic Algorithms

Genetic algorithm (GA) is a heuristic search – based optimization method that uses historical information about previous population to produce a new population. GA concept is motivated by natural evolution, so it is part of a broad range of Evolution Algorithms (EA).

Initially, a GA algorithm generates a random population (set of random solutions) that consists of different hypotheses (members of the population). Then, in each iteration, each hypothesis of the population is assigned a score using a fitness function. Part of the population which have the highest fitness score are used to generate the new population after applying crossover and mutation operations. The GA algorithm is described in Table 2.2 following (Mitchell, 1997):

Table 2.2. Genetic Algorithm. Pseudo code for the GA algorithm as described in (Mitchell, 1997)

Let:

P is the initial population consists of n hypotheses (members h)

r is the fraction of the population to be replaced by crossover operation

m is the mutation rate

Initialization: Generate a population (P) consists of (n) random hypotheses (h)

Evaluation: For each hypothesis in P compute $Fitness(h)$

While ($\max_h Fitness(h) < stop_threshold$)

Generate new population P_s as follows:

1. **Select:** $(1-r)$ n members of P to be added to the new generation of population

such that:
$$Pr(h_i) = \frac{Fitness(h_i)}{\sum_{k=1}^n Fitness(h_k)}$$

2. **Crossover:** select $\frac{r * n}{2}$ pairs of hypotheses from P , and for each pair of hypotheses (h_i, h_j) apply Crossover operator to generate two offspring which later is added to P_s
3. **Mutate:** for each of $(m * n)$ hypotheses of P_s selected at random, mutate a random bit in its representation
4. **Update:** $P \leftarrow P_s$
5. **Evaluate:** For each hypothesis in P compute $Fitness(h)$

Return the hypothesis having the highest $Fitness(h)$

The flowchart for this algorithm is shown in Figure 2.3:

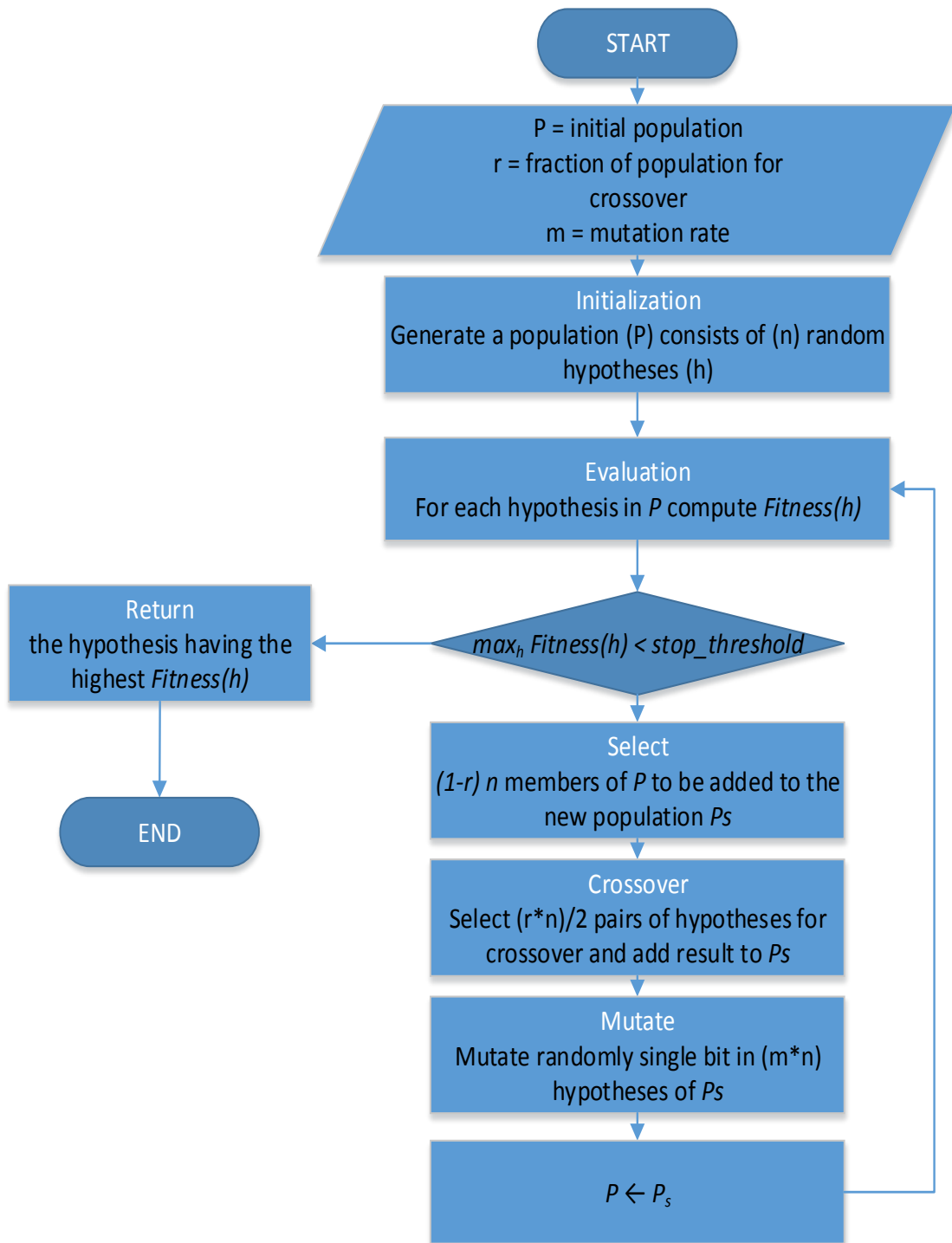
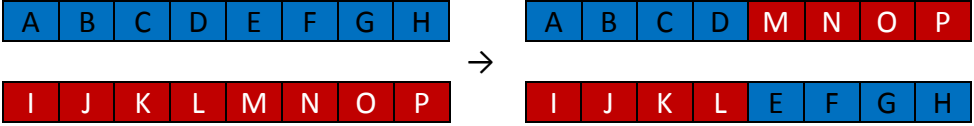
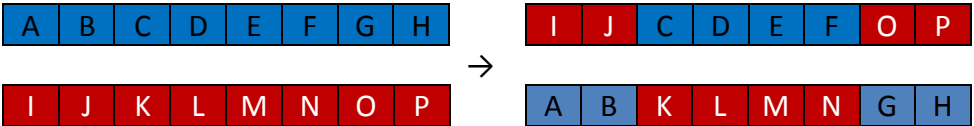
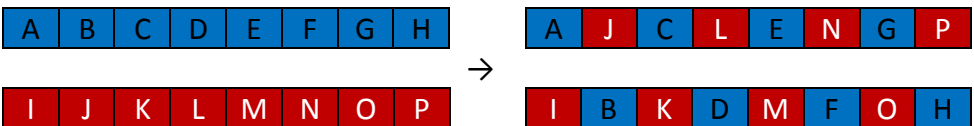


Figure 2.3. Genetic Algorithm Flowchart. A flowchart shows different steps of the genetic algorithm.

The crossover operation aims to create in the new population from a pair of parental hypotheses a pair of successor hypotheses, called “offspring”. This is achieved by exchanging bits between the parents to make the offspring. There are different techniques for making the crossover as summarized in Table 2.3.

Table 2.3. Crossover in genetic algorithm. Three common types of crossover operation during the genetic algorithm processing.

Single-point crossover	
Two-point crossover	
Uniform crossover	

To ensure diversity among the hypotheses of the new generation, the mutation method is implemented. A simple form of the mutation is the “Single bit mutation” where a single bit is flipped in the new generation, as follows:

1 1 0 1 0 0 0 1 → 1 1 0 0 0 0 0 1

2.5 Machine Learning Classification Models

Given a collection of labeled samples represented by feature vectors (training data), the classification task aims to build a model for the class label of the data as a function of features used to compile feature vectors. Later, this learned model is used to predict the class label for previously unseen samples (testing data). There are numerous classification models, such as, Decision Trees (Quinlan, 1987), Neural Networks (NN) (Rumelhart, et al., 1986), Bayesian Network Classifiers (Friedman, et al., 1997), K-Nearest Neighbors (Altman, 1992), Support Vector Machines (SVM) (Cortes and Vapnik, 1995), etc..

2.5.1 K-Nearest Neighbors

The K-Nearest Neighbors method is a simple classification algorithm where a test sample belongs to the same class as its K-nearest samples. The determination of the k-nearest samples to the test sample is performed by computing the distance between the test sample and each of the training samples. Euclidean distance is an example of such distance metric between two samples A and B, defined as in the following:

$$Distance(A, B) = \sqrt{\sum_{i=1} (A_i - B_i)^2}, \quad (2.10)$$

2.5.2 Artificial Neural Networks

The Artificial Neural Network (NN) model is a collection of nodes (arranged in layers) and connected with weighted links. In the example in Figure 2.4 in the output node its input the weights of its input links are summed up generating an output signal:

$$f(W, X) = \left(\sum_i w_i X_i \right), \quad (2.11)$$

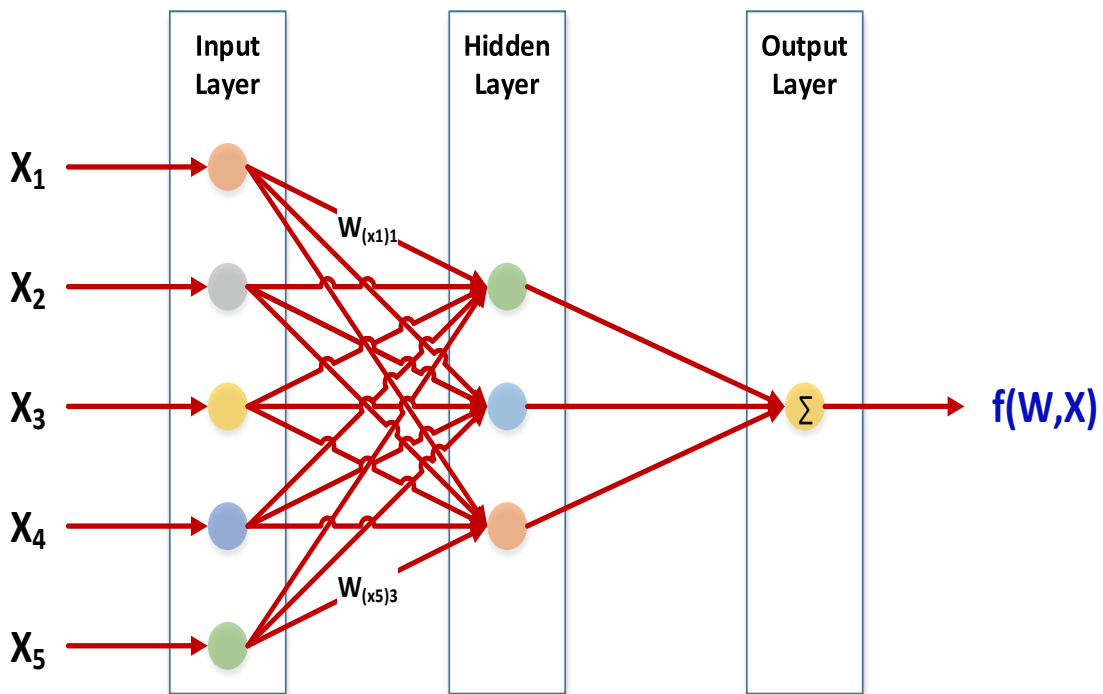


Figure 2.4 An Example of a Neural Network Structure. A typical NN is composed of different layers, each of which consists of several nodes. Nodes within different layers are connected by links that have assigned weights.

In the beginning, the NN learning algorithm initializes the weights w_i of the links in the NN. Then, it adjusts the weights to minimize the following objective function:

$$E = \sum_i (Y_i - f(w_i, X_i))^2, \quad (2.12)$$

Where Y_i represents the desired output of the NN and $f(w_i, X_i)$ is the predicted value from the NN when the input is X_i . This objective function can be optimized using different algorithms. The back-propagation algorithm is widely used for this purpose.

2.5.3 Deep Learning

Deep learning (Deng and Yu, 2014) is a branch of machine learning that aims to model data abstraction by building a complex structure composed of numerous processing layers. The deep learning may discover the information structures within the large datasets by building distributed representation for the data. Deep learning proved to improve in some cases the performance in different applications based on large datasets, such as, speech recognition and image processing.

2.5.4 Ensemble Methods

The idea of the ensemble methods is to construct a set of classifiers from the training data and then to combine their predictions into the final prediction. An apparent advantage of the ensemble of classifiers is the frequently the overall improvement of prediction performance. There are different types of the ensemble methods. The first is based on the Bagging (Bootstrap) algorithm. In the Bagging algorithm, N classifiers are

trained on subset of samples (sampling with replacement) and all classifiers are trained on the same set of features (M). A test sample is assigned to the class with the most (or averaged) votes by the N classifiers.

The second type of ensemble methods is Random Subspace Method, which in training uses subset of features (m) for a subset of the samples. A special case of it is known as the Random Forest method (Ho, 1998) where decision trees are used together with the whole set of training samples but with different features for each decision tree.

The third type of ensemble methods is Boosting which focuses on the misclassified samples. Initially, all samples are equally weighted and trained by a classifier. In the second phase, another classifier is trained on the samples but the misclassified samples by the first classifier are given higher weight. This process continues to create N classifiers. The prediction of a test sample is the weighted sum of all the predictions of the N classifiers. Boosting yields higher accuracy than bagging, however, it has a risk of overfitting the data.

2.5.4.1 Random Forests

Random forests are simply an ensemble of decision trees. The training data is divided between N trees (Figure 2.5) and each tree is trained with subset of the data. This division on the data occurs on the features (not the samples), such that each tree received the entire training data for M randomly selected features (with replacement). In the testing phase, the class prediction of an unknown sample is provided by each tree

and the random forest counts the votes and assigns the class label to the most voted class. As most of the ensemble methods, random forests usually yield better accuracy than other classifiers. It is fast and can handle high dimensional data easily.

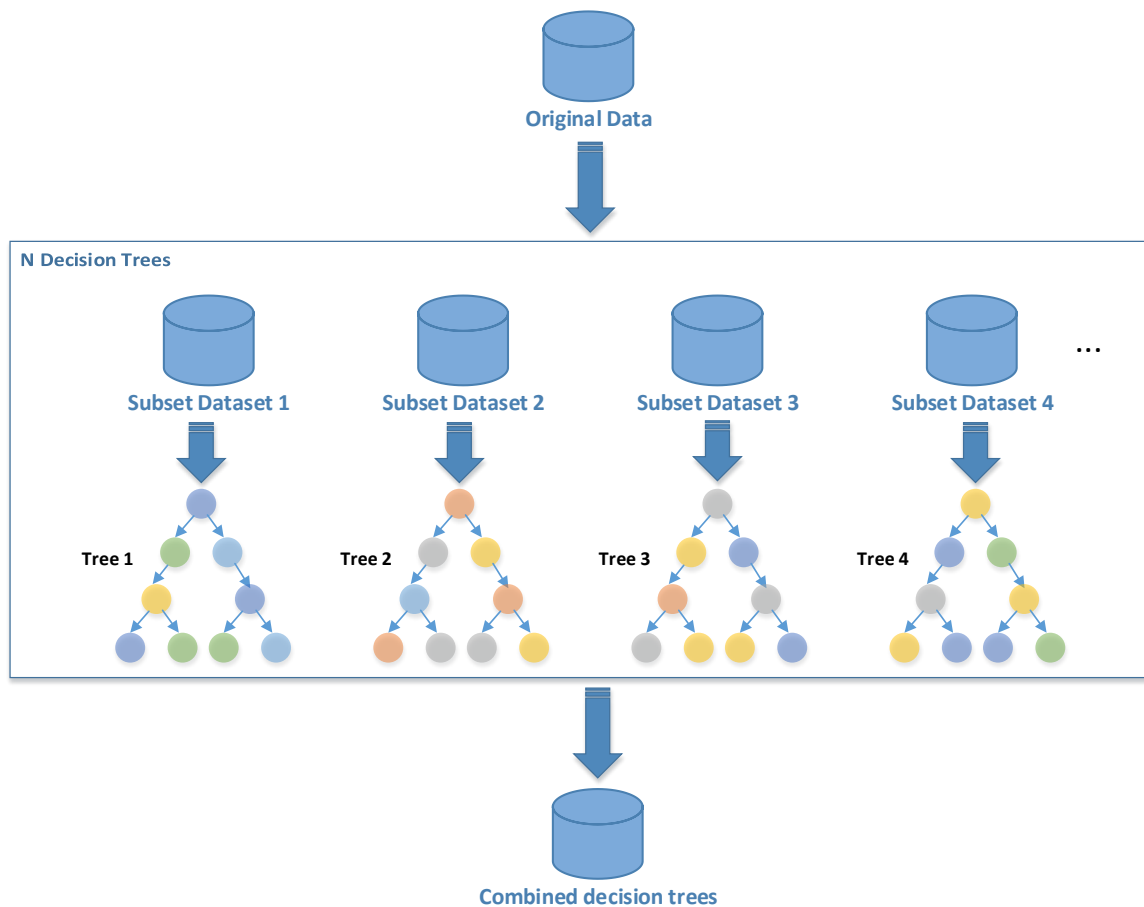


Figure 2.5 Illustration of a Typical Structure of Random Forests. The Random Forest model is an ensemble of N trees, each of which learn from data using part of the features. All predictions from these sub-trees are combined to produce the Random Forest model prediction.

Chapter 3

Distinct Profiling of Antimicrobial Peptides Using their Compositional and Physicochemical Properties

This chapter is based on a published study:

Distinct profiling of antimicrobial peptide families (Khamis AM, Essack M, Gao X, Bajic VB) (2015) Distinct Profiling of Antimicrobial Peptide Families. *Bioinformatics* 31(6):849-856. doi:10.1093/bioinformatics/btu738

3.1 Summary

The spread of multi-drug resistant (MDR) pathogens raised the demand to design novel antimicrobial peptides (AMPs). AMPs kill rapidly and show broad-spectrum efficiency against MDR pathogens, therefore making AMPs being considered as a possible complement and substitute for conventional antibiotics. It is challenging to use current *in-silico* techniques to design novel AMPs due to numerous design parameters, absence of models, testing procedures, cost and production time of the design. AMPs are categorized into families/sub-families based on their sequences, 3D structures and functions. The capability to identify properties that can discriminate each of the AMP families from all other AMPs can facilitate discovering the main characteristics of these AMP families. This will help in the *in-silico* design of synthetic AMPs. In this study we considered 14 different AMP families and sub-families. We described all peptides that belong to these 14 different AMP families and sub-families using their amino acid sequence compositional and physicochemical properties. Then, using Genetic Algorithm

(GA), we identified the subset of properties that accurately discriminate each of the AMP families from all other AMPs with an average specificity, sensitivity and precision of 99.86%, 92.88% and 95.96%, respectively. We further explored the selected discriminative properties using information available in literature and found that many of them have been shown to be characteristics (either compositional or functional) of these AMP families. This set of the selected properties could assist as guides for novel synthetic AMPs *in-silico* designs. The method that we developed is generic as it can be applied to characterize different groups of properties families. This work was published in (Khamis, et al., 2015).

3.2 Introduction

With the evolution and rapid increase of multi-drug resistant (MDR) pathogens, it has become important to look for alternatives medications that can effectively attack these pathogens (Saxena and Gomber, 2010). Due to the broad-spectrum activity against Gram-positive and Gram-negative bacteria, protozoa, fungi and many enveloped viruses (Hancock and Scott, 2000; Thomas, et al., 2010; Zasloff, 2002), antimicrobial peptides (AMPs) have been considered to have potential use in treatment of MDR pathogens because of their effectiveness on wide range of pathogens and the ability to rapidly kill these microbes (Brahmachary, et al., 2004; Gordon, et al., 2005; Hancock and Diamond, 2000; Hancock and Lehrer, 1998; Hancock and Sahl, 2006; Peters, et al., 2010; Sundararajan, et al., 2012; Thomas, et al., 2010). Synthetic AMPs were also found to destroy MDR pathogens (Yeaman and Yount, 2003). AMPs permeate and penetrate the

target pathogen (Andreu and Rivas, 1998; Brogden, 2005; Radek and Gallo, 2007), consequently disrupting the formation of the cytoplasmic membrane septum, which leads eventually to killing the microbe (Chen, et al., 2012; Fjell, et al., 2012; Frecer, et al., 2004; Peters, et al., 2010). AMPs can easily differentiate between membrane structure of pathogens and other organisms because of the differences between them in the membrane structure. This makes microbes to be simply known and targeted by AMPs (Thomas, et al., 2010).

Typically, AMPs have relatively short sequence of fewer than 100 amino acids in length (Jenssen, et al., 2006; Pasupuleti, et al., 2012; Peters, et al., 2010; Sang and Blecha, 2008). Majority of these amino acids are cationic and amphipathic (Epand and Vogel, 1999; Lehrer and Ganz, 1999). While AMPs are evolutionarily conserved (Hancock and Sahl, 2006; Yeaman and Yount, 2003), minor variations in their structure can result in major differences in their functional characteristics (Ganz, 2003; Thomas, et al., 2010). Based on their primary sequence and structure, AMPs were categorized into families/sub-families (Kaiser and Diamond, 2000; Yeaman and Yount, 2003).

Several databases focused on AMPS were built and contain hundreds of AMPs, e.g. DAMPD (Sundararajan, et al., 2012), APD2 (Wang, et al., 2009) , CAMP (Thomas, et al., 2010; Waghu, et al., 2014), ANTIMIC (Brahmachary, et al., 2004) and APD (Wang and Wang, 2004). In spite of these numerous databases, the increased demand for AMPs, boosted increased attention to the *in-vitro* design of synthetic AMPs (Marcos, et al., 2008; Nusslein, et al., 2006). It is challenging to design AMPs because of the many

properties that characterize AMP families (Guralp, et al., 2013). Consequently, *in-silico* method is recommended to subselect the properties that can serve as a guide for the design process of novel AMPs (Juretic, et al., 2011). Such computational approaches to identify the properties that characterize different AMP families have gained more interest (Fjell, et al., 2012; Maccari, et al., 2013).

While, several methods were proposed to characterize AMPs using their amino acid compositional characteristics and sequence alignment information (Lata, et al., 2010; Lata, et al., 2007; Wang, et al., 2011), such information is not sufficient to characterize AMPs. Rather, the structural and physicochemical properties of AMPs can provide more insight about the activities of these AMPs (Fjell, et al., 2012). Consequently, several studies were performed on the physicochemical properties of AMPs (Langham, et al., 2008; Maccari, et al., 2013; Porto, et al., 2010; Torrent, et al., 2011). Though, to the best of our knowledge, a comprehensive study of these properties, their effectiveness within different regions of AMP sequences, and using these properties to characterize AMP families is yet to be done.

In this study, a novel computational model is proposed to represent AMP amino acid sequences and a method to determine the most discriminative properties (either compositional or physicochemical) of AMPs that can discriminate each of the AMP families from all other AMPs. For this purpose, we used Genetic Algorithms (GA) to optimize an objective function based on unsupervised clustering. Using the selected properties, enabled high accuracy discrimination of the AMP families and for most of

the AMP families, 100% specificity in peptides separation of a particular AMP family from all other AMPs has been achieved. This is associated with sensitivity range from ~62% to 100%. We found although not all many of the selected properties were reported in the literature as functional or compositional characteristics of the associated AMP family. These findings suggest that the identified properties can serve as a guide during the *in-silico* design process of synthetic novel AMPs.

3.3 Methods

3.3.1 Datasets

We obtained the sequences of 753 non-redundant mature and natural peptides from the DAMPD database (Sundararajan, et al., 2012). These peptides belong to 128 families/sub-families. We selected DAMPD database because all the deposited peptides were curated manually and previously validated experimentally to have antimicrobial activity. These peptides were classified into families and sub-families using the UniProt (UniProt, 2014) annotation. From the initial set of AMPs we eliminated all families/sub-families that have less than 10 peptides. The remaining AMP families we name “target AMP families”. As a result of this filtration process, we considered 14 target AMP families containing a total of 465 peptides. The remaining 114 AMP families/sub-families with total of 288 peptides served as a portion of the “negative dataset”. For each target AMP family, its negative dataset is specific. That is, the negative dataset contains in addition to the above-mentioned 288 peptides determined by the filtering process, all peptides from the other 13 target AMP families (where target AMP family peptides are

eliminated). We used these negative datasets within the iterative procedure to identify the properties that characterize each AMP family. That is, in each iteration of the GA optimization, an unsupervised clustering is performed to compare the peptides of a specific AMP family to its negative dataset composed of all other AMPs. The number of peptides in each of the 14 AMP families/sub-families is shown in Table 3.1, and the composition of amino acids of these families/sub-families is provided in Appendix 1.

Table 3.1. Number of peptides in each of the 14 AMP families/sub-families.

Distribution of the obtained peptides from DAMPD database among 128 AMP families and sub-families.

AMP family/sub-family	Number of peptides
Alpha-defensin	34
Bacteriocin	24
Beta-defensin	41
Bombinin	31
Cathelicidin	27
Cecropin	30
Cyclotide (Bracelet sub-family)	12
DEFL	37
FSAP (Brevinin sub-family)	143
FSAP (Caerin sub-family)	11
FSAP (Dermaseptin sub-family)	30
Invertebrate defensin (Type 1 sub-family)	21
Invertebrate defensin (Type 2 sub-family)	13
Type A lantibiotic	11
All Other AMP Families with less than 10 peptides (114 Families/sub-families)	288
Total	753

For the purpose of comparing AMP vs. non-AMPs, a set of non-AMPs were compiled as in the following:

1. We extracted from UniProt sequences of all proteins whose length was between 10 and 212 amino acids (the same AMP length interval as in our datasets). We retained the sequences whose ontology annotation keywords do not contain any keyword connected to antimicrobial activity (e.g. Antifungal, Antimicrobial, Fungicide, Antibacterial, Antiviral, Defensin, etc.). This resulted in 18,082 non-AMP sequences.

2. To remove identical sequences and reduce redundancy among the selected sequences, we performed clustering by the h-cd-hit tool (Li and Godzik, 2006) in three steps. We used the three identity thresholds ($\geq 90\%$, $\geq 60\%$, $\geq 30\%$). After this, 7,066 non-AMP sequences were obtained.

3. Finally, we removed all sequences that contain any of the non-canonical amino acids (e.g. nonstandard letters O, B, J, X, U and Z). This yielded 6,740 non-AMP sequences.

3.3.2 Peptide/Protein Sequence Models

The representation method that we developed for peptides/proteins is described in what follows. We adopted the protein representation method used to describe numerically proteins of different lengths for the purpose of identifying protein cellular localization (Matsuda, et al., 2005). In this model, different proteins will differ in the model parameters inferred from properties of amino acid composition within different protein regions. We extended Matsuda et. al. model by adding to this representation a

description for the physicochemical properties in different protein sequence regions. That is, in each region identified by the model from Matsuda et. al., we added for that regions its restrictive physicochemical properties (these properties are characterized as being highly invariant within peptides that belong to the same family). Then, using the GA optimization, we selected subset of the properties from the entire description that best discriminate AMP families. Consequently, as explained above, the method that we developed can handle different lengths of protein/peptide sequences and it can represent the sequencing using compositional and physicochemical properties extracted from different regions.

This method is explained as follows. We divided AMP amino acid sequences into three regions. This is supported by the fact that there are numerous properties enriched within the N and C terminals of the amino acid sequences that discriminate peptides of a particular AMP family all other AMP families (Hayes, et al., 2006; Lata, et al., 2010; Minervini, et al., 2003). Then, we encoded each peptide using a two-part feature vector. The first part of the feature vector represents mostly the composition features of the peptide amino acid sequence. The second part contains the restrictive physicochemical properties within predetermined regions in the peptide sequence. We explain each feature vector part as in the following.

Basic Peptide Features Representation. Peptides from all AMP families were represented by the same set of features. As mentioned above, we borrowed the method for protein sequence representation in (Matsuda, et al., 2005). In this representation, the amino

acid sequence of each peptide was separated into three regions, N-terminal (N), middle region (M) and C-terminal (C). Then, the N-terminal is also separated into four sub-regions (n1, n2, n3 and n4). Determination of each region size is based on the sequence length L . Consequently, long N and C termini are assigned for long sequences and vice versa for short sequence. Because AMP peptides have typically shorter protein sequences than most of the other groups of proteins and also because peptides that belong to different AMP families differ in their sequence length, we tested the performance of the model that we developed with different d_n values (length of one sub-region from the four N-terminal sub-regions) and d_c (length of C-terminal) parameters for every AMP family. We selected the values of d_n and d_c (Table 3.2 and Appendix 2) that were experimentally found to be the most suitable for discriminating the AMP families. Then, 184 features were used to represent the peptide sequence as follows: 140 features were used to represent the composition of amino acids in all the regions/sub-regions n1, n2, n3, n4, M, C and the whole sequence (20 features in 7 regions); 20 features for the twin amino acids composition (two consecutive of the identical amino acids, e.g. MM, KK) in the M-region; 6 features for the frequency of different distances of basic amino acids (K, H and R) in each of N and M regions providing altogether 12 features; additional 6 features for the frequency of different distances of hydrophobic amino acids (A, F, G, I, L, M, P, V and W) in the M-region; the latest 6 features were used to represent for the frequency of different distances of other amino acids (C, D, E, N, Q, S, T and Y) in the M-region. The 6 values of frequency of different distances are calculated as in the subsequent text. The distance (H) between

two consecutive amino acids in the stated class is assigned to one of six categories of distance ($H = 1$, $1 < H \leq 6$, $6 < H \leq 11$, $11 < H \leq 16$, $16 < H \leq 21$, $H > 21$). After that, frequency of different distances are represented by the number of incidences of the distances in these six groups. For instance, the distances between basic amino acids (K, H and R) in the sequence (MRAMRSKNNGGNPAKHMTTNNAK) are 3, 2, 8, 1, and 7. The 6 frequency values of these distance values are (1, 3, 2, 0, 0, and 0). This representation from (Matsuda, et al., 2005) can capture the signal sequences from different peptide sequence regions.

Addition of Family Specific Properties. Family specific properties refer to a collection of restrictive properties that exist within different regions of the peptide sequence of the considered AMP family. This set of properties differs from region to region within the sequence. To select properties that are restrictive with different regions, 544 amino acid physicochemical properties provided in the AAIndex (version 9.1) database (Kawashima and Kanehisa, 2000) were used. We reduced this initial set to 294 features by eliminating properties that have mutual Pearson correlation coefficient of 0.9 or higher and selecting randomly selected one property from such mutually highly correlated properties. Then, to extract the features, first we used progressive multiple sequence alignment algorithm (Thompson, et al., 1994) to align all peptide sequences of a particular family. After that, from the aligned peptide sequences, we identified the restrictive physicochemical properties by removing from a particular family one peptide at a time. After that, we tested if the value of the property for all amino acids of the

removed peptide (for each region of the n1, n2, n3, n4, M and C regions) is inside the min/max values range for the same property (min/max values are defined from the other peptides of the same family in that region). We repeated this test for all peptides (using leave one out approach). After that, the physicochemical properties that we found conserved/restricted in at least 90% of peptides for a specific region particular family were used to represent that region. Finally, we used in the feature vector the median values of these properties for individual amino acids in the region. The following algorithm explains this method for selecting the region specific restrictive physicochemical properties of AMP peptides of a specific family.

```

Let N be the set of L peptide sequences in AMP family  $F_i$  such that  $N=(N_1, N_2, \dots, N_L)$ 
Let  $T_i$  be one of the peptide regions  $T=(n1, n2, n3, n4, M, C)$ 
Let  $P_j$  be amino acid property in  $P=(P_1, P_2, \dots, P_{294})$ 
For each region  $T_i$  in T
  Scores=[]

  For each property  $P_j$  in P
    For each sequence  $N_k$  in N
      Let S be all sequences of  $T_i$  region in N excluding  $N_k, S=(x_i, x_i \in N \ \& \ x_i \neq N_k)$ 
      Let Mn be a vector of minimum values of  $P_j$  within individual positions of S
      Let Mx be a vector of maximum values of  $P_j$  within individual positions of S
      If property value  $P(N_{ka})$  of every amino acid  $N_{ka}$  in  $T_i$  region of  $N_k$  satisfy the following
         $P(N_{ka}) \geq Mn_a$  and  $P(N_{ka}) \leq Mx_a$ , then
          Scores[j, k]= 1
        Otherwise
          Scores[j, k]= 0
      END IF
    END FOR
  END FOR
END FOR

```

Let R be the vector of overall scores such that $R_j = (\sum_{k=1} Scores[j,k])/k$ for each property P_j

Select as restrictive properties in T_1 to be those which have overall score $R_j \geq 0.9$

END FOR

3.3.3 Data Preparation

Normalization. Each feature was normalized to eliminate the bias that comes from variant ranges of values as in the following:

$$x'_i = \frac{x_i - u_i}{\sigma_i} \quad (3.1)$$

where x_i is the original value for the feature, and x'_i is the normalized value; u_i is the mean value of the x_i values for all AMP peptides, and σ_i refers to the standard deviation.

Filtering Data. We excluded from the feature vectors the features that have constant values in all peptides of AMP families.

Target and Non-Target Classes. To identify the most discriminant properties of a particular AMP family, we performed unsupervised clustering using k-means algorithm. All the peptides that belong to the target AMP family represent the target class, whereas the peptides from other AMP families (i.e. the remaining 13 target AMP families in addition to the 291 peptides from the other 114 AMP families) represent the non-target class. We repeated this process independently for each of the 14 target AMP families.

3.3.4 AMP Family-Specific Feature Selection

We selected compositional and physicochemical properties that are most discriminant for a particular AMP family using global optimization based on Genetic Algorithm (GA).

This optimization aims to minimize the following fitness function:

$$F = 1 - Fmeasure + Regularization \quad (3.2)$$

where

$$Fmeasure = \frac{2 * (precision * recall)}{(precision + recall)} = \frac{2 * TP}{2 * TP + FN + FP} \quad (3.3)$$

$$Regularization = \frac{Number\ of\ Selected\ Features}{Total\ Number\ of\ Features} \quad (3.4)$$

We ran the GA with the following settings. The population size was 1000 and the number of generations were 1000 as well. The mutation rate was set to 0.01 and the crossover rate was 0.8. The F-measure in the fitness function is calculated for each individual in every generation. This is performed using unsupervised clustering with k-means clustering algorithm using Euclidean distance. Evaluation of the clustering performance was performed using the known class label of each peptide by computing the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). The F-measure is calculated from these quantities. Because k-means algorithm is sensitive to the selection of initial cluster centroids, we removed the bias that may arise from this random selection by initializing the cluster centroids with the actual mean values in the positive and negative classes. The k-means clustering was performed using

different number of clusters (2 to 15 clusters). We selected the number of clusters that provided the highest F-measure. In the case of optimum clustering, peptides that belong to a particular target AMP family will be grouped together in one cluster (target class) while all peptides from all other AMP families will reside in one or more other clusters (non-target class clusters). We added the regularization constraint to the optimization function to help the GA to select the minimum set of features that provided highest F-measure value. The purpose of this study is not to build a predictive AMP model, however, we use this procedure to profile AMP families by identifying the properties that accurately discriminate AMP families.

3.3.5 Clustering AMPs into Antimicrobial Families

After selecting the properties, we examined their capability to group the peptides of each AMP family and discriminate that family from all other AMPs. For a particular AMP family X that we want to evaluate, all peptides of all AMP families were represented using the selected properties for the family X. Then, k-means clustering was performed. In the case of perfect clustering, all peptides that belong to the family X will be grouped in one cluster while all other peptides that belong to other AMP families will be grouped in one or more of the other clusters. We used this criterion to measure the performance of the clustering based on the selected properties.

3.3.6 Evaluation of Model Results

After performing the clustering of AMPs, we selected the cluster that contain the maximum number of target class peptides as the cluster of the target class. All the remaining non-target clusters may group different AMPs from either non-target or target class. We used accuracy, sensitivity, specificity, precision, Jaccard Index and F-measure to evaluate the performance of the clustering. The measures are defined as in the following:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (3.5)$$

$$Sensitivity (Recall) = \frac{TP}{TP + FN} \quad (3.6)$$

$$Specificity = \frac{TN}{FP + TN} \quad (3.7)$$

$$Precision = \frac{TP}{TP + FP} \quad (3.8)$$

$$Jaccard Index = \frac{TP}{TP + FP + FN} \quad (3.9)$$

$$Fmeasure = \frac{2 * TP}{2 * TP + FN + FP} \quad (3.10)$$

TP represents the number of peptides from the target class in the target class cluster; TN represents the number of peptides from the non-target class in the non-target class clusters; FP represents the number of peptides from the non-target class that belong to the target class cluster; and FN represents the number of peptides from the target class that belong to the non-target class clusters.

We added to the evaluation measures the entropy and purity measures defined in (Tan, et al., 2006). The entropy is calculated as follows. First, the probability that a peptide in cluster i is a member of class j is defined as:

$$p_{ij} = \frac{m_{ij}}{m_i}, \quad (3.11)$$

where the number of objects in cluster i were represented by m_i . The number of objects of class j in cluster i were represented by m_{ij} . Then, the entropy of each cluster i is:

$$e_i = -\sum_{j=1}^L p_{ij} \log_2 p_{ij}, \quad (3.12)$$

where L is the number of classes (in our case we have two classes, target and non-target). The total entropy of a set of clusters is:

$$e = \sum_{i=1}^K \frac{m_i}{m} e_i, \quad (3.13)$$

where m is the total number of peptides and K is the number of clusters. The purity is calculated as follows:

$$p_i = \max_j p_{ij} \quad (3.14)$$

$$\text{purity} = \sum_{i=1}^K \frac{m_i}{m} p_i \quad (3.15)$$

We used MATLAB (MATLAB, 2012) to develop all parts of this computational model.

3.4 Results

3.4.1 Using Global Optimization of Unsupervised K-means Clustering for AMP Feature Selection

Two sets of features were used to describe peptides of a particular AMP family (see Methods). 184 features in the first set represent mainly the composition of amino acids in different regions of the AMP peptide sequences. The second set of features represent restrictive physicochemical properties within different regions of the AMP peptide sequences. Because features have different contribution levels to the discrimination of a particular AMP family from all other AMP families, we performed feature selection using a GA for global optimization of unsupervised clustering. As compared to other optimization methods such as Particle Swarm Optimization (PSO) (Kennedy, 2010; Kennedy and Eberhart, 1995) and Differential Evolution (DE) (Chakraborty, 2008), the

GA has better performance than these methods and is more suitable for optimization of discrete variables (see Section 3.4.3). Also, we compared k-means clustering algorithm with other algorithms, e.g. Affinity Propagation (AP) (Frey and Dueck, 2007) and found that it provided better clustering results than AP (Section 3.4.4). In addition, we compared different distance measures (Euclidean, correlation, cosine and city block) within the k-means clustering and found that Euclidean distance yielded better performance than other measures (Section 3.4.5). Because AMP sequences have variant lengths (Figure 3.1), we used different lengths for the N and C terminal to perform feature selection for each AMP family, i.e., $d_n = 10, 12, 14, 16$ and $d_c = 8, 10$. The clustering performance of different settings for the terminal length parameters, d_n and d_c , is shown in (Appendix 2). Furthermore, we used for the feature selection different number of clusters ($K = 2, 3, \dots, 15$). Columns 2-7 in Table 3.2 contain parameter values that provided the highest value for the F-measure.

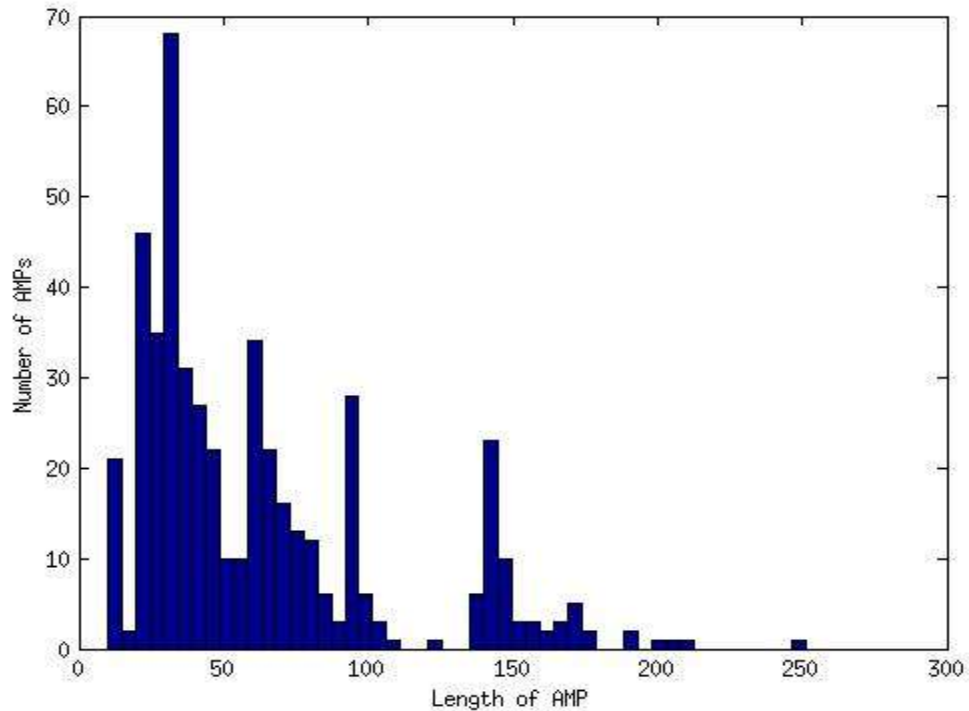


Figure 3.1. AMP length distribution. A histogram of the distribution of lengths of 478 AMPs from 14 target AMP families.

As shown in Table 3.2, a subset of 394 features from the total set of 9162 features used to represent peptide sequences of all 14 target AMP families, were selected to distinguish each AMP family peptides from any other AMP.

Table 3.2. The number of selected properties for AMP families. The number of features that characterize different regions of AMP sequence is selected using GA optimization of unsupervised k-means clustering. Annotations of columns are as follows: N-terminal length (dn), C-terminal length (dc), number of peptides (NP), original number of features

(NF), number of selected features (NSF), number of clusters (NC), number of compositional features (NCF) and number of physicochemical features (NPF).

AMP family/sub-family	dn	dc	NP	NF	NSF	NC	NCF	NPF
Alpha-defensin	10	10	34	299	14	14	12	2
Bacteriocin	14	10	24	225	9	12	7	2
Beta-defensin	10	10	41	261	36	14	29	7
Bombinin	16	10	31	1095	13	9	4	9
Cathelicidin	16	8	27	521	36	11	17	19
Cecropin	12	8	30	835	33	11	8	25
Cyclotide (Bracelet)	12	10	12	350	7	14	2	5
DEFL	12	8	37	252	26	14	20	6
FSAP (Brevinin sub-family)	14	10	143	1945	118	12	18	100
FSAP (Caerin sub-family)	14	10	11	1946	28	14	1	27
FSAP (Dermaseptin)	16	10	30	327	25	15	16	9
Invertebrate def. (Type 1)	10	10	21	402	14	14	8	6
Invertebrate def. (Type 2)	16	8	13	510	9	15	4	5
Type A lantibiotic	16	10	11	194	26	14	26	0
Total				9162	394		172	222

3.4.2 Using Selected Feature Subsets to Cluster AMPs

We performed clustering using the subset of selected features that characterize each target AMP family. Then, we tested if the peptides belong to that family will fall in one cluster and other AMPs will be grouped in other clusters. The performance of the clustering was measured using the known labels of the peptides. The clustering results for each target AMP family are shown in Table 3.3. As noticed in Table 3.3, for 9 of the 14 AMP families, 100% specificity was achieved in the clustering, while the specificity of

the other five AMP families ranged between 99.17% and 99.86%. This provides an evidence of the capability of the selected features to distinguish each of the AMP families. The obtained values of sensitivity, specificity and precision averaged for all 14 target AMP families were 92.88%, 99.86% and 95.96%, respectively. The sensitivity vs. precision results for the clustering using the selected features are shown in Figure 3.2.

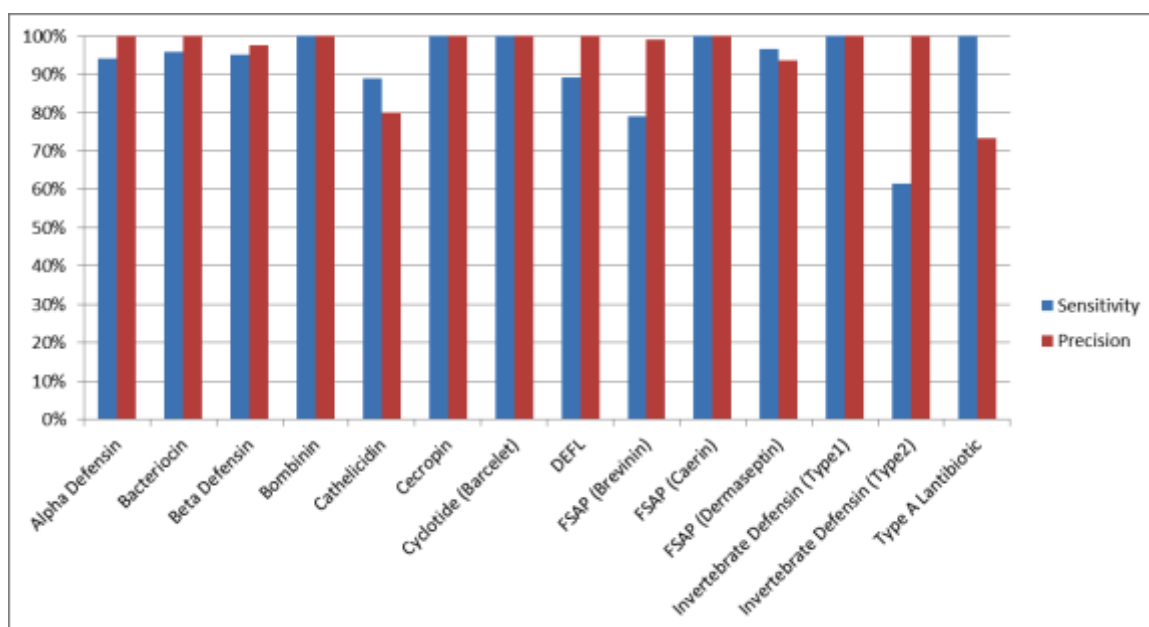


Figure 3.2. K-means clustering performance (sensitivity vs. precision). Plot of sensitivity vs. precision obtained from k-means clustering of 14 target AMP families using optimized set of selected features.

Results in Table 3.3 shows the advantages of using physicochemical properties to characterize peptide sequences. The long distance interactions in the peptide sequence between residues and the similarity in function between distantly related proteins are likely captured at least partly using physicochemical properties (Du and Li, 2006; Liu, et

al., 2012). Consequently, the information of amino acid composition only might be insufficient to accurately distinguish AMP families, specify their target cell interaction, and characterize their activities (Maccari, et al., 2013; Pushpanathan, et al., 2013). To test the importance of characterizing AMP peptide families using physicochemical properties, we used 184 features that represent mainly amino acid composition (part1 of the feature vector only) to represent AMP peptides. Table 3.4 shows the clustering performance using this representation. The average obtained values of sensitivity, specificity and precision for all 14 target AMP families were 73.08%, 94.50% and 40.65%, respectively. We notice that both of the sensitivity and precision are significantly weaker than when using the restrictive physicochemical and compositional properties selected with GA (Table 3.3).

Table 3.3. K-means clustering performance using selected properties by GA algorithm.

K-means clustering performance of 14 target AMP families using GA selected features.

Target AMP family	Number of features	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
Alpha-defensin	14	99.73%	94.12%	100.00%	100.00%	94.12%	96.97%	0.0165	99.73%
Bacteriocin	9	99.87%	95.83%	100.00%	100.00%	95.83%	97.87%	0.0116	99.87%
Beta-defensin	36	99.60%	95.12%	99.86%	97.50%	92.86%	96.30%	0.0289	99.60%
Bombinin	13	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
Cathelicidin	36	98.80%	88.89%	99.17%	80.00%	72.73%	84.21%	0.058	98.80%
Cecropin	33	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
Cyclotide (Bracelet subfamily)	7	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
DEFL	26	99.47%	89.19%	100.00%	100.00%	89.19%	94.29%	0.0325	99.47%

FSAP (Brevinin subfamily)	118	95.88%	79.02%	99.84%	99.12%	78.47%	87.94%	0.1478	95.88%
FSAP (Caerin subfamily)	28	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
FSAP (Dermaseptin subfamily)	25	99.60%	96.67%	99.72%	93.55%	90.62%	95.08%	0.0254	99.60%
Invertebrate defensin (Type 1 subfamily)	14	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
Invertebrate defensin (Type 2 subfamily)	9	99.34%	61.54%	100.00%	100.00%	61.54%	76.19%	0.0426	99.34%
Type A lantibiotic	26	99.47%	100.00%	99.46%	73.33%	73.33%	84.62%	0.0167	99.47%
Average		99.41%	92.88%	99.86%	95.96%	89.19%	93.82%	2.71%	99.41%

Table 3.4. K-means clustering performance using amino acid composition properties.

The k-means clustering performance of 14 target AMP families using 184 features that represent information about amino acid composition.

Target AMP family	Number of features	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
Alpha-defensin	184	96.55%	85.29%	97.08%	58.00%	52.73%	69.05%	0.1085	96.55%
Bacteriocin	184	98.54%	91.67%	98.77%	70.97%	66.67%	80.00%	0.0589	98.54%
Beta-defensin	184	93.89%	68.29%	95.37%	45.90%	37.84%	54.90%	0.1646	94.56%
Bombinin	184	96.81%	70.97%	97.92%	59.46%	47.83%	64.71%	0.1248	96.81%
Cathelicidin	184	87.12%	37.04%	88.98%	11.11%	9.35%	17.09%	0.1983	96.41%
Cecropin	184	91.37%	100.00%	91.01%	31.58%	31.58%	48.00%	0.1135	96.02%
Cyclotide (Bracelet subfamily)	184	94.42%	100.00%	94.33%	22.22%	22.22%	36.36%	0.0548	98.41%
DEFL	184	92.70%	54.05%	94.69%	34.48%	26.67%	42.11%	0.1928	95.09%
FSAP (Brevinin subfamily)	184	90.17%	71.33%	94.59%	75.56%	57.95%	73.38%	0.3618	90.17%
FSAP (Caerin subfamily)	184	87.92%	72.73%	88.14%	8.33%	8.08%	14.95%	0.0809	98.54%
FSAP (Dermaseptin subfamily)	184	93.23%	86.67%	93.50%	35.62%	33.77%	50.49%	0.1267	96.02%
Invertebrate defensin (Type 1 subfamily)	184	97.21%	85.71%	97.54%	50.00%	46.15%	63.16%	0.075	97.21%
Invertebrate defensin (Type 2 subfamily)	184	98.54%	53.85%	99.32%	58.33%	38.89%	56.00%	0.0656	98.54%
Type A lantibiotic	184	91.10%	45.45%	91.78%	7.58%	6.94%	12.99%	0.0907	98.54%
Average		93.54%	73.08%	94.50%	40.65%	34.76%	48.80%	12.98%	96.53%

Additionally, we demonstrated that using both parts of the feature vector in the clustering (i.e. the compositional features along with the restrictive features), but without feature subset selection, did not distinguish well peptides of any AMP family, as shown in (Table 3.5). The average obtained values of sensitivity, specificity and precision for all 14 target AMP families were 73.55%, 91.74% and 31.12%, respectively. We notice again the significant drop in sensitivity and precision as compared when using the selected restrictive physicochemical and compositional properties (Table 3.3). These findings indicated the value of feature subset selection to determine family-specific amino acid composition and physicochemical properties that characterize the peptides of the family and discriminate the AMP family from other AMPs.

Table 3.5. Performance of k-means clustering using all properties. The k-means clustering performance of 14 target AMP families using the entire set of features to represent each family.

Target AMP family	Number of features	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
Alpha-defensin	299	95.62%	85.29%	96.11%	50.88%	46.77%	63.74%	0.1182	95.62%
Bacteriocin	225	98.54%	87.50%	98.90%	72.41%	65.62%	79.25%	0.0659	98.54%
Beta-defensin	261	90.44%	85.37%	90.73%	34.65%	32.71%	49.30%	0.1752	94.56%
Bombinin	1095	89.77%	83.87%	90.03%	26.53%	25.24%	40.31%	0.1486	95.88%
Cathelicidin	521	88.31%	33.33%	90.36%	11.39%	9.28%	16.98%	0.1949	96.41%
Cecropin	835	93.23%	80.00%	93.78%	34.78%	32.00%	48.48%	0.1457	96.02%
Cyclotide (Bracelet subfamily)	350	94.02%	100.00%	93.93%	21.05%	21.05%	34.78%	0.0562	98.41%
DEFL	252	92.43%	56.76%	94.27%	33.87%	26.92%	42.42%	0.1699	95.09%
FSAP (Brevinin subfamily)	1945	81.54%	43.36%	90.49%	51.67%	30.85%	47.15%	0.4433	83.53%

FSAP (Caerin subfamily)	1946	91.63%	63.64%	92.05%	10.61%	10.00%	18.18%	0.0786	98.54%
FSAP (Dermaseptin subfamily)	327	89.51%	76.67%	90.04%	24.21%	22.55%	36.80%	0.1584	96.02%
Invertebrate defensin (Type 1 subfamily)	402	95.09%	90.48%	95.22%	35.19%	33.93%	50.67%	0.0872	97.21%
Invertebrate defensin (Type 2 subfamily)	510	96.02%	61.54%	96.62%	24.24%	21.05%	34.78%	0.0775	98.27%
Type A lantibiotic	194	71.98%	81.82%	71.83%	4.13%	4.09%	7.86%	0.0944	98.54%
Average		90.58%	73.55%	91.74%	31.12%	27.29%	40.76%	14.39%	95.90%

Our representation method of peptide sequences is characterized by the identification of restrictive physicochemical properties in different regions of the peptide sequence. We compared this method of representation with another simple representation method that uses all physicochemical and compositional properties in different regions within the peptide sequence. Consequently, the entire set of 294 amino acids features in each of the 6 peptide regions (n1,n2,n3,n4,M and C) were used. This resulted into a total of $1948=184 + (294 \times 6)$ features without performing any additional feature selection. The clustering results in (Table 3.6) provided average sensitivity, specificity and precision for all 14 target AMP families of 63.91%, 93.54% and 32.51%, respectively. We again noticed significance drop in sensitivity and precision as compared with the values obtained when features were selected by our method (Table 3.3). This confirms the importance of identifying the restrictive physicochemical properties in different regions of the peptide sequence followed by feature subset selection to select subset of restrictive properties. Since huge percentage of properties do not contribute to distinguishing the AMP families, if we use these properties they will affect the distance

measure. The precision and the sensitivity obtained by four different methods used to represent the AMPs are compared in (Figure 3.3 and Figure 3.4).

Table 3.6. K-means clustering performance using all physicochemical properties. The k-means clustering performance of 14 target AMP families using all the 294 physicochemical properties in each of the 6 regions (n1,n2,n3,n4,M and C).

Target AMP family	Number of features	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
Alpha-defensin	1948	94.82%	67.65%	96.11%	45.10%	37.10%	54.12%	0.157	95.48%
Bacteriocin	1948	96.55%	70.83%	97.39%	47.22%	39.53%	56.67%	0.116	96.81%
Beta-defensin	1948	87.92%	36.59%	90.87%	18.75%	14.15%	24.79%	0.241	94.56%
Bombinin	1948	88.71%	80.65%	89.06%	24.04%	22.73%	37.04%	0.1598	95.88%
Cathelicidin	1948	96.02%	33.33%	98.35%	42.86%	23.08%	37.50%	0.1611	96.41%
Cecropin	1948	93.49%	66.67%	94.61%	33.90%	28.99%	44.94%	0.1487	96.02%
Cyclotide (Bracelet subfamily)	1948	99.07%	100.00%	99.06%	63.16%	63.16%	77.42%	0.024	99.07%
DEFL	1948	90.04%	54.05%	91.90%	25.64%	21.05%	34.78%	0.1996	95.09%
FSAP (Brevinin subfamily)	1948	81.54%	43.36%	90.49%	51.67%	30.85%	47.15%	0.4429	83.53%
FSAP (Caerin subfamily)	1948	91.63%	63.64%	92.05%	10.61%	10.00%	18.18%	0.0786	98.54%
FSAP (Dermaseptin subfamily)	1948	95.22%	80.00%	95.85%	44.44%	40.00%	57.14%	0.1234	96.02%
Invertebrate defensin (Type 1 subfamily)	1948	91.24%	100.00%	90.98%	24.14%	24.14%	38.89%	0.0921	97.21%
Invertebrate defensin (Type 2 subfamily)	1948	85.13%	61.54%	85.54%	6.96%	6.67%	12.50%	0.0925	98.27%
Type A lantibiotic	1948	96.41%	36.36%	97.30%	16.67%	12.90%	22.86%	0.0876	98.54%
Average		91.99%	63.91%	93.54%	32.51%	26.74%	40.28%	15.17%	95.82%

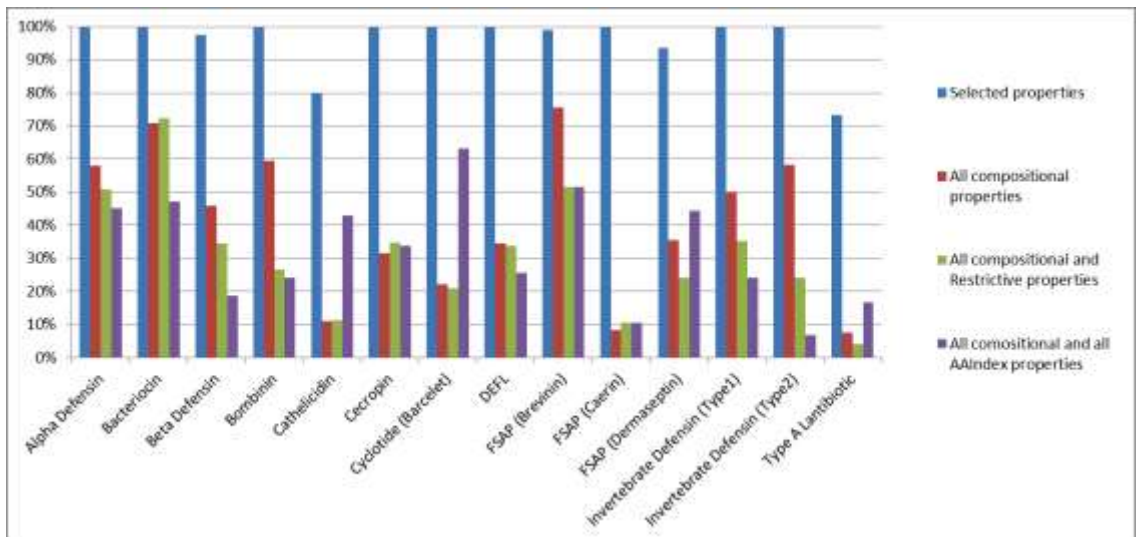


Figure 3.3. Comparison between AMP representation methods (precision). Bar plots of the precision of four AMP representation methods.

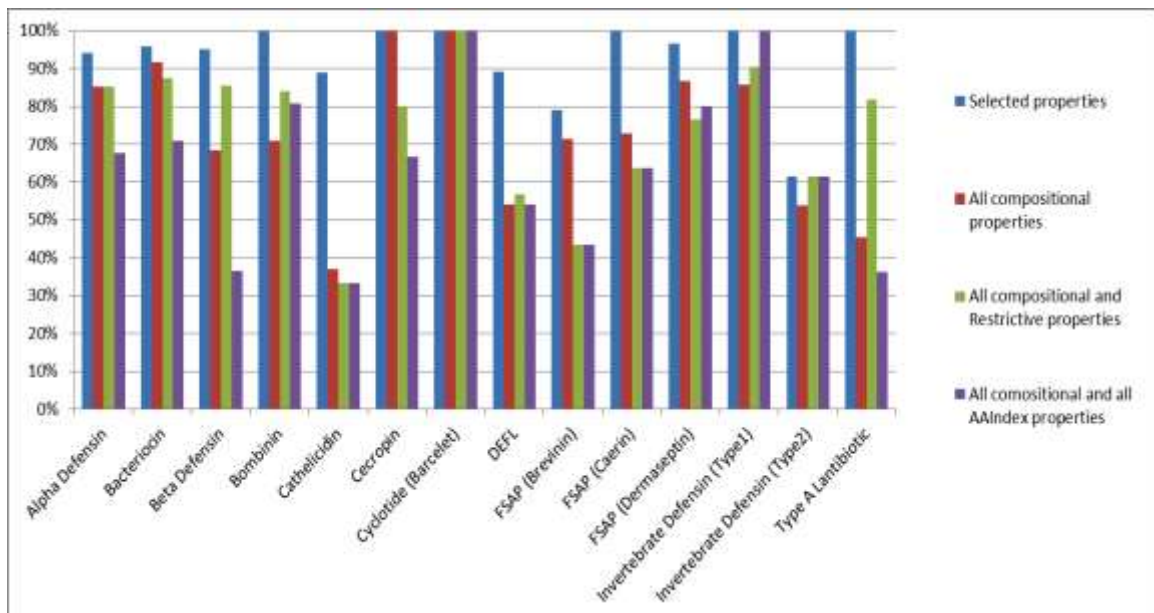


Figure 3.4. Comparison between AMP representation methods (sensitivity). Bar plots of the sensitivity of four AMP representation methods.

3.4.3 Comparison between Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Differential Evolution (DE) optimization algorithms

We compared the clustering performance using the features selected by three different global optimization algorithms, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Differential Evolution (DE). We found the following:

1. The majority of the features selected by GA were among features selected by PSO and DE.
2. While we tested many different parameter values for both of PSO and DE, and selected the best results for both of them, GA gives the minimum set of features that yielded the highest accuracy among the three methods.

The results of PSO and DE are provided below in Tables 3.7 and 3.8, respectively, while the results of GA are provided in Table 3.3 in the previous section.

Table 3.7. K-means clustering performance using PSO algorithm. The performance of the k-means clustering of 14 target AMP families using features selected by the Particle Swarm Optimization (PSO) algorithm.

Target AMP family	Number of features	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
Alpha-defensin	101	99.34%	91.18%	99.72%	93.94%	86.11%	92.54%	0.0442	99.34%
Bacteriocin	60	99.73%	91.67%	100.00%	100.00%	91.67%	95.65%	0.0226	99.73%
Beta-defensin	92	98.80%	85.37%	99.58%	92.11%	79.55%	88.61%	0.0676	98.80%
Bombinin	134	99.61%	95.12%	99.86%	97.50%	92.86%	96.30%	0.0271	99.61%
Cathelicidin	204	97.21%	59.26%	98.62%	61.54%	43.24%	60.38%	0.1318	97.21%

Cecropin	143	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
Cyclotide (Bracelet subfamily)	113	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
DEFL	90	98.80%	89.19%	99.30%	86.84%	78.57%	88.00%	0.0596	98.80%
FSAP (Brevinin subfamily)	485	88.56%	53.15%	96.65%	78.35%	46.34%	63.33%	0.3161	91.81%
FSAP (Caerin subfamily)	876	98.94%	81.82%	99.19%	60.00%	52.94%	69.23%	0.036	98.94%
FSAP (Dermaseptin subfamily)	121	99.07%	90.00%	99.45%	87.10%	79.41%	88.52%	0.0498	99.07%
Invertebrate defensin (Type 1 subfamily)	168	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
Invertebrate defensin (Type 2 subfamily)	209	98.94%	61.54%	99.59%	72.73%	50.00%	66.67%	0.0548	98.94%
Type A lantibiotic	74	98.94%	90.91%	99.06%	58.82%	55.56%	71.43%	0.0299	98.94%
Average		98.42%	84.94%	99.36%	84.92%	75.45%	84.33%	6.00%	98.66%

Table 3.8. K-means clustering performance using DE algorithm. The performance of the k-means clustering of 14 target AMP families using features selected by the Differential Evolution (DE) optimization algorithm.

Target AMP family	Number of features	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
Alpha-defensin	73	99.20%	88.24%	99.72%	93.75%	83.33%	90.91%	0.0474	99.20%
Bacteriocin	57	99.20%	91.67%	99.45%	84.62%	78.57%	88.00%	0.0443	99.20%
Beta-defensin	85	98.01%	80.49%	99.02%	82.50%	68.75%	81.48%	0.0991	98.01%
Bombinin	246	99.87%	96.77%	100.00%	100.00%	96.77%	98.36%	0.0097	99.87%
Cathelicidin	129	97.08%	55.56%	98.62%	60.00%	40.54%	57.69%	0.135	97.08%
Cecropin	211	98.94%	86.67%	99.45%	86.67%	76.47%	86.67%	0.0581	98.94%
Cyclotide (Bracelet subfamily)	126	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
DEFL	72	97.88%	72.97%	99.16%	81.82%	62.79%	77.14%	0.1123	97.88%
FSAP (Brevinin subfamily)	299	93.63%	72.73%	98.40%	91.23%	67.97%	80.93%	0.26	94.54%
FSAP (Caerin subfamily)	392	99.74%	81.82%	100.00%	100.00%	81.82%	90.00%	0.0165	99.74%
FSAP (Dermaseptin subfamily)	73	97.61%	86.67%	98.06%	65.00%	59.09%	74.29%	0.0835	97.61%
Invertebrate defensin (Type 1 subfamily)	126	99.07%	95.24%	99.18%	76.92%	74.07%	85.11%	0.0359	99.07%
Invertebrate defensin (Type 2 subfamily)	133	99.07%	61.54%	99.73%	80.00%	53.33%	69.57%	0.0478	99.07%
Type A lantibiotic	52	95.88%	81.82%	96.09%	23.68%	22.50%	36.73%	0.0612	98.54%

Average	98.23%	82.30%	99.06%	80.44%	69.00%	79.78%	7.22%	98.48%
----------------	--------	--------	--------	--------	--------	--------	-------	--------

3.4.4 Comparison of k-means clustering with Affinity Propagation (AP) clustering Algorithm

We compared the performance of clustering 14 target AMP families using Affinity Propagation (AP) and k-means algorithms. The clustering results in Table 3.9 for AP as compared to the clustering results of k-means algorithm in Table 3.3 shows the k-means clustering provided higher accuracy than AP for most of the AMP families.

Table 3.9. Affinity Propagation clustering performance. The performance of the Affinity Propagation (AP) clustering of 14 target AMP families using features selected by the Genetic Algorithm.

Target AMP family	Number of features	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
Alpha-defensin	68	98.67%	79.41%	99.58%	90.00%	72.97%	84.38%	0.0719	98.67%
Bacteriocin	70	99.20%	83.33%	99.73%	90.91%	76.92%	86.96%	0.0583	99.20%
Beta-defensin	62	96.81%	85.37%	97.47%	66.04%	59.32%	74.47%	0.1165	96.81%
Bombinin	113	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
Cathelicidin	71	98.54%	70.37%	99.59%	86.36%	63.33%	77.55%	0.0891	98.54%
Cecropin	190	98.94%	93.33%	99.17%	82.35%	77.78%	87.50%	0.0469	98.94%
Cyclotide (Bracelet subfamily)	93	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
DEFL	51	97.34%	67.57%	98.88%	75.76%	55.56%	71.43%	0.1484	97.34%
FSAP (Brevinin subfamily)	254	96.15%	83.92%	99.02%	95.24%	80.54%	89.22%	0.2278	96.15%
FSAP (Caerin subfamily)	241	98.67%	81.82%	98.92%	52.94%	47.37%	64.29%	0.0423	98.67%

FSAP (Dermaseptin subfamily)	95	98.54%	83.33%	99.17%	80.65%	69.44%	81.97%	0.0739	98.54%
Invertebrate defensin (Type 1 subfamily)	71	99.60%	90.48%	99.86%	95.00%	86.36%	92.68%	0.0276	99.60%
Invertebrate defensin (Type 2 subfamily)	78	99.47%	69.23%	100.00%	100.00%	69.23%	81.82%	0.0269	99.47%
Type A lantibiotic	48	99.20%	54.55%	99.87%	85.71%	50.00%	66.67%	0.0452	99.20%
Average		98.65%	81.62%	99.38%	85.78%	72.06%	82.78%	6.96%	98.65%

3.4.5 Comparison of clustering using different distance measures

We performed k-means clustering of 14 target AMP families using three different distance measures (correlation, cosine and city block) as shown in Tables 3.10-3.12. The comparison of these results with the clustering results using Euclidean distance (Table 3.3) shows that the clustering results using these measures are comparable to those obtained by the Euclidean distance, but better results with minimum number of features were obtained with the Euclidean distance.

Table 3.10. K-means clustering performance using city block distance. The performance of the k-means clustering (using city block distance measure) of 14 target AMP families using features selected by the Genetic Algorithm.

Target AMP family	Number of features	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
Alpha-defensin	13	99.73%	94.12%	100.00%	100.00%	94.12%	96.97%	0.0209	99.73%
Bacteriocin	11	99.87%	95.83%	100.00%	100.00%	95.83%	97.87%	0.0102	99.87%
Beta-defensin	28	99.60%	92.68%	100.00%	100.00%	92.68%	96.20%	0.0251	99.60%
Bombinin	17	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
Cathelicidin	39	99.47%	88.89%	99.86%	96.00%	85.71%	92.31%	0.037	99.47%

Cecropin	42	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
Cyclotide (Bracelet subfamily)	4	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
DEFL	17	99.34%	86.49%	100.00%	100.00%	86.49%	92.75%	0.0485	99.34%
FSAP (Brevinin subfamily)	110	97.92%	90.21%	99.68%	98.47%	88.97%	94.16%	0.1268	97.92%
FSAP (Caerin subfamily)	484	99.60%	81.82%	99.87%	90.00%	75.00%	85.71%	0.0228	99.60%
FSAP (Dermaseptin subfamily)	29	99.07%	90.00%	99.45%	87.10%	79.41%	88.52%	0.0524	99.07%
Invertebrate defensin (Type 1 subfamily)	9	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
Invertebrate defensin (Type 2 subfamily)	8	99.34%	61.54%	100.00%	100.00%	61.54%	76.19%	0.0455	99.34%
Type A lantibiotic	16	99.73%	90.91%	99.87%	90.91%	83.33%	90.91%	0.0156	99.73%
Average		99.55%	90.89%	99.91%	97.32%	88.79%	93.69%	2.89%	99.55%

Table 3.11. K-means clustering performance using cosine distance. The performance of the k-means clustering (using cosine distance measure) of 14 target AMP families using features selected by the Genetic Algorithm.

Target AMP family	Number of features	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
Alpha-defensin	23	99.73%	94.12%	100.00%	100.00%	94.12%	96.97%	0.0187	99.73%
Bacteriocin	23	99.87%	95.83%	100.00%	100.00%	95.83%	97.87%	0.0096	99.87%
Beta-defensin	43	98.94%	85.37%	99.72%	94.59%	81.40%	89.74%	0.062	98.94%
Bombinin	33	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
Cathelicidin	51	98.27%	77.78%	99.04%	75.00%	61.76%	76.36%	0.0899	98.27%
Cecropin	55	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
Cyclotide (Bracelet subfamily)	17	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
DEFL	29	99.47%	89.19%	100.00%	100.00%	89.19%	94.29%	0.038	99.47%
FSAP (Brevinin subfamily)	261	92.43%	70.63%	97.54%	87.07%	63.92%	77.99%	0.3356	92.43%
FSAP (Caerin subfamily)	199	99.60%	81.82%	99.87%	90.00%	75.00%	85.71%	0.022	99.60%
FSAP (Dermaseptin subfamily)	36	99.20%	96.67%	99.31%	85.29%	82.86%	90.62%	0.0367	99.20%
Invertebrate defensin (Type 1 subfamily)	17	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
Invertebrate defensin (Type 2 subfamily)	44	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
Type A lantibiotic	37	99.60%	100.00%	99.60%	78.57%	78.57%	88.00%	0.0139	99.60%

Average	99.08%	92.24%	99.65%	93.61%	87.33%	92.68%	4.47%	99.08%
----------------	--------	--------	--------	--------	--------	--------	-------	--------

Table 3.12. K-means clustering performance using correlation distance. The performance of the k-means clustering (using correlation distance measure) of 14 target AMP families using features selected by the Genetic Algorithm.

Target AMP family	Number of features	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
Alpha-defensin	24	99.73%	94.12%	100.00%	100.00%	94.12%	96.97%	0.0195	99.73%
Bacteriocin	26	99.87%	95.83%	100.00%	100.00%	95.83%	97.87%	0.0095	99.87%
Beta-defensin	41	99.60%	92.68%	100.00%	100.00%	92.68%	96.20%	0.0285	99.60%
Bombinin	45	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
Cathelicidin	45	98.67%	88.89%	99.04%	77.42%	70.59%	82.76%	0.0603	98.67%
Cecropin	36	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
Cyclotide (Bracelet subfamily)	30	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
DEFL	38	99.60%	94.59%	99.86%	97.22%	92.11%	95.89%	0.0291	99.60%
FSAP (Brevinin subfamily)	370	88.84%	45.45%	99.02%	91.55%	43.62%	60.75%	0.3225	89.91%
FSAP (Caerin subfamily)	112	99.87%	90.91%	100.00%	100.00%	90.91%	95.24%	0.0094	99.87%
FSAP (Dermaseptin subfamily)	41	99.73%	96.67%	99.86%	96.67%	93.55%	96.67%	0.0183	99.73%
Invertebrate defensin (Type 1 subfamily)	30	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
Invertebrate defensin (Type 2 subfamily)	35	99.60%	92.31%	99.73%	85.71%	80.00%	88.89%	0.0205	99.60%
Type A lantibiotic	35	99.34%	90.91%	99.46%	71.43%	66.67%	80.00%	0.0255	99.34%
Average		98.92%	91.60%	99.78%	94.29%	87.15%	92.23%	3.88%	98.99%

3.4.6 Testing the Selected Properties on Non-AMPs and other AMP Databases

We checked if the features determined for AMP families would separate AMPs from non-AMPs. If this is not possible, then the non-AMPs would have similar feature profiles as some classes of AMPs. We found that this distinguishing of AMPs from non-AMPs was

possible with an average accuracy, sensitivity, specificity and cluster purity of 96.72%, 67.62%, 96.85% and 99.56%, respectively, confirming that features selected for AMP families are highly specific to AMPs. One should note that our selection of features was not made with the aim to distinguish AMP families from non-AMPs. Rather, they were selected to distinguish different AMP families from other AMPs.

Furthermore, we evaluated another database CAMP (Waghu, et al., 2014). We considered only experimentally validated mature AMP peptide sequences with UNIPROT IDs from CAMP. We tested if the selected features based on DAMPD database entries would separate AMP families from CAMP and found that this is possible with an average accuracy, sensitivity, specificity and cluster purity of 94.03%, 76.91%, 94.58% and 97.96%, respectively. This suggests that the features selected based on DAMPD entries discriminate well between different AMP families in the CAMP database. It should be mentioned, however, that the criteria for inclusion of AMPs into CAMP and into DAMPD are not necessarily the same, so very strict comparison is not possible.

3.4.7 Selected Properties that Discriminate AMP Families

We identified, in total, 394 properties to discriminate each family of the 14 target AMP families from all other AMPs (Table 3.2, column 6). The whole set of these properties is provided in (Appendix 3). Different amounts of properties were recognized as being characteristic for different AMP families. For instance, seven properties can distinguish cyclotides (bracelet subfamily) from all other AMPs with 100% accuracy. Though, additional number of properties were required to discriminate certain AMP families

from all other AMPs, especially sub-families belong to the same super family. E.g., s brevinin, caerin and dermaseptin sub-families of the frog skin active peptide (FSAP) family needed 118, 28 and 25 properties, respectively.

The selected properties within the peptide sequence are region-specific. Specifically, the properties are determined from the N terminal, M region, C terminal or the entire AMP sequence. Part of the selected properties are associated to compositional features of amino acid sequence, while other part of the selected properties are associated to physicochemical properties. The selected properties distribution among these two groups of features is illustrated in (Table 3.2, columns 8-9). Amino acid composition properties can distinguish highly separable AMP family from other AMPs, such as type A lantibiotic family. Though, other 13 AMP families need a mixture of amino acid composition and physicochemical properties to discern their peptides from other AMPs.

Certain selected properties show the enrichment of a particular amino acid(s) or physicochemical property in a specific region, these properties are so-called “enriched”. Conversely, some other properties indicate the importance of the depletion or absence of a specific amino acid(s) or physicochemical property in a specific region, which we name these properties as “depleted”. For instance, in alpha defensins, arginine is enriched and lysine is depleted in the C terminal, and these two composition properties were identified as significant properties to characterize this family.

3.5 Discussion

We developed here a novel computational model for identification of AMP characteristics that discern peptides of AMP families using compositional and physicochemical properties. The identified properties can serve as potential design guides for synthetic AMPs development. In addition, these properties can be utilized for the development of a classification model to predict the class (category) of a new candidate AMP. In the following text, enrichment and depletion are considered related to one target AMP family relative to all other AMP families. We comment here on the selected properties for all of the 14 AMP families.

Alpha-defensins. We found enrichment of cysteine (C) in the n3 and n4 regions, arginine (R) in the M region, and arginine and tyrosine (Y) in the C region, while glycine (G) was found depleted in the n2 region as well as lysine (K) in the C region. Numerous studies have demonstrated that bactericidal activity is independent of highly conserved features, such as invariant disulfide array, Arg-Glu salt bridge, or Gly residue at CysIII+8 I, with exception to the high arginine content relative to lysine (Lehrer, 2007; Rajabi, et al., 2008; Schmidt, et al., 2012), which complies with our finding. Moreover, Schmidt *et al.* (2012) demonstrated that the replacement of arginine with lysine decreases the activity of these peptides (Schmidt, et al., 2012). Also, AMPs disrupt membranes through a combination of electrostatic interactions between cationic amino acid side chains and electronegative components of the microbial cell envelope, followed by the insertion of hydrophobic patches into the nonpolar interior of the membrane bilayer (Brogden,

2005). The mouse alpha-defensin cryptdin-4 (Crp4) was demonstrated to induce bactericidal activity via this mechanism (Satchell, et al., 2003). NMR structure of Crp4 demonstrated its cationic amino acids to be arginine, lysine and histidine (H) and its hydrophobic patches to include isoleucine (I), leucine (L), valine (V), phenylalanine (F) and tyrosine. Similar to these findings, we found for alpha-defensins the enrichment of cysteine in the N region, which, taking into account their hydropathy index, may suggest that it is a key component of the hydrophobic patch. The enrichment of arginine and tyrosine likely adds to the electrostatic interaction that contributes to membrane disruption.

Bacteriocins. We found an enrichment of glycine, valine, asparagine and tyrosine in the n2, n2, n3 and n3 regions, respectively. Alanine (A) was found depleted in the N region. If we only consider the amino acids identified in our study as enriched, we pinpoint valine, glycine and tyrosine as key hydrophobic residues for bacteriocins. Findings reported by (Jabrane, et al., 2002) partially support our results as they demonstrated that leucine and asparagine are enriched in bacteriocin serracin P. Oppegard et al. additionally demonstrated by mutation analysis that substituting tyrosine with glycine or leucine significantly decreases bactericidal activity (Oppegard, et al., 2008).

Beta-defensins. We found arginine, valine, phenylalanine, asparagine (N), cysteine and glycine enriched in the n1, n2, n3, n4, n4, and n4 regions, respectively. Also, for the properties extracted from the n4 sub-region an enrichment for the parameter of charge transfer donor capacity was found (Charton and Charton, 1983). The M region exhibited

enrichment, of cysteine, while the C region had enrichment of arginine and cysteine. It was proposed that the beta-defensin N-terminal helix with many hydrophobic residues is inserted inside the micelle, while the C-terminal helix with one large positive charge patch is located outside the micelle and interacts with the charged head groups of the micelle (Chandrababu, et al., 2009). Of the array of amino acids, we identified valine, phenylalanine and cysteine as key hydrophobic residues enriched in the N-terminal that likely facilitate the insertion of the beta-defensins N-terminal helix into the micelles, whilst for the C-terminal helix arginine is the key amino acid forming the positive charge patch. Finally, if we only consider amino acids that our study identified as enriched, we observe cysteine glycine and arginine most enriched. Our findings are partly corroborated by the results of (Midorikawa, et al., 2003) and (Chandrababu, et al., 2009) as they demonstrated that the beta-defensins are characterized by the enrichment of cysteine. That is the existence of six conserved cysteine residues (Midorikawa, et al., 2003) and that the arrangement of cysteine residues in the three-dimensional space is important to the antimicrobial selectivity and salt-dependent activity by mutating all six cysteine residues of human beta-defensin-3 (HBD-3) (Chandrababu, et al., 2009).

Bombinins. We only identified a few properties related to composition and distance such as the enrichment of glycine and threonine (T) in the n1 and n4 regions, respectively, and of glycine and leucine in the C region. Our findings are in part supported by bombinins being characterized as glycine-rich and Zangger et al.

demonstrated via NMR the presence of a glycine ridge that is believed to provide unique functionality (Zangger, et al., 2008).

Cathelicidins. We identified the enrichment of arginine in the n1, and n2 regions, respectively. Lysine was identified depleted in the C region, while proline (P) was enriched in the C region. Also, for the properties extracted from the n3 sub-region, the enrichment for the linker propensity from 1-linker and normalized frequency of beta-sheet, were found (George and Heringa, 2002; Levitt, 1978). Our findings are supported by cathelicidins being characterized as proline-rich and having a highly conserved N-terminal preprosequences followed by variable C-terminal sequences that are biologically active effectors (Chan, et al., 2001; Zanetti, et al., 1995). Moreover, the proline was proved to sustain the antimicrobial activity of mammalian cathelicidins by resisting serine proteases cleavage of the scissile bond (Shinnar, et al., 2003). Our finding are further supported by cathelicidins from hagfish exhibiting four arginines positioned between the cathelin domain and the antimicrobial sequences (Uzzell, et al., 2002). The arginine tetrads of these latent zymogens are believed to be specifically processed by prohormone convertases such as furin proteases in specific cells as an activity switch (Rockwell, et al., 2002; Steiner, 1998).

Cecropins. We identified enrichment of lysine and glutamic acid (E) in the n3 and n4 regions, respectively, while alanine was enriched in the C region. Andreu *et al.* produced synthetic cecropin A that induces comparable antibacterial activity and is indistinguishable by chemical and physical criteria from the naturally occurring cecropin

A (Andreu, et al., 1983). In partial corroboration with our findings, it has been demonstrated that cecropin analogues with an impaired N-terminal helix, such as cecropin A-(3-37) with removed lysine and tryptophan has reduced membrane disrupting abilities that correlate with their lower antibacterial activity that was rationalized in terms of reduced binding to bacteria (Andreu, et al., 1983; Steiner, et al., 1988). Similarly, Fink *et al.* demonstrated via a chemically synthesized cecropin D analog (9-37) that no activity is observed without phenylalanine and glutamic acid in the N-terminal (Fink, et al., 1989). Moreover, Lee *et al.* demonstrated that lysine, glutamic acid, and arginine are conserved in cecropins and that alanine is enriched (Lee, et al., 1989).

Cyclotides. We only identified a few properties related to composition, such as enrichment of glutamic acid and glycine in the n4 and C regions, respectively. For the properties extracted from the C region enrichment for normalized frequency of turn in alpha+beta class was found (Palau, et al., 1982). Work by Hermann *et al.* partially supports these finding as they demonstrated that methylation of charged glutamic acid residue of cyclotide cycloviolacin O2 decreased its potency 48-fold. They additionally showed conserved cysteine residues and demonstrated that acetylation of the two lysine residues also reduced the potency 3-fold (Herrmann, et al., 2006). Koehbach *et al.* elucidated the structure of kalata B7 to determine its associated ligand–receptor interaction. They inferred an interaction with the oxytocin receptor owing to loop 3 of kalata B7 (-CYTQGC-) being homologous to the six-residue ring sequence of oxytocin.

They further exhibited the crucial role of the tyrosine and glutamine residues (loop 3) by generating mutated variants (Y replaced by A, S, or F; Q was replaced by A or E), all of which were inactive or did not bind to the receptor (Koehbach, et al., 2013). Moreover, Rosengren *et al.* demonstrated that most cyclotides have a glycine positioned before the cysteine residue to form the ϕ angle required for the type II β -turn needed for cyclization (Rosengren, et al., 2003).

Defensin-like (DEFL) peptides. We found glutamic acid, serine, threonine, cysteine, cysteine, glutamic acid, cysteine, glycine and serine to be enriched in the n1, n1, n1, n2, n3, n3, n4, n4, n4 regions, respectively. Cysteine was enriched in the C region. We also found for the properties extracted from the n4 sub-region such as conformational parameter of inner helix (Beghin and Dirckx, 1975) and ratio of average and computed composition (Nakashima, et al., 1990) to be depleted. If we only consider the amino acids that our study identified as enriched we observe the N-terminal is negatively charged (glutamic acid) and enriched with cysteine and serine. Our findings are in part supported by Correa and Oguiura findings in which they produced a phylogenetic analysis of beta-defensin-like genes of Bothrops, Crotalus and Lachesis snakes and observed that these have conserved cysteine residues (Correa and Oguiura, 2013).

Frog skin active peptides. We additionally discriminated sub-families, such as the brevinin, caerin and dermaseptin, belonging to the frog skin active peptide (FSAP) family. For the brevinin sub-family, we found enrichment of phenylalanine in the n1 region, and cysteine and isoleucine in the C region. Glutamic acid and serine were

identified depleted in the n1 and n4 sub-regions, respectively. We also found the enrichment of physicochemical properties such as normalized positional residue frequency at helix termini N (n1 region), weights for alpha-helix at the window position of -6 (n2 region), normalized composition from fungi and plant (n2 region), normalized composition from mt-proteins (n3 region), pK(-COOH) (n3 region), normalized composition of membrane proteins (n4 region), weights for alpha-helix at the window position 1 (M region). Work by Pal *et al.* partially corroborates these findings as they demonstrated that replacement of the cysteine residues with serine in brevinin-1BYa, a cationic alpha-helical peptide present in skin secretions of the foothill yellow-legged frog *Rana boylei*, reduced its haemolytic activity and activities against Gram-negative bacteria and yeast species. However, high potency against Gram-positive bacteria was retained (Pal, et al., 2006). Conlon *et al.* (2009) demonstrated via structure-activity relationship of the brevinin family peptides, that brevinin-1BLc is more potent than brevinin-1Ya and -1Yc and that the appreciably lower antimicrobial potencies of brevinin-1Ya and -1Yc correlates with decreased cationicity produced by the amino acid substitutions Lys(11)-->Asn (brevinin-1Ya) and Pro(14)-->Glu (brevinin-1Yc) (Conlon, et al., 2009).

For the Caerin sub-family we identified fewer properties related to composition and distance of amino acid residues, such as the enrichment of valine in the C region, and enrichment of physicochemical properties such as normalized hydrophobicity scales for alpha-proteins (n1 region), weights for alpha-helix at the window position of 5 (n1 region), van der Waals parameter R0 (n1 region), surrounding hydrophobicity in turn (n1

region), Optical rotation (C region) and Normalized positional residue frequency at helix termini N" (C region). Our findings are supported in part as valine (Wong, et al., 1997) was shown to play a role in the activity of caerins.

Lastly, dermaseptin sub-family showed enrichment of tryptophan in n1 and n4 regions. Alanine is enriched in a negatively charged M region, as well as glutamine and leucine in the C region. Properties that were depleted in this family include phenylalanine in the n1 region. We found enrichment of physicochemical properties such as weights for alpha-helix at the window position of -6 (n4 region), average interactions per side chain atom (M region) and AA composition of EXT of multi-spanning proteins (C region). Our finds are partially corroborated by Moll *et al.* demonstrating that tryptophan is important for this peptides ability to penetrate membranes (Moll, et al., 2000). Lequin *et al.* further showed via comparison of dermaseptin B2 and S9 structures its common cationic amino acids to be lysine and glutamic acid, and key hydrophobic residues to be isoleucine, leucine and valine (Lequin, et al., 2006). Moreover, Cao *et al.* compared the antimicrobial activities of recombinant adenoregulin with C-amidated terminus to that without an amidated C-terminus and demonstrated that the amide of glutamine at C-terminus increased its potency against microorganisms such as *Tritirachium album* and *Saccharomyces cerevisiae* (Cao, et al., 2005).

Type A lantibiotics. We found valine and threonine enriched in the n2 and n4 regions, respectively, methionine in the M region and asparagine, cysteine, phenylalanine, serine and threonine enriched in the C region. Cysteine and methionine are also enriched in

the entire sequence. Properties that appear depleted in Type A lantibiotics include alanine in the N region and arginine in the entire sequence as well as the frequency of (>1 aa and ≤6 aa) Distances of basic amino acids (RHK) in N region. Work by Sloatweg *et al.* demonstrated that C-terminal modification of nisin does not deteriorate biological activity in sharp contrast to N-terminal modification (Sloatweg, et al., 2013). Since lantibiotics are a class of more extensively modified bacteriocins, characterized by the presence of lanthionine (Lan) and methyllanthionine (MeLan) 'hinge' regions that originate from cysteine and serine (Dischinger, et al., 2014; Lohans and Vederas, 2014), our findings of enriched serine and threonine in this family compared to others are partially supported. Moreover, Chen *et al.* demonstrated using mutacin II that these hinge regions are essential for biological activity and biosynthesis or export of the peptide (Chen, et al., 1998).

Invertebrate defensins. We additionally discriminated sub-families (Type 1 and Type 2) of invertebrate defensins. For the invertebrate defensin Type 1 sub-family we found enrichment of aspartic acid, alanine, cysteine, leucine, cysteine and valine in the n1, n2, n2, n4, n4, C and C regions, respectively. For the invertebrate defensin Type 2 sub-family (arthropods) we found histidine, cysteine, cysteine and threonine enriched in the n1, n2, n3 and C regions, respectively.

3.6 Conclusion

The rapid increase of MDR pathogens fostered more interest in design of novel AMPs using *in-silico* methods. In this work, we developed a method to profile different AMP families using compositional and physicochemical properties extracted from AMP sequences. We used GA to optimize an objective function based on unsupervised k-means clustering to select the properties that strongly discriminate each AMP family from all other AMPs. Our results suggest that the properties that we identified to profile each AMP family can be a useful guide during the *in-silico* design process for novel synthetic AMPs. The method that we developed is generic and it can be applied to profile different protein families.

Chapter 4

Novel ML Method of Prediction of Transcription Factor DNA Binding Sites

This chapter is prepared as a manuscript:

New method for transcription factor binding site prediction (Abdullah M. Khamis, Olaa Motwalli, Romina Oliva, Boris R. Jankovic, Yulia A. Medvedeva, Xin Gao, Vladimir B. Bajic), intended for submission

4.1 Summary

Identification of transcription factor (TF) binding sites (TFBSs) is important in computational inference of gene regulation. Classical widely used computational TFBSs predictions based on variants of Position Weight Matrix (PWM), frequently suffer from high false positive rates. In a typical computational study of transcription regulation in higher organisms, numerous TFBS models are used due to a large number of involved TFs. To overcome these problems we developed a novel method, DRAF, for TFBS prediction that requires only 14 prediction models for 232 human TFs and dramatically improves accuracy. DRAF combines information from TFBS sequence and physicochemical properties of the TF DNA binding domains. Computational evaluation of DRAF on 321 human CHIP-seq datasets shows 14-, 15- and 16-fold reduction of false positives relative to the models from HOCOMOCO, TRANSFAC and DeepBind, respectively, suggesting that conventional PWM models for TFBS prediction can be efficiently replaced by small number of models that significantly improve prediction accuracy.

4.2 Introduction

Information on regulation of transcription forms a basis for understanding regulatory mechanisms of gene activation or repression in living organisms. Transcription factors (TFs) are a key component of gene regulatory networks. They bind promoters and other DNA regulatory regions in a sequence-specific manner and control gene expression through such interactions (Blancafort, et al., 2004). TF binding sites (TFBSs) on DNA are short sequences located in the gene regulation regions (Cawley, et al., 2004) being usually 5-20 base-pairs (bp) in length (Das and Dai, 2007). Accurate detection of TFBSs is frequently a central step in computational reconstruction of gene regulatory networks.

Both computational and experimental methods have been used for TFBS detection. For experimental approaches, there are numerous *in vivo* and *in vitro* high-throughput methods that have been developed (Geertz and Maerkl, 2010; Weirauch, et al., 2013), such as, for example, DNA microarray (Bulyk, et al., 1999) and microfluidic technologies (Maerkl and Quake, 2007) for *in vitro* approaches, and ChIP-chip (Horak and Snyder, 2002) and the more recent ChIP-seq (Park, 2009) technologies for *in vivo* approaches. High-throughput ChIP-seq experiments from the ENCODE project (Consortium, 2012; Gerstein, et al., 2012) have investigated only about 200 human TFs in less than a hundred cell lines. Despite the progress that has been made, these numbers are far lower than the estimated number of TFs that are encoded in the human genome or that might regulate a single cell type (Vaquerizas, et al., 2009). Therefore, the need for efficient computational methods to predict TFBS remains (Hoglund and Kohlbacher,

2004). Indeed, computational approaches for identifying TFBSs have been used successfully (Bulyk, 2003; Qiu, 2003) varying from simple pattern matching methods to more complex models (Elnitski, et al., 2006). As an illustration, cursory search for articles on “*computational transcription factor binding site models*” in Web of Science of Thomson Reuters yielded over 43,000 citations on December 07, 2015. The actual number of citations is significantly higher.

Pattern matching methods attempt to predict a TFBS by screening a candidate sequence of interest with a model derived from experimentally determined binding sites for a TF. Although suggested three decades ago (Stormo, et al., 1982), position-specific weight matrix (PWM) type models still remains the most widely used models for TFBS predictions, primarily due to their simplicity. However, a PWM model has several disadvantages. First, it is very sensitive to the quality and size of the set of TFBSs DNA sequences used to derive the PWM model (Roulet, et al., 1998). Secondly, the PWM prediction models of TFBSs frequently result in a high rate of false-positive predictions (Bi, et al., 2011). Thirdly, conventional PWMs do not model dependencies between individual positions within the TFBS (Stormo, 2000). Fourthly, usually for a TF one or more models for their TFBSs are developed to capture variability among TFBS sequences and to improve model performance. This results in a large number of TFBS models in major resources. For Example, 426 TFBS models are used to represent 401 TFs in HOCOMOCO (Kulakovskiy, et al., 2013), while 1,082 TFBS models in JASPAR (Mathelier, et al., 2015) represent 1,059 TFs. In TRANSFAC, for 5,760 TFs a total of 2,170 TFBS models are used. Further, PWM models of TFBSs do not utilize any information about

the TFs that bind to them. As a consequence, these models may not discriminate between slightly different binding sites that belong to different TFs with different structures and DNA affinity properties.

Obviously, there has been a challenge to develop models that predicts TFBSs with high specificity and sensitivity. Classical TFBS PWM models have been improved to incorporate nucleotide k-mer relationships (Gershenson, et al., 2005; Mordelet, et al., 2013). Also, more flexible approaches have been implemented to develop customized models of TFBSs, such as those based on Markov Chain (Ellrott, et al., 2002), Bayesian networks (Ben-Gal, et al., 2005), undirected graphs (Reddy, et al., 2007), Hidden Markov Models (HMM) (Mathelier and Wasserman, 2013) and most recently deep learning (Alipanahi, et al., 2015). Various methods incorporated sequence-specific and structural features of DNA for prediction of TFBSs, for example, DNA shape (Zhou, et al., 2015) and local chemical and structural properties (Bauer, et al., 2010; Meysman, et al., 2011).

However, the above approaches did not use information from TFs that bind TFBSs. A lot of research was done on incorporating suitable TFs properties into models for TFBS predictions with a hope to improve models and their prediction accuracy. Some examples of such work are the use of empirical protein–DNA binding energies (Alamanova, et al., 2010; Chen, et al., 2012; Gabdoulline, et al., 2012; Kono and Sarai, 1999; Liu, et al., 2008), also based on structural knowledge (Endres, et al., 2004; Kaplan, et al., 2005).

In addition, a variety of computational approaches have been developed based on modeling TF-TFBS interactions. Qian and colleagues (Qian, et al., 2006) used gene

ontology (GO) annotations of TFs, represented in a binary vector to denote presence or absence of each GO term in the TF, and a binary representation of TFBSs (Bhasin, et al., 2005) to describe TF-TFBS pairs. This work was later extended (Qian, et al., 2007) to include GO annotations of the TF target genes (TFT) to the TF-TFBS pairs, resulting in the use of TF-TFT-TFBS triplets. This improved the accuracy of predictions. An apparent deficiency of this approach occurs when two TFs share the same GO features but have different binding sites. An associated problem is that GO annotation of TFs does not have sufficient resolution, so this additionally reduces capability to predict distinct TFBSs. Moreover, such methods are not applicable for studies of TFs that do not possess enough GO functional annotations. Another approach that includes the amino acid properties of TFs in a model was implemented in (Cai, et al., 2009), where only six physicochemical properties of amino acids were used to for feature vectors that describe a TF.

The previously mentioned deficiencies of PWM models for TFBS predictions reduce significantly utility of such models. In order to overcome these deficiencies, in this work, we developed a novel method, DRAF, for predicting TFBSs based on the physicochemical properties of the DNA binding domains of TFs and the nucleotides sequence characteristics of target TFBSs of TFs. DRAF dramatically reduces both the number of required TFBS models and the false positive rate of TFBS prediction. It required only 14 prediction models for 232 TFs. In a comprehensive evaluation we demonstrate that DRAF models, on the ChIP-seq data of 321 human cell types obtained from ENCODE (Consortium, 2012), at the same sensitivity level generate many fold higher specificity

than PWMs from HOCOMOCO and TRANSFAC (Matys, et al., 2006) databases, or the DeepBind (Alipanahi, et al., 2015) models.

4.3 Methods

4.3.1 Datasets

TF and TFBS Sequences. We used TFBS sequences from the HOCOMOCO (Kulakovskiy, et al., 2013) database version 9, where TFBSs were selected based on the PWM thresholds with P-value < 0.0005 (as explained in (Kulakovskiy, et al., 2013)). P-values were computed by the MACRO-APE (<http://autosome.ru/macroape>, (Vorontsov, et al., 2013)). Consequently, 139,085 TFBS sequences of 426 TFBS models corresponding to 401 human TFs were obtained. Due to the large number of parameters in the DRAF models as compared to PWM models, we requested that the minimum required number of TFBSs per a single TF is 15. We further discarded all TFs that did not have DNA binding domains in the Pfam database (Finn, et al., 2014). This reduced the initial set of 426 TFBS models (associated with 401 TFs) to 250 TFBS models (associated with 232 TFs) with a total of 110,399 corresponding TFBS sequences (Table 4.1). The amino acid sequences for these 232 TFs were obtained from UniProt (UniProt, 2015).

Table 4.1. TFBS model distribution according to TFBS length. Total of 250 TFBS models distributed in 14 models represent 14 distinct TFBS lengths.

TFBS Length	Number of Tfs	Total Number of TFBSs
7	18	3,521
8	13	4,927
9	37	20,818
10	28	7,338
11	29	12,645
12	27	13,211
13	33	11,672
14	21	7,534
15	12	7,721
16	8	4,389
17	9	4,245
18	7	3,946
19	5	6,719
20	3	1,713
Total	250	110,399

TF Domains: Protein domain information was obtained from the Pfam database. We used domains that are annotated as “DNA binding domain” in at least three out of total

five annotation sections used in Pfam (Pfam, Seq-info, Pdb, GO and Interpro). We restricted our study to manually curated DNA binding domains (that is Pfam-A) having the highest significance score and E-value less than 0.1. Finally, each TF was represented by the amino acid sequence of its DNA binding domains.

4.3.2 Modeling TF-TFBS sequence pairs

Encoding TF Properties. Each TF was encoded by three sets of characteristics: a/ the physicochemical properties of amino acids obtained from the AAindex database (Kawashima and Kanehisa, 2000), last database update March 31, 2008 (Kawashima, et al., 2008), b/ the DNA binding domain family classification, and c/ the amino acid binding mode preference to DNA bases obtained from (Luscombe and Thornton, 2002). For the first set of properties, we used numerical values of 544 physicochemical properties of amino acids available in the AAindex database version 9.1. A feature i of TF_j is the average value of the physicochemical property i in the sequence of the DNA binding domain of TF_j weighted by the relative occurrences of individual amino acids in the sequence:

$$\text{Feature}_i(TF_j) = \sum_{\text{Amino Acid } k=1}^{20} \frac{\text{Freq}_k}{\text{length}(TF_j)} * \text{Property Value}_i(k), \quad (4.1)$$

where Freq_k is the number of times amino acid k is found in the sequence of the DNA binding domain of TF_j ; $\text{Property Value}_i(k)$ is the numerical value of physicochemical

property i for amino acid k ; $length(TF_j)$ is the number of amino acids in the sequence of the DNA binding domain of TF_j . We used the same formula for each of the 544 features, which resulted in a 544-dimensional vector for each TF_j .

Since all TF DNA binding domains obtained from the Pfam database belong to 72 domain families, we used 7-binary digits to encode them and this represents the second set of properties of TFs that we used. While 55% of TFs have only a single DNA binding domain, the remaining 45% have more than one. In such cases, we encode the corresponding part of the feature vector with the domain with the lowest E-value, as such domains should be more statistically significant. The third set of TF properties were determined and encoded as follows. Amino acids were classified into three categories according to their binding mode preference to the DNA bases as in (Luscombe and Thornton, 2002) (Table 4.2). These categories are: i) they bind to DNA bases through hydrogen bonds, ii) they bind to DNA bases through van der Waals contacts, or iii) they do not interact with DNA bases in significant numbers. As the last three features to describe a TF we used the weighted occurrence of amino acids in these three categories of amino acid binding preferences to DNA bases:

$$\text{Feature}_{k=1,2,3}(TF_j) = \frac{Freq_k}{length(TF_j)}, \quad (4.2)$$

where $Freq_k$ is the total number occurrences of amino acids that belong to category k (there are three categories, Table 4.2) in the sequence of the DNA binding domain of

TF_j ; $length(TF_j)$ is the number of amino acids in the sequence of the DNA binding domain of TF_j . The final set of properties to represent a TF consists thus of 554 (544+7+3) properties.

Table 4.2. Amino acid classification. Distribution of 20 Amino acids according to their binding mode preference to DNA bases, adopted from Table 4 in (Luscombe and Thornton, 2002).

Category	Amino acids	Mode of interaction
i)	Arg, Lys, His, Ser, Asn, Gln, Asp, Glu	Hydrogen bonds
ii)	Phe, Pro, Thr, Gly, Ala, Val, Leu, Iso, Tyr	van der Waals contacts
iii)	Cys, met, Trp	No base contact

TFBS Representation. Each TFBS that consists of L nucleotides was represented using a vector of length $4*L$ obtained as follows. Each of the four nucleotides (A, C, G, T) in the TFBS sequence was encoded by a 4-digits binary number as follows: A as 0001, C as 0010, G as 0100 and T as 1000. A TFBS is then represented as a vector of length $4*L$ by concatenating the binary sequences corresponding to its nucleotide sequence as described. For example, 'ACTCCGAT' will be represented by '00010010100000100010010000011000'. The TFBSs of the selected 232 TFs (associated with 250 TFBS models) have 14 distinct lengths $L = 7 nt, 8 nt, 9 nt, \dots, 19 nt, 20 nt$.

Combining TF and TFBS Descriptions. Both TF and TFBS properties were combined in one TF-TFBS feature vector as follows. Suppose that T_i and B_j are the feature row-vectors for TF_i and $TFBS_j$, respectively. We define the combined TF-TFBS feature vector D as:

$$D = [T_i, B_j], \quad (4.3)$$

For example, when $L = 12$, the TF-TFBS pair is coded by a 602-dimensional (554 TF properties plus 12×4 TFBS properties) vector. If a TF is associated with N TFBSs, then we will have N TF-TFBS pair vectors, where the first part, TF vector, remains the same across all N vectors.

4.3.3 Data preparation

Normalization. To remove the bias that arises from different ranges of values used for TF and TFBS features, we normalized each feature by scaling minimum and maximum values to 0 and 1, respectively, as follows:

$$x'_i = \frac{x_i - \min(X)}{\max(X) - \min(X)} \quad (4.4)$$

where x_i is the original feature value, x'_i is the value after normalization. X is the list of feature values x_i across all samples. $\min(X)$ and $\max(X)$ are the minimum and the maximum values of X , respectively. In addition, the feature vectors are of different lengths due to varying TFBS lengths. Since a separate model is built for each of the 14 TFBS lengths, this does not cause any problems in the analysis.

TF Feature Selection. It is widely acknowledged that irrelevant and weakly relevant features may decrease the accuracy of predictors (Kohavi, 1994). Among different feature selection methods that we examined, namely, the Minimum Redundancy Maximum Relevance (mRMR) method, individual features providing highest AUC (Area Under Curve) and forward sequential subset feature selection methods, we found that mRMR method yielded the highest accuracy. Consequently, we used the mRMR method to identify TF properties of relevance to distinguish between the two classes of data, binder to a TFBS and non-binders to considered TFBS (see the next section). We select the top $N=150$ features with the highest mRMR scores out of the initial set of 554 features. This threshold of 150 top ranked features was subjective, but we are focusing on the best features suggested by mRMR.

4.3.4 Positive (true) and negative (false) data

Positive Data. The ‘positive’ set consists of 110,399 TF-TFBS pairs (‘positive’ pairs) that correspond to 232 TFs and their associated 110,399 TFBS targets.

Negative Data. For each TF-TFBS pair we produced a presumably ‘false’ TF-TFBS pair by preserving the TF feature part of the feature vector, but randomly selecting sequences from human chromosomes 4 and 22 to correspond to the ‘TFBS’ sequence part in the feature vector. These two chromosomes were used because chromosome 4 has the lowest (~38%) GC content, while chromosome 22 is one of the two chromosomes that have the highest (~48%) GC content across all human chromosomes. From the initial ‘negative’ TF-TFBS set, we excluded all TF-TFBS pairs that were also contained in the

'positive' dataset, which resulted in the final 'negative' data. 'Positive' and 'negative' data were given different class labels. Finally, for each 'positive' TF-TFBS pair, we created in this way 10 'negative' TF-TFBS pairs to make the ratio between the number of 'negative' and 'positive' samples 10 to 1.

Training and test sets. We split all data into 14 groups corresponding to the 14 different TFBS lengths that we considered. Then, separately for each of these groups, we generated training and test data, and based on that we developed one prediction model for each of the groups. Note that for 214 out of the 232 TFs, each of the groups was associated with mutually distinct sets of TFs assigned to the group based on length of their TFBSs, i.e. if a TF was associated with one of the groups, it did not appear associated with any of the other groups. However, for the remaining 18 TFs, they were associated with two groups because they have two sets of TFBSs corresponding to different TFBS lengths. We pooled 'positive' and 'negative' data together, and used 70% of the data for training and the 30% for testing. This division was made at random on the TF level such that 70% of the TF-TFBS pairs of a particular TF were used for training and the remaining 30% were used for testing. Within the group, such training and testing data were pooled separately. Separately from this test, we performed 5-fold and 10-fold cross-validation on the whole datasets, and reported the obtained results from each experiment. We set thresholds on the model outputs that yield the highest accuracy on the training data and used these thresholds when evaluated the model performance on the test data. The same is done in cross-validation.

4.3.5 Random forests TFBS prediction model

DRAF uses random forests (RF) (Breiman, 2001) to model the relationship between TFs and their TFBSs represented as TF-TFBSs pairs. The 250 TFBS models associated with 232 TFs fall into 14 groups according to the length of their TFBS sequences (Table 4.1). We built 14 prediction models accordingly, such that one model represents all TF-TFBS pairs with a common TFBS length. Each prediction model is represented as “random forests” composed of an ensemble of 80 decision trees. We tested a range of decision trees in the ensemble (10, 20,..., 150) and found that a random forest composed of 80 decision trees demonstrated the highest accuracy on the training data. In the training phase, the model was trained with all TF-TFBS pairs in the training set that belong to ‘positive’ and ‘negative’ classes. In the testing phase, the trained model was used to predict the ‘positive’ or ‘negative’ class of a particular TF-TFBS pair sample in the test set after the features in the test set were normalized using parameters of obtained from scaling on the training set. Tests by cross-validation were done in a standard way using the scaling as explained above, and the test results reported here were the average across all the folds.

4.3.6 Model evaluation metrics

The quality of the model was evaluated by accuracy, sensitivity, specificity, precision, F-measure and Matthew’s correlation coefficient (MCC), partially motivated by an objective to allow for comparison with other existing prediction methods. These performance measures are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (4.5)$$

$$Sensitivity (Recall) = \frac{TP}{TP + FN} \quad (4.6)$$

$$Specificity = \frac{TN}{FP + TN} \quad (4.7)$$

$$Precision = \frac{TP}{TP + FP} \quad (4.8)$$

$$Fmeasure = \frac{2 * (precision * recall)}{(precision + recall)} = \frac{2 * TP}{2 * TP + FN + FP} \quad (4.9)$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4.10)$$

TP (true positive) represents correctly predicted 'positive' TF-TFBS pairs, TN (true negative) represents correctly predicted 'negative' pairs, FP (false positive) represents 'negative' pairs incorrectly predicted as 'positive' pairs, and FN (false negative) represents 'positive' pairs incorrectly predicted as 'negative' pairs.

4.3.7 Comparison of DRAF RF models with other model types

We compared the prediction results of the DRAF models with three other types of machine learning models, namely Neural Networks (NN), Support Vector Machines

(SVM) and Gaussian Mixture Regression (GMR) models. For each model, we set parameters to provide the highest accuracy on the training data. For Neural Networks, we tested feed-forward back-propagation network with two and three hidden layers each with either 100 or 200 neurons, and using sigmoid functions for hidden layers and linear transfer function for output layer. Based on the accuracy obtained from testing these options on the training data, we used feed-forward back-propagation network with three hidden layers using the sigmoid transfer function, each with 100 neurons, and a linear output layer. The maximum number of epochs to train was set to 500 and the learning rate was set to 0.05 with the performance goal of 1×10^{-5} . For Support Vector Machines (Cortes and Vapnik, 1995), we tried four different types of kernels namely, linear, polynomial, radial basis and sigmoid. We used the radial basis function (Burges, 1998; Schölkopf and Smola, 2002) within the LIBSVM (Chang and Lin, 2011) implementation of SVM which provided the highest accuracy on the training data. We tested different values for the gamma (0.0625, 0.125, 0.25, 0.5, 1 and 2) and the regularization (cost) (0.5, 1, 2, 4 and 8) parameters. The kernel parameter values, which provided highest accuracy on the training data, were set to 0.125 and 8 for the gamma and regularization parameters, respectively. For Gaussian Mixture Regression, we used (Calinon, 2009; Calinon, et al., 2007) implementation and tested different number of Gaussian components (5, 10, 15, 20 and 25). We finally set the number of Gaussian components to 20 as this provided the best performance on the training data.

4.3.8 DRAF model validation on ChIP-seq data

ChIP-seq data. To measure the capability of the DRAF models to predict TFBSs with high sensitivity and specificity, we evaluated the DRAF models using independent ChIP-seq datasets. For this purpose, we used sequences of all human ENCODE ChIP-seq peaks that were processed and assigned signal scores by the ENCODE uniform processing pipeline (Consortium, 2012). Consequently, we retrieved 690 ChIP-seq datasets that related to 165 unique TFs evaluated in different cell types, and 58 of these 165 TFs were among the 232 TFs we used to construct the DRAF models. For each of these 58 TFs, we selected all corresponding ChIP-seq datasets from all available cell types. From each dataset, we used the top 500 sequences having the highest peak enrichment scores. This resulted in a total of 321 ChIP-seq datasets, each of which consists of 500 sequences.

TFBS extraction from ChIP-seq data. TFBS part in the TF-TFBS feature vector was constructed from each ChIP-seq peak of length N by extracting sequences of length L starting from the first position until the end of the sequence and moving each time by one nucleotide. This resulted in a total of $N-L+1$ TFBS sequence parts. The same number of TFBS sequence parts was extracted from the reverse complement sequence of the ChIP-seq peak, resulting in a total of $2*(N-L+1)$ TFBS sequence parts. For example, if a particular ChIP-seq peak has ($N=100$ bp), and the TFBS length ($L=10$), then we will extract 91 TFBS sequence parts from this ChIP-seq peak and 91 TFBS sequence parts from the reverse complement sequence. Then, each of these TFBS sequence parts was represented in the same binary representation explained above. The TF part of the

feature vector consists of the properties of the TF for which this ChIP-seq data belongs to. Then, TF-TFBS pairs were constructed by associating the TF part of the feature vector with each of the TFBS parts. Finally, we used DRAF model to examine all these TF-TFBS pairs and predict correct TF-TFBS associations. A ChIP-seq peak was declared to be correctly predicted (i.e. true positive) if at least one of the $2^{*(N-L+1)}$ TF-TFBS pairs were identified by the DRAF model to be a correct pair (i.e. the TFBS is a correct binding site for the associated TF). If none of the $2^{*(N-L+1)}$ TF-TFBS pairs that belong to a particular ChIP-seq peak were identified by the DRAF model to be a correct pair, then this ChIP-seq peak was considered as a false negative. To visually evaluate the quality of DRAF model predictions on the ChIP-seq data, we plotted the sequence logo for all the ChIP-seq predictions made by the DRAF model for each TF using WebLogo tool (Crooks, et al., 2004) and compared the sequence logo with the standard TF sequence logo obtained from HOCOMOCO.

'Negative' data. The ChIP-seq data enables the estimation of the DRAF model sensitivity. However, to estimate the model's specificity, we constructed 'negative' (background, false) TFBS sequence parts from human chromosome 21 (average CG content ~41%). First, we excluded from chromosome 21 all regions covered by ENCODE ChIP-seq peaks that belong to any cell type for a specific TF. For this reason, preparation of the 'negative' data from chromosome 21 differs from one TF to another. Then, TFBS sequence parts of length L were extracted starting from the first position until the end of chromosome 21 sequence and moving each time by one nucleotide. This extraction

process was repeated on the reverse complement sequence of chromosome 21. TFBS sequence parts that contained ambiguous nucleotides were not considered. Feature vectors describing TF-TFBS pairs were compiled in the same way as explained previously by associating the TF properties with TFBS representation for all TFBS sequence parts. After that, we removed from TF-TFBS collection all pairs that were found among the training set of the corresponding DRAF model. Consequently, the number of TF-TFBS pairs in the ‘negative’ set varies from one TF to another depending on the excluded regions from chromosome 21 due to the overlap with CHIP-seq peaks, the excluded TF-TFBS pairs due to the overlap with the training set and the TFBS length L . Finally, we used the DRAF models to examine all these TF-TFBS pairs. If a particular TF-TFBS pair of the ‘negative’ set was incorrectly predicted as ‘positive’, we considered this pair as a false positive prediction. Similarly, if the TF-TFBS pair was correctly predicted as a ‘negative’ pair, we considered this pair as a true negative prediction.

4.3.9 Comparison between DRAF models, PWM models (HOCOMOCO, TRANSFAC) and DeepBind models on CHIP-seq datasets

Position Weight Matrix (PWM) models. We compared the predictive performance of the DRAF models with the PWM models obtained from the HOCOMOCO (version 9) and TRANSFAC (version 2012.2) databases. For each of the 58 TFs that we considered, we used the corresponding HOCOMOCO and TRANSFAC PWMs that model the respective TFBSs. Then, we scanned the CHIP-seq peaks and the chromosome 21 (as explained in the previous section) using MEME FIMO (Grant, et al., 2011) to report PWM matching

scores on these sequences. At different sensitivity levels (10%, 20%, ..., 90%) we compared the scores obtained from the DRAF models with those obtained from the PWM models by monitoring the change in specificity and the average distance in nucleotides (nt) between false positive prediction occurrences on the 'negative' sequences (derived from chromosome 21).

DeepBind Models: We repeated the same comparison that we performed between DRAF models and PWM models but this time with DeepBind (Alipanahi, et al., 2015) models. We found 54 out of the 58 TFs that we tested using ENCODE ChIP-seq data to have a DeepBind model. This resulted in a comparison of the DRAF models with the DeepBind models in 302 out of the total 321 ChIP-Seq datasets that we retrieved from ENCODE data.

4.4 Results

4.4.1 Selected properties of TFs

To describe each TF, we used 150 features including AAindex properties, DNA binding domain family classification, and amino acid binding mode preference to DNA bases. These properties were selected as top 150 ranked ones based on the mRMR method (Peng, et al., 2005). In the 14 models corresponding to different TFBS lengths, on average, out of these 150 selected features, 145 are AAindex properties, while 5 reflect other properties we introduced (see below). Therefore, on average, one out of four AAindex properties (27% = 145/544) and one out of two other features were selected

through filtering. Of the 145 selected AAindex properties, on average, 115 can be classified into six groups, while 30 are 'unclassified', according to (Kawashima, et al., 2008) (see Figure 4.1). We notice from Figure 4.1 that almost half of the selected features from AAindex (46%) can be classified as belonging to the hydrophobicity or alpha and turn propensity groups. Although this may highlight the role of these two categories of features relevant to our models for predicting the TF affinity to TFBSs, we also notice that the distribution in groups of our AAindex selected features reflects well distribution of all the 544 AAindex properties (Kawashima, et al., 2008).

The remaining 5 selected features represent the other properties we introduced, namely the DNA binding domain family classification and the amino acid binding mode preference to DNA bases. In particular, on average, 4 out of 7 features (57%), used to describe the DNA binding domain family, and 1 out of 3 features (33%), used to represent the amino acid binding mode preference, were selected in the 150 top ranked ones. This suggests a high information value of the features we added to those from AAindex for predicting TF-TFBS links in the method we used.

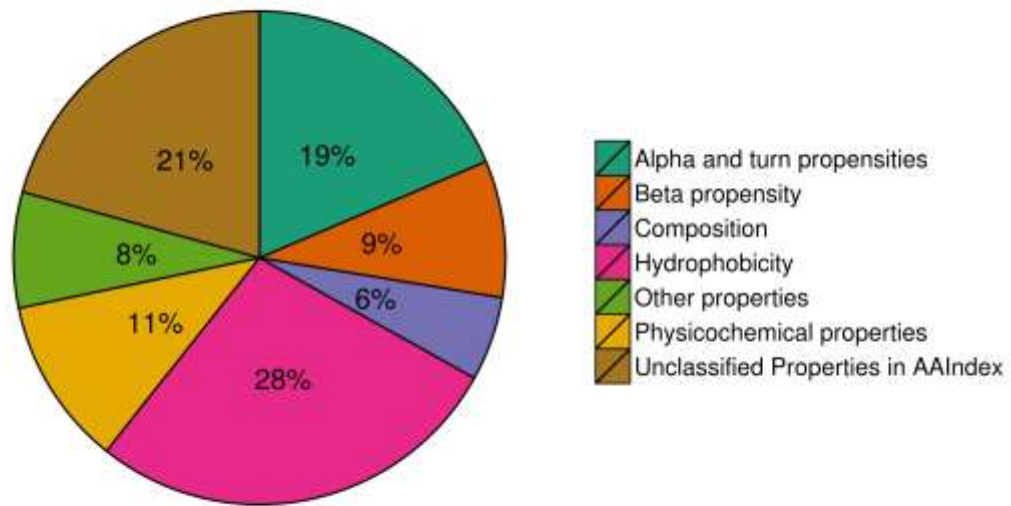


Figure 4.1. Distribution of the selected properties used by DRAF models. Distribution of 145 selected AAindex properties; 115 (79%) properties belong to six biological classes, while other 30 (21%) properties are ‘unclassified’ according to (Kawashima, et al., 2008).

4.4.2 TF-TFBS predictions by DRAF models

DRAF models were first trained using sets of TF-TFBS pairs and then used to predict TFBSs of TFs (Figure 4.2). In order to measure the capability of the DRAF models in predicting the TF-TFBS pairs, we applied two different testing strategies using holdout and cross-validation methods. In the first experiment, based on the holdout approach, the average accuracy, sensitivity, specificity and precision obtained from applying all the

14 DRAF models to the test data were 99.16%, 92.53%, 99.86% and 98.57%, respectively (Figure 4.3, Appendix Table A4.1). These results were obtained using thresholds on the model output scores that provided the highest accuracy on the training data. We repeated the same experiment using the thresholds that provided the highest specificity and the highest sensitivity on the training data (Figure 4.3, Appendix Table A4.2 and A4.3).

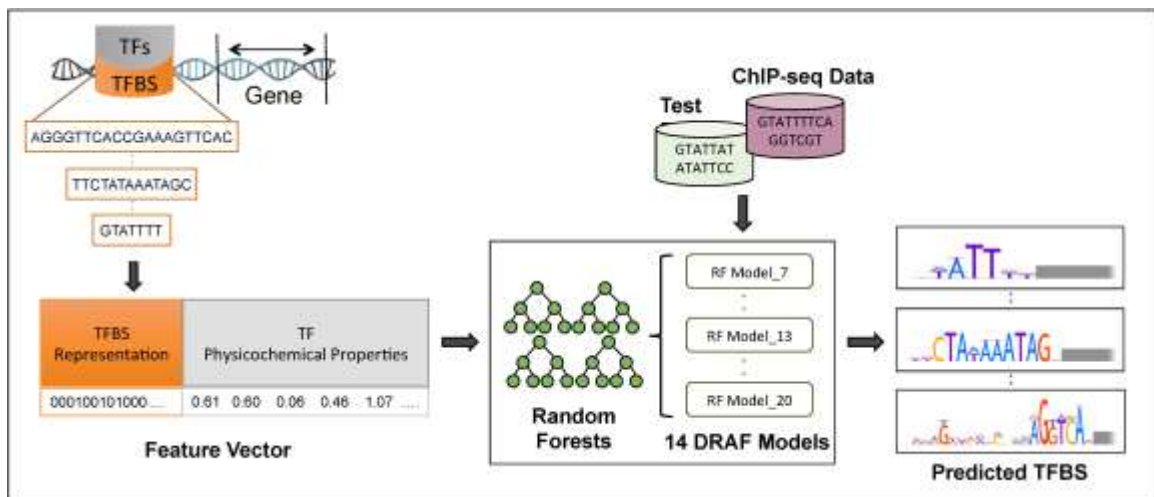


Figure 4.2. The input data, training procedure and usage of DRAF models for prediction of TF-TFBS pairs. Sequences of TFs and their TFBS are represented in TF-TFBS pairs using physicochemical properties of TFs and binary representation of TFBSs. Then, DRAF models were constructed for each group of TFs depending on the TFBS length. Finally, DRAF models were tested using the test data and another set of independent ChIP-seq validation datasets. The DRAF models predict which TF-TFBS pair represent a valid target TFBS for a particular TF.

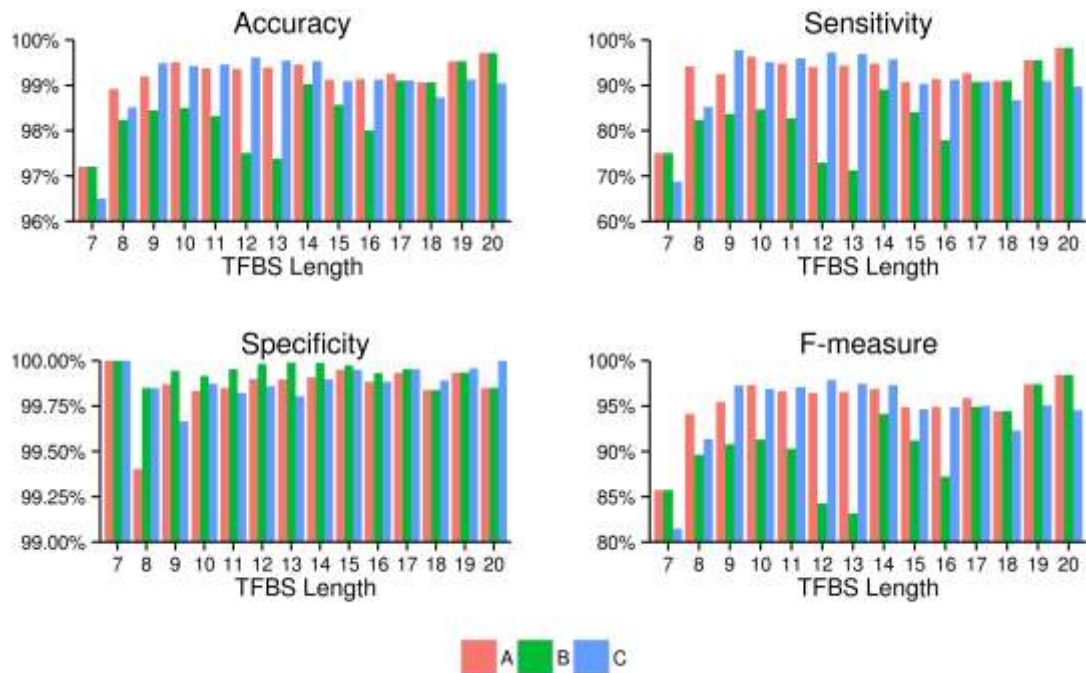


Figure 4.3. The prediction performance of DRAF models on the test data. DRAF models were applied on the test data using different settings for selecting thresholds on the models' prediction score that provided (A) the highest accuracy, (B) the highest sensitivity and (C) the highest specificity, on the training set.

In addition, Figure 4.4 demonstrates that receiver operating characteristic (ROC) curve for 14 models (AUC=0.9991). This high value of AUC and the ROC curve in Figure 4 suggest that the DRAF models could predict the TF-TFBS relationship with very high accuracy for all modeled TFs.

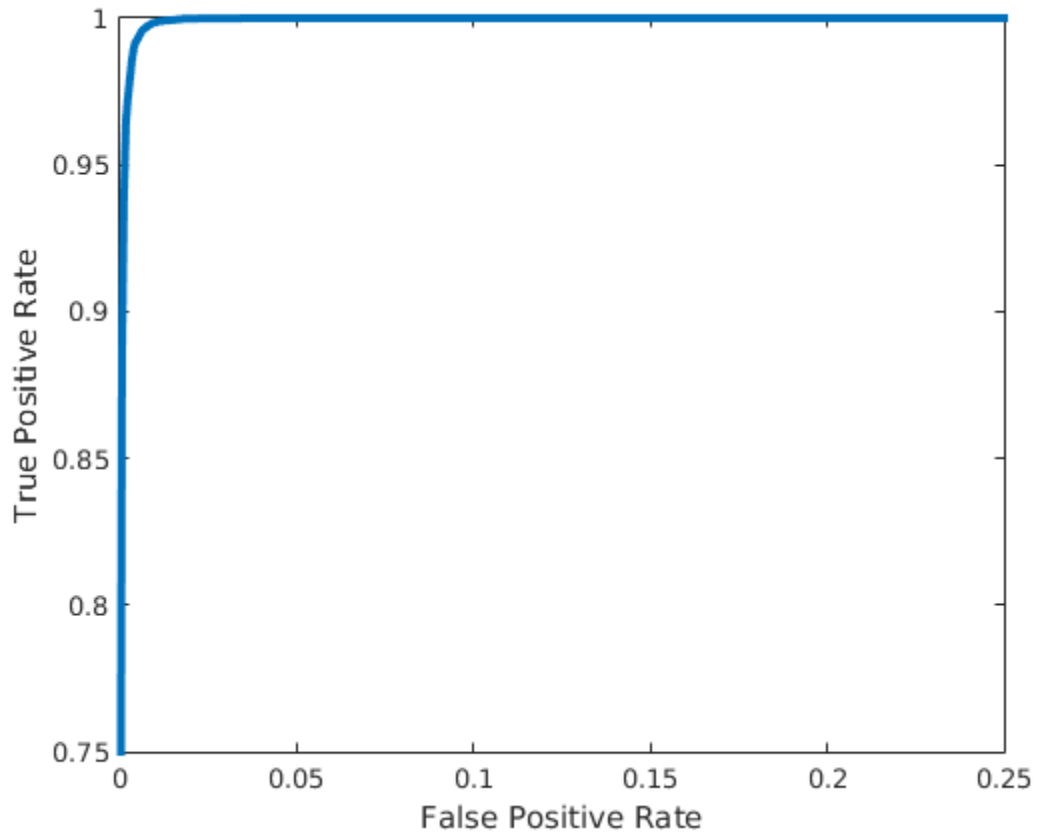


Figure 4.4. The ROC curve for the evaluation of the DRAF models on the test set. The ROC curve (true positive rate vs. false positive rate) was performed for the prediction outputs obtained from all the 14 DRAF models on the test set. The AUC for the DRAF models is 0.9991.

We repeated the evaluation experiments using 5-fold and 10-fold cross validation in addition to the holdout method. Appendix Tables A4.4-6 shows the prediction results using 10-fold cross validation for all the 14 models using the thresholds giving the highest accuracy, specificity and sensitivity, respectively, on the training data. Appendix

Tables A4.7-9 show similar results to those viewed in Appendix Tables A4.4-6 but obtained using 5-fold cross validation.

Finally, we compared the prediction results of the DRAF models with NN, SVM and GMR models (see Methods). DRAF models outperformed other models in terms of accuracy, specificity and precision (Figure 4.5, Table 4.3). It, however, yielded lower sensitivity than NN and SVM but higher than the GMR models.

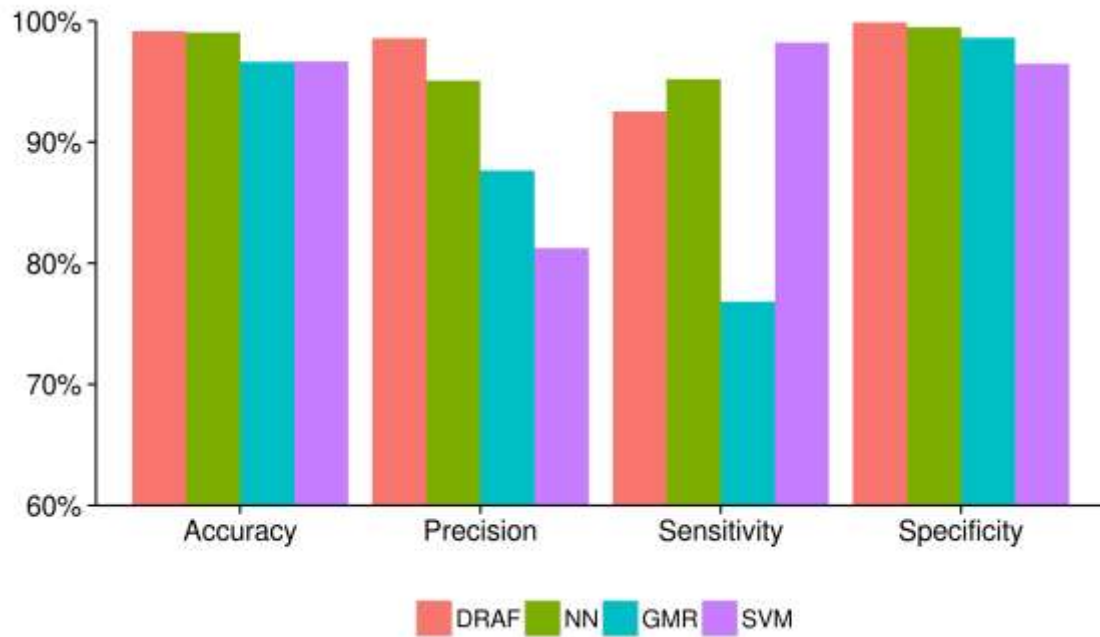


Figure 4.5. Comparison between RF, NN, GMR and SVM models. All the four types of models were trained using the same training data and applied on the same test data. The models were compared using performance measures calculated on the test data.

Table 4.3. Comparison between DRAF and other model types. The average prediction results on the test data using 4 model types, DRAF, Neural Networks (NN), Gaussian Mixture Regression (GMR) and Support Vector Machines (SVM).

	DRAF	NN	GMR	SVM
Accuracy	99.16%	99.06%	96.62%	96.66%
Sensitivity	92.53%	95.19%	76.83%	98.20%
Specificity	99.86%	99.49%	98.63%	96.48%
Precision	98.57%	95.08%	87.63%	81.25%

4.4.3 Model evaluation on ChIP-seq data

We evaluated the predictive performance of the DRAF models on 321 ENCODE ChIP-seq datasets derived from 321 different human cell types (see Methods). We changed the thresholds on the model output scores to obtain the predictions at different sensitivity levels (10%, 20%... 90%) and monitored the capability of the DRAF models to recognize background sequences (this was measured using the average distance between predictions on the background sequences). The results show that the DRAF models accurately predict TFBSs on sequences of ChIP-seq peaks, while maintaining high specificity of predictions on the background sequences.

We examined the similarity of the predicted TFBSs on the ChIP-seq peaks to the known TFBSs of each TF. The sequence logos for TFBS predictions at different sensitivity levels showed high similarity to known sequence logos obtained from HOCOMOCO

(Kulakovskiy, et al., 2013) for the corresponding TFs (Figure 4.6, Appendix Table A5.1).

This suggests that each of the 14 DRAF models was capable to capture the DNA binding patterns of the TFs encoded by that model.


























TF	Cell Type	Original TFBS Sequence Logo	Sensitivity 60%	Sensitivity 70%	Sensitivity 80%	Sensitivity 90%
ATF1	K562					
ATF3	A549					
BATF	GM12878					
BRCA1	GM12878					
CEBPB	A549					

Figure 4.6. Sequence logos for the predicted TFBS sequences on the human ChIP-seq datasets using DRAF models. The figure shows different sequence logos obtained from the DRAF predicted TFBS sequences from ChIP-seq datasets at different sensitivity levels. The complete set of sequence logos for the 321 ENCODE ChIP-seq datasets is provided in Appendix Table A5.1.

We next compared the performance of the DRAF models with HOCOMOCO PWM models, TRANSFAC PWM models and DeepBind models on the 321 ChIP-seq datasets

(see Methods). For this purpose, we set thresholds on the DRAF model prediction scores that yielded highest F-measure scores on the training data and used these thresholds when evaluated the model performance on the ChIP-seq datasets (see Methods). The results show that DRAF models got higher specificity on the background data as compared to the other three types of models, while having better or at least the same sensitivity levels as the other methods. That is, the DRAF models provided (on average of all sensitivity levels) 14-, 15- and 16-folds less frequent false positive predictions on the background data than HOCOMOCO PWMs, TRANSFAC PWMs and DeepBind models, respectively (Figure 4.7).

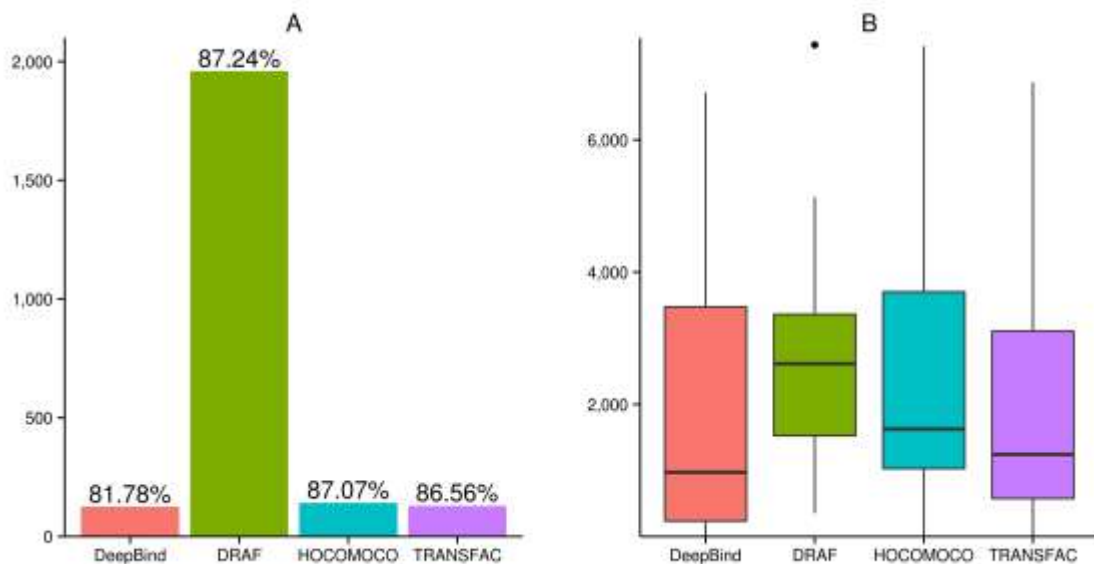


Figure 4.7. Comparison of DRAF, HOCOMOCO, TRANSFAC and DeepBind models using thresholds on the DRAF model output scores that provided the highest F-measure.

Comparison of performance of 14 DRAF models and 321 HOCOMOCO PWMs, 319 TRANSFAC PWMs and 302 DeepBind models on 321 ChIP-seq datasets obtained from

ENCODE using thresholds on the DRAF model prediction scores that yielded highest F-measure scores on the training data. In A, the Y-axis represents the logarithm of distance (in nt) between false positive prediction occurrences on the background sequences (chromosome 21) averaged over all the tested ChIP-seq datasets. The labels on the top of the blue bars indicate the average sensitivity on the ChIP-seq datasets using the corresponding model. In B, the box plots show distance (in nt) between false positive prediction occurrences on the background sequences (chromosome 21) across all the tested ChIP-seq datasets.

The previous results were obtained by comparing DRAF and other models on ENCODE ChIP-seq datasets using thresholds that provided the highest F-measure scores on the training data. We repeated this comparison between all types of models by monitoring their performance at different sensitivity levels (see Methods). At all studied sensitivity levels the DRAF models outperformed HOCOMOCO, TRANSFAC and DeepBind models, providing higher specificity on the background datasets. That is, the DRAF models yielded higher specificity (averaged over all sensitivity levels) than each of the HOCOMOCO and TRANSFAC and DeepBind models in 89.41% (287 out of 321 ChIP-seq datasets), 93.35% (298 out of 319 ChIP-seq datasets) and 91.91% (278 out of 302 ChIP-seq datasets), respectively, of the tested ChIP-seq datasets by each model type. DRAF models provided higher specificity (averaged over all sensitivity levels) than any of the other three model types together in 80.24% of the tested ChIP-seq datasets. Another

measure that we used to compare the models was by calculating the average distance between false positive predictions on the background datasets (see Methods) and averaged that over all the tested background datasets. We found that at each of the tested sensitivity levels, DRAF models provided smaller number of false positive predictions than all of the other three models (Figure 4.8 and 4.9). These results show that DRAF models provided false positive predictions (averaged over all sensitivity levels) 2-, 3- and 5-folds less frequent than HOCOMOCO PWMs, TRANSFAC PWMs and DeepBind models, respectively (Figure 4.10, Table 4.4).

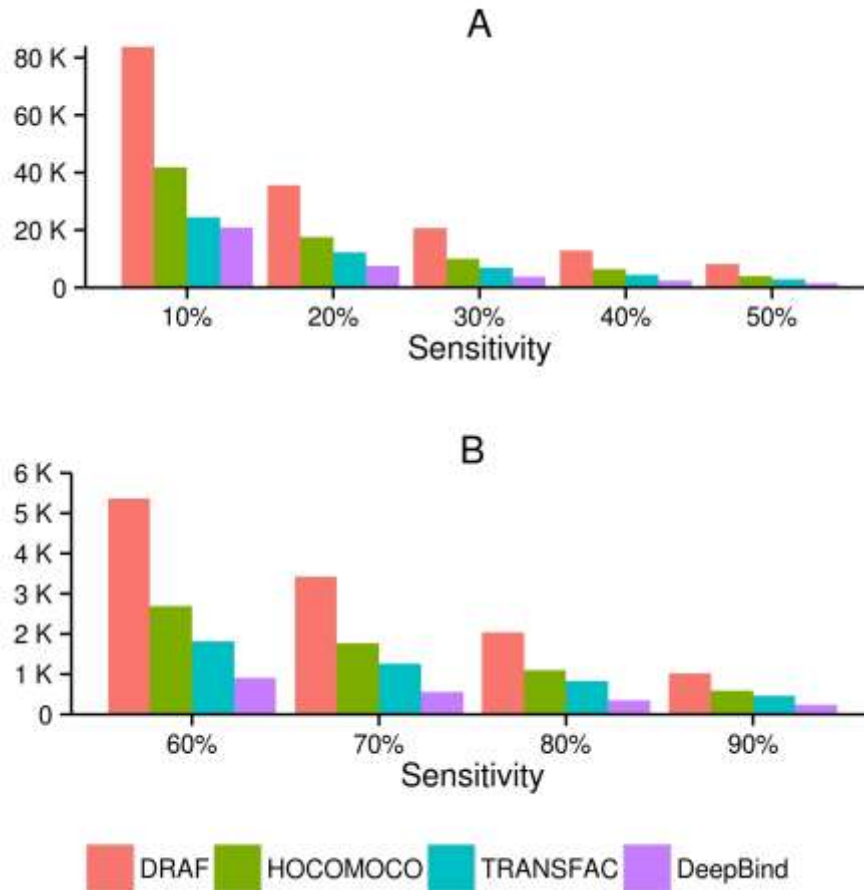


Figure 4.8. Comparison of DRAF, HOCOMOCO, TRANSFAC and DeepBind models at different sensitivity levels (averaged over all ChIP-seq datasets). Comparison of 14 DRAF models and 321 HOCOMOCO PWMs, 319 TRANSFAC PWMs and 302 DeepBind models on 321 ChIP-seq datasets obtained from ENCODE. The X-axis represents different sensitivity levels (A: 10%, 20%...50% and B: 60%, 70%...90%) and the Y-axis represents the distance (in nt) between false positive prediction occurrences on the background sequences (chromosome 21) averaged over all the tested ChIP-seq datasets.

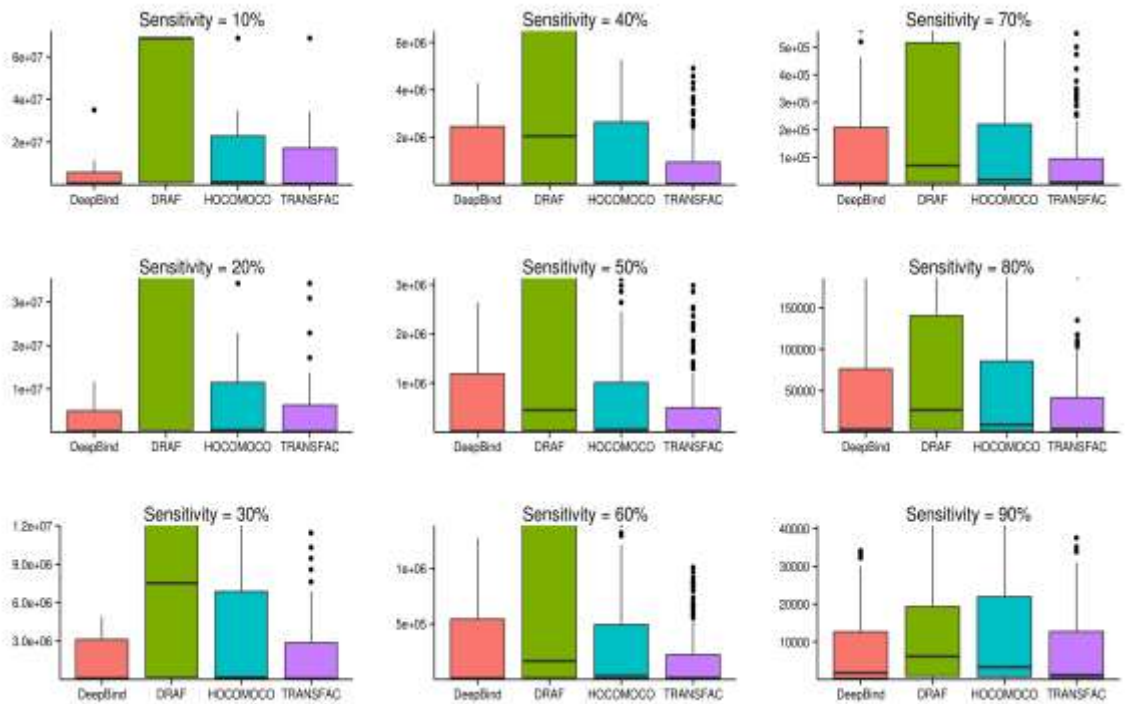


Figure 4.9. Comparison of DRAF, HOCOMOCO, TRANSFAC and DeepBind models at different sensitivity levels. Comparison of 14 DRAF models and 321 HOCOMOCO PWMs, 319 TRANSFAC PWMs and 302 DeepBind models on 321 ChIP-seq datasets obtained from ENCODE at different sensitivity levels (10%, 20%...90%). The Y-axis represents the logarithm of the distance (in nt) between false positive prediction occurrences on the background sequences (chromosome 21) across all the tested ChIP-seq datasets.

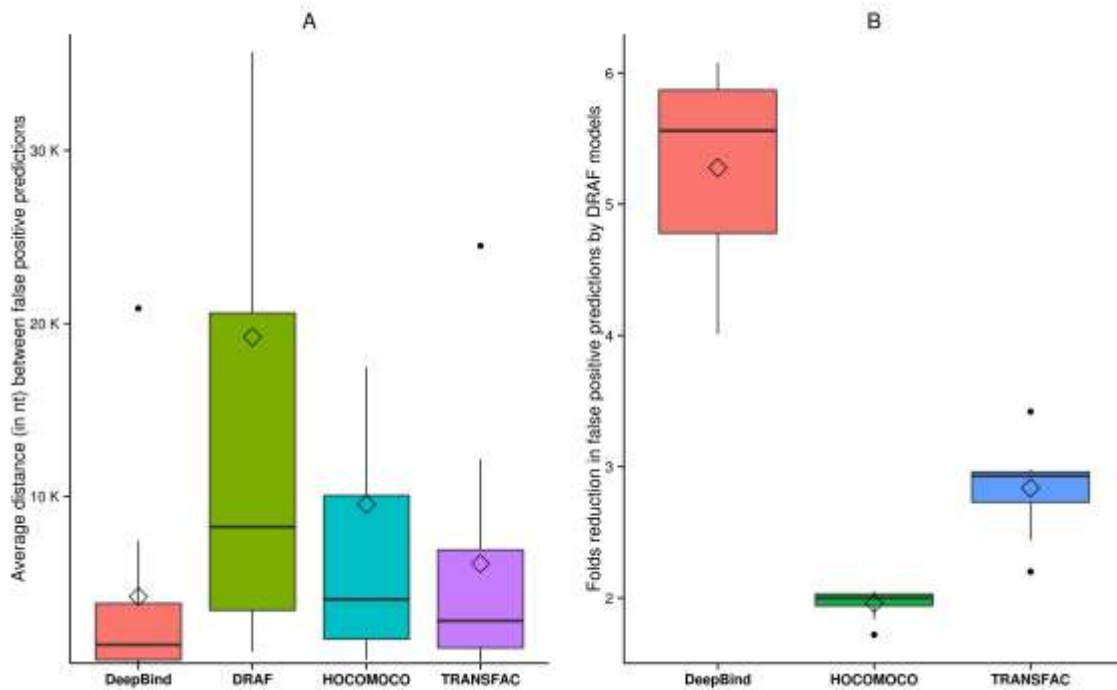


Figure 4.10. Summary of Comparison between DRAF, HOCOMOCO, TRANSFAC and DeepBind models. A: boxplots show the average distance (in nt) between false positive prediction occurrences on the background sequences (chromosome 21) at different sensitivity levels obtained from testing corresponding models on the ChIP-seq datasets and averaged over all the datasets at each sensitivity level. B: boxplots show the folds-reduction in false positive predictions obtained in by DRAF models as compared to the corresponding HOCOMOCO, TRANSFAC and DeepBind models.

Table 4.4. Average distance (in nt) between false prediction occurrences on the background sequences.

A	B	C	D	E			
Sensitivity	DRAF	TRANSFAC	HOCOMOCO	DeepBind	B/C	B/D	B/E
10%	83,682	24,488	41,782	20,871	3.42	2	4.01
20%	35,682	12,161	17,471	7,464	2.93	2.04	4.78
30%	20,601	6,916	10,064	3,825	2.98	2.05	5.39
40%	12,874	4,369	6,377	2,315	2.95	2.02	5.56
50%	8,250	2,820	4,063	1,436	2.93	2.03	5.75
60%	5,365	1,815	2,695	913	2.96	1.99	5.88
70%	3,431	1,256	1,771	564	2.73	1.94	6.08
80%	2,038	834	1,107	347	2.44	1.84	5.87
90%	1,021	464	593	245	2.2	1.72	4.17
				Average	2.84	1.96	5.28

4.4.4 Comparative performance of DRAF models

Comparison of the performance of our models with other similar works is not straightforward, due to the differences between approaches and criteria used in different studies. Some models were assessed only for individual TFs or specific TF families. For example, the model created by Ellrott and colleagues (Ellrott, et al., 2002) was evaluated on HNF4 α ; the model developed by Endress (Endres, et al., 2004) was evaluated only on Zif268; Alamanova and colleagues (Alamanova, et al., 2010) tested their model on few TFs such as P53 and NF- κ B. Liu and Bader (Liu and Bader, 2007)

reported results on Mat- α 2 and GCN4 bZIP. Chen and others used six TFs to test their model (Chen, et al., 2012).

Therefore, we focused on the comparison of the DRAF performance and the reported results in (Cai, et al., 2009; Qian, et al., 2006; Qian, et al., 2007) as these studies used more than 100 TFs each. The summary of this comparison is given in Table 4.5. For this, we evaluated DRAF performance on a much more comprehensive and significantly larger test datasets (Table 4.5, Columns 2 and 3) than in these publications and results show that the DRAF models outperformed other models reported in the above-mentioned three studies (Table 4.5).

Table 4.5. Comparison of prediction results from different studies. This table shows the prediction accuracy of DRAF models and other models on different TF-TFBS test datasets.

Study	Number of TFs	Number of unique TFBSs	'Positive' data	'Negative' data	Highest accuracy
(Qian, et al., 2006)	480	2,341	3,356	6,850	76.6%
(Qian, et al., 2007)	143	571	3,430	7,000	87.9%
(Cai, et al., 2009)	599	2,402	3,541	31,869	91.1%
DRAF models	232	44,710	110,399	1,103,990	99.16%

4.5 Discussion

We developed DRAF models to capture the relationship between TFs and their TFBSs. One DRAF model is developed to capture relationships of all TFs whose associated TFBSs have the same length. Consequently, we needed 14 DRAF models to represent TF-TFBS pairs that covered 250 TFBS sets and 232 TFs. These TFBSs have 14 distinct lengths (7 nt, 8 nt, ..., 20 nt). PWMs and other model types (e.g. DeepBind models) are usually developed for TFBSs for individual TFs, resulting in numerous models (one or more models for a single TF). This is the reason that major TFBS model databases (HOCOMOCO, TRANSFAC or JASPAR) have a very large number of TFBS models. DRAF, however, successfully reduces the number of required models to only 14 models, based on the length of the TFBSs. So one such model captures many TFs. For 232 TFs, DRAF requires only 14 models, ~18 times less than the corresponding 250 TFBS models in HOCOMOCO, ~54 times less than the corresponding 749 TFBS models in TRANSFAC (231 TFs of the 232 TFs were found TRANSFAC) and ~50 times less than the 704 TFBS models in DeepBind (124 TFs of the 232 TFs were found in DeepBind).

This dramatic reduction in the number of models did not decrease the performance. On the contrary, the results showed that DRAF models significantly outperformed all three: HOCOMOCO PWMs, TRANSFAC PWMs and the DeepBind models in 80.24% of the 321 ENCODE ChIP-seq data of human cell types. Even more, DRAF models increased specificity by generating on average false positive predictions 14-, 15- and 16-folds less frequent than the HOCOMOCO PWMs, TRANSFAC PWMs and DeepBind models,

respectively. This confirmed the capability of DRAF models for prediction of TFBSs of TFs.

It should be noted that our testing methodology involved validating models on background data composed of the entire chromosome 21 (excluding data used in training and those overlapped with ChIP-seq peaks when specific cell type was used). Such chromosome-wide testing is useful to assess in an unbiased way TFBS prediction models, as it does not involve creation of artificial background sequences. For example, the background sequences used by DeepBind models were the top 500 even-numbered ChIP-seq peaks that were randomly shuffled.

Although our model demonstrates the lowest false positive rates among the tested models, the absolute number of false positive predictions is still noticeable. It is worth mentioning that our models predict binding sites in a non-cell-specific manner, thus some of them may not be available for binding in a given cell type and therefore do not intersect with ChIP-seq peaks for that cell type. Chromatin structure interferes with the TF binding via modifications of histones (Consortium, 2012; Wang, et al., 2012) and to a less extent via DNA methylation (Medvedeva, et al., 2014). Commonly used computational strategies to compensate for the specifics of the chromatin structure would be a subsequent filtering of unavailable binding sites using histone modification data (e.g. (Ramsey, et al., 2010)) or DNase I hypersensitivity regions (DNase-seq) for a cell type of interest (Alam, et al., 2014; Boyle, et al., 2008).

4.6 Conclusion

In this work, we modeled TF-TFBS interactions using properties extracted from sequences of DNA binding domains of TFs and TFs' DNA binding sites using a new methodology (DRAF). A random forests DRAF model was built for all TFs sharing a common TFBS length. That is, for all 250 TFBS models obtained from the HOCOMOCO database we developed 14 DRAF models representing 14 distinct TFBS lengths. The average prediction accuracy of 99.16%, which, to the best of our knowledge is the highest of those currently reported, clearly demonstrates the advantages of our methodology for TFBS prediction.

Using our method we reduced the number of required models approximately 18, 54 and 50 times compared to the HOCOMOCO, TRANSFAC and DeepBind models, respectively. Yet, we demonstrated higher specificity of DRAF models than with the HOCOMOCO, TRANSFAC and DeepBind models for >80% of ChIP-seq datasets for all sensitivity levels.

Chapter 5

Conclusions and Future Work

5.1 Introducing remarks

In this study, we described proteins by feature vectors derived from converting sequence properties and physicochemical properties of amino acids from specific protein regions into numerical values. In specific applications we sub-selected features of so described proteins in order make the description more suitable for the problems analyzed. Such description allowed for very discriminating characterization of proteins that in some cases may be related to protein functions. We used this approach in two different applications. In the first application, we aimed to characterize AMPs using their physicochemical properties and sub-select important properties that discriminate between AMP families. In the other application, we use the same method to extract TF properties that help in identifying TF affinity to its binding sites.

5.2 Comments on the Developed Methods

It has been a challenge to represent proteins by a suitable numerical form that is also efficient for the types of problems we analyzed. Various number of the previous studies represented proteins using their amino acid composition properties. However, such representations do not include the physicochemical features that capture different aspects of the amino acid properties such as their structure and hydrophobicity. Such

properties are important and highly correlated with some characteristics and functions that proteins perform. Consequently, in the methods that we developed in this study we used physicochemical properties of amino acids to represent the proteins numerically.

Also, not all amino acids within the protein sequence are of the same importance in characterizing protein families and in predicting protein functions. Consequently, the functional regions within the protein sequence are likely more important than other regions. These regions vary depending on the specific function of the protein we intend to study. For example, the N-terminal and C-terminal regions are very important to model the protein cellular location and in characterizing AMPs. The protein DNA binding domains regions are very important to control the TF affinity to its DNA binding sites. Our numerical representation of the proteins focuses on these functional regions during the feature representation for the proteins under study.

In addition, it is very important to select the subset of amino acid properties that are more relevant to the prediction of protein function because some properties are more important than others. For this purpose, our method relies on feature selection after representing proteins by feature vectors. The feature selection was aimed to reduce the total number of features to a smaller and likely more relevant set of features for the problems we studied. For this reason, in Chapter 3, we performed a pre-filtering step in our representation to maintain in the protein numerical representation only those properties that are conserved with the protein families (restricted properties). Such properties are more likely to characterize well the protein families under study.

We did not use all possible protein properties such as protein 3D structure properties in our representation of the proteins. Such properties are very important to understand the overall protein characteristics and function, but we did not include them as they are missing for a large number of proteins. Consequently, we resorted using physicochemical properties of individual amino acids instead.

5.3 Contribution summary

The usage of the protein 3D structure properties is very important to understand the protein characteristics and predict its function. However, as mentioned earlier, this information is available only for very small percentage of known proteins. Consequently, in such a situation we resorted to the use of physicochemical properties of amino acids in the protein functional regions.

We developed a novel method using computational representation of proteins based on characteristics of different protein regions (N-terminal, M-region and C-terminal) and combined these with the compositional and physicochemical properties of protein amino acids sequences. We show that this description provides important biological insight about characterization of the protein functional groups. Using feature selection techniques, we identified key properties of proteins that allow for very accurate characterization of different protein families. We demonstrated efficiency of our method in application to a number of AMP families.

We developed another novel method that uses a combination of amino acids physicochemical properties of DNA binding domains of TFs and their TFBS properties to develop machine learning models for predicting TFBSs. Feature selection is used to identify the most relevant characteristics of the amino acids for such modeling. In addition to reducing the number of required models to only 14 for several hundred TFs, the final prediction accuracy of our models appears dramatically better than with other methods.

Overall, we did show how to efficiently utilize properties of proteins in deriving more accurate solutions for two important problems of computational biology and bioinformatics.

5.4 Future Research

The research performed in this study can be extended using the following strategies:

- a. **Include protein structural properties:** While we have shown that physicochemical properties of amino acids when combined with sequence properties are very useful for characterizing proteins and inferring their functions, accurate prediction of protein functions require information from the 3D protein structure. Consequently, in addition to the physicochemical properties of the amino acids and their sequences, it will be beneficial to include the 3D structure properties of the proteins into the models we developed for the two problems we analyzed.

- b. **Correlation between selected properties:** By applying our method on the AMPs, we selected properties that characterize different AMP families. These properties can help in the *in-silico* design of novel AMPs. However, a very useful extension to our work and ahead of the *in-silico* design of novel AMPs, is the study of the correlation between different properties in order to determine which properties may be redundant in relation to the problem in question and to rank these properties accordingly. This will be useful information during the design process of novel AMPs.
- c. **Prediction of DNA binding sites for new TFs:** In this study we build models that learn from known binding sites of TFs and then can be used to detect new binding sites for these TFs. A normal extension to this work is to predict binding sites for new TFs that do not have known binding sites yet. The DRAF modeling is generic, so it can be used to detect binding sites for TFs not used to train the model. To achieve this, we may look for the family class of the DNA binding domain of the new TF and search for the most homologous DNA binding domain in the TFs used in the training. Then, we can use the corresponding DRAF model to scan for novel TFBSs for this new TF. The motivation is as follows. TFs usually have DNA binding domains and these domains that belong to different classes. If two TFs have their DNA binding domains in the same class and are similar in their sequence, then we assume they are more likely binding to the same/similar TFBSs.

BIBLIOGRAPHY/REFERENCES

- Alam, T., *et al.* Promoter analysis reveals globally differential regulation of human long non-coding RNA and protein-coding genes. *PLoS One* 2014;9(10):e109443.
- Alamanova, D., Stegmaier, P. and Kel, A. Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies. *BMC Bioinformatics* 2010;11:225.
- Alipanahi, B., *et al.* Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;33(8):831-838.
- Altman, N.S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am Stat* 1992;46(3):175-185.
- Altschul, S.F., *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 1997;25(17):3389-3402.
- Andreu, D., *et al.* Solid-phase synthesis of cecropin A and related peptides. *Proceedings of the National Academy of Sciences of the United States of America* 1983;80(21):6475-6479.
- Andreu, D. and Rivas, L. Animal antimicrobial peptides: an overview. *Biopolymers* 1998;47(6):415-433.
- Apweiler, R., *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic acids research* 2001;29(1):37-40.
- Ashburner, M., *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 2000;25(1):25-29.
- Bauer, A.L., *et al.* Using sequence-specific chemical and structural properties of DNA to predict transcription factor binding sites. *PLoS computational biology* 2010;6(11):e1001007.
- Beghin, F. and Dirkx, J. [Proceedings: A simple statistical method to predict protein conformations]. *Archives internationales de physiologie et de biochimie* 1975;83(1):167-168.
- Ben-Gal, I., *et al.* Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics* 2005;21(11):2657-2666.
- Berg, J.M., *et al.* Biochemistry. New York: W.H. Freeman; 2002.
- Berman, H., Henrick, K. and Nakamura, H. Announcing the worldwide Protein Data Bank. *Nature structural biology* 2003;10(12):980.
- Berman, H.M., *et al.* The Protein Data Bank. *Nucleic acids research* 2000;28(1):235-242.
- Bhasin, M., *et al.* Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett.* 2005;579(20):4302-4308.
- Bi, Y., *et al.* Tree-based position weight matrix approach to model transcription factor binding site profiles. *PLoS One* 2011;6(9):e24210.
- Blancafort, P., Segal, D.J. and Barbas, C.F., 3rd. Designing transcription factor architectures for drug discovery. *Mol Pharmacol* 2004;66(6):1361-1371.

- Boyle, A.P., *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* 2008;132(2):311-322.
- Brahmachary, M., *et al.* ANTIMIC: a database of antimicrobial sequences. *Nucleic acids research* 2004;32(Database issue):D586-589.
- Breiman, L. Random Forests. *Machine Learning* 2001;45(1):5-32.
- Brogden, K.A. Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? *Nature reviews. Microbiology* 2005;3(3):238-250.
- Bromberg, Y., *et al.* New in protein structure and function annotation: hotspots, single nucleotide polymorphisms and the 'Deep Web'. *Current opinion in drug discovery & development* 2009;12(3):408-419.
- Bulyk, M.L. Computational prediction of transcription-factor binding site locations. *Genome Biol* 2003;5(1):201.
- Bulyk, M.L., *et al.* Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nat Biotechnol* 1999;17(6):573-577.
- Burges, C.J. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery* 1998;2(2):121-167.
- Cai, Y., *et al.* A novel computational approach to predict transcription factor DNA binding preference. *Journal of proteome research* 2009;8(2):999-1003.
- Cai, Y., *et al.* A novel computational approach to predict transcription factor DNA binding preference. *Journal of proteome research* 2009;8(2):999-1003.
- Cai, Y., *et al.* A novel sequence-based method of predicting protein DNA-binding residues, using a machine learning approach. *Molecules and cells* 2010;30(2):99-105.
- Calinon, S. Robot Programming by Demonstration: A Probabilistic Approach. EPFL/CRC Press; 2009.
- Calinon, S., Guenter, F. and Billard, A. On learning, representing, and generalizing a task in a humanoid robot. *IEEE Trans Syst Man Cybern B Cybern* 2007;37(2):286-298.
- Cao, W., *et al.* Expression and purification of antimicrobial peptide adenoregulin with C-amidated terminus in Escherichia coli. *Protein expression and purification* 2005;40(2):404-410.
- Carninci, P., *et al.* The transcriptional landscape of the mammalian genome. *Science* 2005;309(5740):1559-1563.
- Cawley, S., *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 2004;116(4):499-509.
- Chakraborty, U.K. Advances in differential evolution. Berlin: Springer Verlag; 2008.
- Chan, Y.R., *et al.* Anti-microbial activity and cell binding are controlled by sequence determinants in the anti-microbial peptide PR-39. *The Journal of investigative dermatology* 2001;116(2):230-235.
- Chandrababu, K.B., Ho, B. and Yang, D. Structure, dynamics, and activity of an all-cysteine mutated human beta defensin-3 peptide analogue. *Biochemistry* 2009;48(26):6052-6061.
- Chang, C.-C. and Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2011;2(3):1-27.

- Charton, M. and Charton, B.I. The dependence of the Chou-Fasman parameters on amino acid side chain structure. *Journal of theoretical biology* 1983;102(1):121-134.
- Chen, C.Y., *et al.* Predicting target DNA sequences of DNA-binding proteins based on unbound structures. *PLoS One* 2012;7(2):e30446.
- Chen, L., *et al.* How the antimicrobial peptides kill bacteria: Computational physics insights. *Communications in Computational Physics* 2012;11(3):709.
- Chen, P., *et al.* Structure-activity study of the lantibiotic mutacin II from *Streptococcus mutans* T8 by a gene replacement strategy. *Applied and environmental microbiology* 1998;64(7):2335-2340.
- Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 2001;43(3):246-255.
- Clark, W.T. and Radivojac, P. Analysis of protein function and its prediction from amino acid sequence. *Proteins* 2011;79(7):2086-2096.
- Conlon, J.M., *et al.* Peptides with potent cytolytic activity from the skin secretions of the North American leopard frogs, *Lithobates blairi* and *Lithobates yavapaiensis*. *Toxicon : official journal of the International Society on Toxinology* 2009;53(7-8):699-705.
- Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489(7414):57-74.
- Consortium, F., *et al.* A promoter-level mammalian expression atlas. *Nature* 2014;507(7493):462-470.
- Correa, P.G. and Oguiura, N. Phylogenetic analysis of beta-defensin-like genes of Bothrops, Crotalus and Lachesis snakes. *Toxicon : official journal of the International Society on Toxinology* 2013;69:65-74.
- Cortes, C. and Vapnik, V. Support-Vector Networks. 1995(3):273-297.
- Cortes, C. and Vapnik, V. Support-Vector Networks. *Mach Learn* 1995;20(3):273-297.
- Crooks, G.E., *et al.* WebLogo: a sequence logo generator. *Genome Res* 2004;14(6):1188-1190.
- Das, M.K. and Dai, H.K. A survey of DNA motif finding algorithms. *BMC Bioinformatics* 2007;8 Suppl 7:S21.
- Deng, L. and Yu, D. Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing* 2014;7(3-4):197-387.
- Dischinger, J., Basi Chipalu, S. and Bierbaum, G. Lantibiotics: promising candidates for future applications in health care. *International journal of medical microbiology : IJMM* 2014;304(1):51-62.
- Du, P. and Li, Y. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. *BMC bioinformatics* 2006;7:518.
- Ellrott, K., *et al.* Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics* 2002;18 Suppl 2:S100-109.
- Elnitski, L., *et al.* Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res.* 2006;16(12):1455-1464.
- Emanuelsson, O., *et al.* Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* 2000;300(4):1005-1016.

- Endres, R.G., Schulthess, T.C. and Wingreen, N.S. Toward an atomistic model for predicting transcription-factor binding sites. *Proteins* 2004;57(2):262-268.
- Epand, R.M. and Vogel, H.J. Diversity of antimicrobial peptides and their mechanisms of action. *Biochimica et biophysica acta* 1999;1462(1-2):11-28.
- Fink, J., *et al.* The chemical synthesis of cecropin D and an analog with enhanced antibacterial activity. *The Journal of biological chemistry* 1989;264(11):6260-6267.
- Finn, R.D., *et al.* Pfam: the protein families database. *Nucleic Acids Res* 2014;42(Database issue):D222-230.
- Fjell, C.D., *et al.* Designing antimicrobial peptides: form follows function. *Nature reviews. Drug discovery* 2012;11(1):37-51.
- Freder, V., Ho, B. and Ding, J.L. De novo design of potent antimicrobial peptides. *Antimicrobial agents and chemotherapy* 2004;48(9):3349-3357.
- Frey, B.J. and Dueck, D. Clustering by passing messages between data points. *Science* 2007;315(5814):972-976.
- Friedberg, I. Automated protein function prediction--the genomic challenge. *Briefings in bioinformatics* 2006;7(3):225-242.
- Friedman, N., Geiger, D. and Goldszmidt, M. Bayesian network classifiers. *Mach Learn* 1997;29(2-3):131-163.
- Gabdoulline, R., *et al.* 3DTF: a web server for predicting transcription factor PWMs using 3D structure-based energy calculations. *Nucleic Acids Res* 2012;40(Web Server issue):W180-185.
- Ganz, T. The role of antimicrobial peptides in innate immunity. *Integrative and comparative biology* 2003;43(2):300-304.
- Garcia, I., Fall, Y. and Gomez, G. Review of synthesis, biological assay, and QSAR studies of HMGR inhibitors. *Current topics in medicinal chemistry* 2012;12(8):895-919.
- Geertz, M. and Maerkl, S.J. Experimental strategies for studying transcription factor-DNA binding specificities. *Brief Funct Genomics* 2010;9(5-6):362-373.
- George, R.A. and Heringa, J. An analysis of protein domain linkers: their classification and role in protein folding. *Protein engineering* 2002;15(11):871-879.
- Gershenson, N.I., Stormo, G.D. and Ioshikhes, I.P. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. *Nucleic Acids Res* 2005;33(7):2290-2301.
- Gerstein, M.B., *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* 2012;489(7414):91-100.
- Golub, G.H. and Van Loan, C.F. Matrix computations. Baltimore: Johns Hopkins University Press; 1983.
- Gordon, Y.J., Romanowski, E.G. and McDermott, A.M. A review of antimicrobial peptides and their therapeutic potential as anti-infective drugs. *Current eye research* 2005;30(7):505-515.
- Grant, C.E., Bailey, T.L. and Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* 2011;27(7):1017-1018.
- Gromiha, M.M. Chapter 3 - Protein Structure Analysis. In: Gromiha, M.M., editor, *Protein Bioinformatics*. Singapore: Academic Press; 2010. p. 63-105.

- Guralp, S.A., *et al.* From design to screening: a new antimicrobial peptide discovery pipeline. *PloS one* 2013;8(3):e59305.
- Guyon, I. and Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 2003;3:1157-1182.
- Hancock, R.E. and Diamond, G. The role of cationic antimicrobial peptides in innate host defences. *Trends in microbiology* 2000;8(9):402-410.
- Hancock, R.E. and Lehrer, R. Cationic peptides: a new source of antibiotics. *Trends in biotechnology* 1998;16(2):82-88.
- Hancock, R.E. and Sahl, H.G. Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. *Nature biotechnology* 2006;24(12):1551-1557.
- Hancock, R.E. and Scott, M.G. The role of antimicrobial peptides in animal defenses. *Proceedings of the National Academy of Sciences of the United States of America* 2000;97(16):8856-8861.
- Hayes, M., *et al.* Casein-derived antimicrobial peptides generated by *Lactobacillus acidophilus* DPC6026. *Applied and environmental microbiology* 2006;72(3):2260-2264.
- Hegyí, H. and Gerstein, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *Journal of Molecular Biology* 1999;288(1):147-164.
- Hensen, U., *et al.* Exploring protein dynamics space: the dynasome as the missing link between protein structure and function. *PloS one* 2012;7(5):e33931.
- Herrmann, A., *et al.* Key role of glutamic acid for the cytotoxic activity of the cyclotide cycloviolacin O2. *Cellular and molecular life sciences : CMLS* 2006;63(2):235-245.
- Ho, T.K. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence* 1998;20(8):832-844.
- Hoglund, A. and Kohlbacher, O. From sequence to structure and back again: approaches for predicting protein-DNA binding. *Proteome Sci* 2004;2(1):3.
- Holland, J.H. Adaptation in natural and artificial systems. MIT Press; 1992.
- Horak, C.E. and Snyder, M. ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol* 2002;350:469-483.
- Hvidsten, T.R., *et al.* A comprehensive analysis of the structure-function relationship in proteins based on local structure similarity. *PloS one* 2009;4(7):e6266.
- Jabrane, A., *et al.* Characterization of serracin P, a phage-tail-like bacteriocin, and its activity against *Erwinia amylovora*, the fire blight pathogen. *Applied and environmental microbiology* 2002;68(11):5704-5710.
- Jensen, H., Hamill, P. and Hancock, R.E. Peptide antimicrobial agents. *Clinical microbiology reviews* 2006;19(3):491-511.
- Jolliffe, I.T. Principal component analysis. New York: Springer; 2002.
- Juretic, D., *et al.* Knowledge-based computational methods for identifying or designing novel, non-homologous antimicrobial peptides. *European biophysics journal : EBJ* 2011;40(4):371-385.
- Kaiser, V. and Diamond, G. Expression of mammalian defensin genes. *Journal of leukocyte biology* 2000;68(6):779-784.

- Kaplan, T., Friedman, N. and Margalit, H. Ab initio prediction of transcription factor targets using structural knowledge. *PLoS computational biology* 2005;1(1):e1.
- Kawashima, S. and Kanehisa, M. AAindex: amino acid index database. *Nucleic acids research* 2000;28(1):374.
- Kawashima, S., *et al.* AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 2008;36(Database issue):D202-205.
- Kennedy, J. Particle Swarm Optimization. In: Sammut, C. and Webb, G., editors, *Encyclopedia of Machine Learning*. Springer US; 2010. p. 760-766.
- Kennedy, J. and Eberhart, R. Particle swarm optimization. *1995 IEEE International Conference on Neural Networks Proceedings, Vols 1-6* 1995:1942-1948.
- Khamis, A.M., *et al.* Distinct profiling of antimicrobial peptide families. *Bioinformatics* 2015;31(6):849-856.
- Kidera, A., *et al.* Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J Protein Chem* 1985;4(1):23-55.
- Koehbach, J., *et al.* Oxytocin plant cyclotides as templates for peptide G protein-coupled receptor ligand design. *Proceedings of the National Academy of Sciences of the United States of America* 2013;110(52):21183-21188.
- Kohavi, R. Feature subset selection as search with probabilistic estimates. In, *Proceedings AAAI Fall Symposium on Relevance*. New Orleans, LA; 1994. p. 122–126.
- Kono, H. and Sarai, A. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins* 1999;35(1):114-131.
- Kulakovskiy, I.V., *et al.* HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res* 2013;41(Database issue):D195-202.
- Landreh, M., *et al.* New developments in protein structure-function analysis by MS and use of hydrogen-deuterium exchange microfluidics. *The FEBS journal* 2011;278(20):3815-3821.
- Langham, A.A., *et al.* Correlation between simulated physicochemical properties and hemolysis of protegrin-like antimicrobial peptides: predicting experimental toxicity. *Peptides* 2008;29(7):1085-1093.
- Larkin, M.A., *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* 2007;23(21):2947-2948.
- Lata, S., Mishra, N.K. and Raghava, G.P. AntiBP2: improved version of antibacterial peptide prediction. *BMC bioinformatics* 2010;11 Suppl 1:S19.
- Lata, S., Sharma, B.K. and Raghava, G.P. Analysis and prediction of antibacterial peptides. *BMC bioinformatics* 2007;8:263.
- Lee, D., Redfern, O. and Orengo, C. Predicting protein function from sequence and structure. *Nature reviews. Molecular cell biology* 2007;8(12):995-1005.
- Lee, J.Y., *et al.* Antibacterial peptides from pig intestine: isolation of a mammalian cecropin. *Proceedings of the National Academy of Sciences of the United States of America* 1989;86(23):9159-9162.
- Lehrer, R.I. Multispecific myeloid defensins. *Current opinion in hematology* 2007;14(1):16-21.

- Lehrer, R.I. and Ganz, T. Antimicrobial peptides in mammalian and insect host defence. *Current opinion in immunology* 1999;11(1):23-27.
- Lequin, O., *et al.* Dermaseptin S9, an alpha-helical antimicrobial peptide with a hydrophobic core and cationic termini. *Biochemistry* 2006;45(2):468-480.
- Leszczynski, J. Handbook of computational chemistry. Dordrecht ; New York: Springer; 2012.
- Letunic, I., Doerks, T. and Bork, P. SMART: recent updates, new developments and status in 2015. *Nucleic acids research* 2015;43(Database issue):D257-260.
- Levitt, M. Conformational preferences of amino acids in globular proteins. *Biochemistry* 1978;17(20):4277-4285.
- Li, W. and Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;22(13):1658-1659.
- Lieberman, M., Marks, A.D. and Peet, A. Marks' basic medical biochemistry : a clinical approach. Philadelphia: Wolter Kluwer Health/Lippincott Williams & Wilkins; 2013.
- Liu, B., *et al.* Using amino acid physicochemical distance transformation for fast protein remote homology detection. *PLoS one* 2012;7(9):e46633.
- Liu, L.A. and Bader, J.S. Ab initio prediction of transcription factor binding sites. *Pac. Symp. Biocomput.* 2007:484-495.
- Liu, Z., *et al.* Structure-based prediction of transcription factor binding sites using a protein-DNA docking approach. *Proteins* 2008;72(4):1114-1124.
- Lohans, C.T. and Vederas, J.C. Structural characterization of thioether-bridged bacteriocins. *The Journal of antibiotics* 2014;67(1):23-30.
- Luscombe, N.M. and Thornton, J.M. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J Mol Biol* 2002;320(5):991-1009.
- Maccari, G., *et al.* Antimicrobial peptides design by evolutionary multiobjective optimization. *PLoS computational biology* 2013;9(9):e1003212.
- Maerkl, S.J. and Quake, S.R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 2007;315(5809):233-237.
- Marchler-Bauer, A., *et al.* CDD: NCBI's conserved domain database. *Nucleic acids research* 2015;43(Database issue):D222-226.
- Marcos, J.F., *et al.* Identification and rational design of novel antimicrobial peptides for plant protection. *Annual review of phytopathology* 2008;46:273-301.
- Mathelier, A., *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2015.
- Mathelier, A. and Wasserman, W.W. The next generation of transcription factor binding site prediction. *PLoS Comput Biol* 2013;9(9):e1003214.
- MATLAB. version 8.0.0.783 (R2012b). Natick, Massachusetts, United States: The MathWorks Inc.; 2012.
- Matsuda, S., *et al.* A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein science : a publication of the Protein Society* 2005;14(11):2804-2813.
- Matys, V., *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research* 2006;34(Database issue):D108-110.

- McKee, T. and McKee, J.R. *Biochemistry : the molecular basis of life*. Oxford ; New York: Oxford University Press; 2012.
- Medvedeva, Y.A., *et al.* Effects of cytosine methylation on transcription factor binding sites. *BMC Genomics* 2014;15:119.
- Meysman, P., *et al.* Use of structural DNA properties for the prediction of transcription-factor binding sites in *Escherichia coli*. *Nucleic Acids Research* 2011;39(2):e6.
- Midorikawa, K., *et al.* Staphylococcus aureus susceptibility to innate antimicrobial peptides, beta-defensins and CAP18, expressed by human keratinocytes. *Infection and immunity* 2003;71(7):3730-3739.
- Minervini, F., *et al.* Angiotensin I-converting-enzyme-inhibitory and antibacterial peptides from *Lactobacillus helveticus* PR4 proteinase-hydrolyzed caseins of milk from six species. *Applied and environmental microbiology* 2003;69(9):5297-5305.
- Mitchell, T.M. *Machine Learning*. New York: McGraw-Hill; 1997.
- Moll, G.N., *et al.* Comparison of the membrane interaction and permeabilization by the designed peptide Ac-MB21-NH2 and truncated dermaseptin S3. *Biochemistry* 2000;39(39):11907-11912.
- Mordelet, F., *et al.* Stability selection for regression-based models of transcription factor-DNA binding specificity. *Bioinformatics* 2013;29(13):i117-125.
- Nakashima, H., Nishikawa, K. and Ooi, T. Distinct character in hydrophobicity of amino acid compositions of mitochondrial proteins. *Proteins* 1990;8(2):173-178.
- Nusslein, K., *et al.* Broad-spectrum antibacterial activity by a novel abiogenic peptide mimic. *Microbiology* 2006;152(Pt 7):1913-1918.
- Oppgaard, C., *et al.* Mutational analysis of putative helix-helix interacting GxxxG-motifs and tryptophan residues in the two-peptide bacteriocin lactococcin G. *Biochemistry* 2008;47(18):5242-5249.
- Pal, T., *et al.* Brevinin-1BYa: a naturally occurring peptide from frog skin with broad-spectrum antibacterial and antifungal properties. *International journal of antimicrobial agents* 2006;27(6):525-529.
- Palau, J., Argos, P. and Puigdomenech, P. Protein secondary structure. Studies on the limits of prediction accuracy. *International journal of peptide and protein research* 1982;19(4):394-401.
- Palliser, C.C. and Parry, D.A.D. Quantitative comparison of the ability of hydropathy scales to recognize surface β -strands in proteins. *Proteins: Structure, Function, and Bioinformatics* 2001;42(2):243-255.
- Park, P.J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009;10(10):669-680.
- Pascual-Garcia, A., *et al.* Quantifying the evolutionary divergence of protein structures: the role of function change and function conservation. *Proteins* 2010;78(1):181-196.
- Pasupuleti, M., Schmidtchen, A. and Malmsten, M. Antimicrobial peptides: key components of the innate immune system. *Critical reviews in biotechnology* 2012;32(2):143-171.

- Peng, H., Long, F. and Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE transactions on pattern analysis and machine intelligence* 2005;27(8):1226-1238.
- Peters, B.M., Shirliff, M.E. and Jabra-Rizk, M.A. Antimicrobial peptides: primeval molecules or future drugs? *PLoS pathogens* 2010;6(10):e1001067.
- Porto, W., Fernandes, F. and Franco, O. An SVM Model Based on Physicochemical Properties to Predict Antimicrobial Activity from Protein Sequences with Cysteine Knot Motifs. In: Ferreira, C., Miyano, S. and Stadler, P., editors, *Advances in Bioinformatics and Computational Biology*. Springer Berlin Heidelberg; 2010. p. 59-62.
- Pushpanathan, M., Gunasekaran, P. and Rajendhran, J. Antimicrobial peptides: versatile biological properties. *International journal of peptides* 2013;2013:675391.
- Qian, Z., Cai, Y.D. and Li, Y. A novel computational method to predict transcription factor DNA binding preference. *Biochem. Biophys. Res. Commun.* 2006;348(3):1034-1037.
- Qian, Z., *et al.* An approach to predict transcription factor DNA binding site specificity based upon gene and transcription factor functional categorization. *Bioinformatics* 2007;23(18):2449-2454.
- Qiu, P. Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. *Biochem Biophys Res Commun* 2003;309(3):495-501.
- Quinlan, J.R. Simplifying decision trees. *Int. J. Man-Mach. Stud.* 1987;27(3):221-234.
- Radek, K. and Gallo, R. Antimicrobial peptides: natural effectors of the innate immune system. *Seminars in immunopathology* 2007;29(1):27-43.
- Rajabi, M., *et al.* The conserved salt bridge in human alpha-defensin 5 is required for its precursor processing and proteolytic stability. *The Journal of biological chemistry* 2008;283(31):21509-21518.
- Ramsey, S.A., *et al.* Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics* 2010;26(17):2071-2075.
- Ravasi, T., *et al.* An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* 2010;140(5):744-752.
- Raven, P.H. and Johnson, G.B. *Biology*. Boston: McGraw-Hill; 2002.
- Reczko, M. and Hatzigerrorgiou, A. Prediction of the subcellular localization of eukaryotic proteins using sequence signals and composition. *Proteomics* 2004;4(6):1591-1596.
- Reddy, T.E., DeLisi, C. and Shakhnovich, B.E. Binding site graphs: a new graph theoretical framework for prediction of transcription factor binding sites. *PLoS computational biology* 2007;3(5):e90.
- Rockwell, N.C., *et al.* Precursor processing by kex2/furin proteases. *Chemical reviews* 2002;102(12):4525-4548.
- Rosengren, K.J., *et al.* Twists, knots, and rings in proteins. Structural definition of the cyclotide framework. *The Journal of biological chemistry* 2003;278(10):8606-8616.
- Roulet, E., *et al.* Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA. *In Silico Biol* 1998;1(1):21-28.
- Rumelhart, D.E., Hinton, G.E. and Williams, R.J. Learning Representations by Back-Propagating Errors. *Nature* 1986;323(6088):533-536.

- Saeys, Y., Inza, I. and Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007;23(19):2507-2517.
- Sang, Y. and Blecha, F. Antimicrobial peptides and bacteriocins: alternatives to traditional antibiotics. *Animal health research reviews / Conference of Research Workers in Animal Diseases* 2008;9(2):227-235.
- Sangar, V., *et al.* Quantitative sequence-function relationships in proteins based on gene ontology. *BMC bioinformatics* 2007;8:294.
- Satchell, D.P., *et al.* Interactions of mouse Paneth cell alpha-defensins and alpha-defensin precursors with membranes. Prosegment inhibition of peptide association with biomimetic membranes. *The Journal of biological chemistry* 2003;278(16):13838-13846.
- Saxena, S. and Gomber, C. Surmounting antimicrobial resistance in the Millennium Superbug: *Staphylococcus aureus*. *cent.eur.j.med* 2010;5(1):12-29.
- Schmidt, N.W., *et al.* Arginine in alpha-defensins: differential effects on bactericidal activity correspond to geometry of membrane curvature generation and peptide-lipid phase behavior. *The Journal of biological chemistry* 2012;287(26):21866-21872.
- Schölkopf, B. and Smola, A.J. Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT press; 2002.
- Shinnar, A.E., Butler, K.L. and Park, H.J. Cathelicidin family of antimicrobial peptides: proteolytic processing and protease resistance. *Bioorganic chemistry* 2003;31(6):425-436.
- Slootweg, J.C., *et al.* Synthesis, antimicrobial activity, and membrane permeabilizing properties of C-terminally modified nisin conjugates accessed by CuAAC. *Bioconjugate chemistry* 2013;24(12):2058-2066.
- Sra, S. and Inderjit, S.D. Generalized Nonnegative Matrix Approximations with Bregman Divergences. 2006:283--290.
- Steiner, D.F. The proprotein convertases. *Current opinion in chemical biology* 1998;2(1):31-39.
- Steiner, H., Andreu, D. and Merrifield, R.B. Binding and action of cecropin and cecropin analogues: antibacterial peptides from insects. *Biochimica et biophysica acta* 1988;939(2):260-266.
- Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* 2000;16(1):16-23.
- Stormo, G.D., *et al.* Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res* 1982;10(9):2997-3011.
- Sundararajan, V.S., *et al.* DAMPD: a manually curated antimicrobial peptide database. *Nucleic acids research* 2012;40(Database issue):D1108-1112.
- Tan, P.-N., Steinbach, M. and Kumar, V. Introduction to data mining. Boston: Pearson Addison Wesley; 2006.
- Tatusov, R.L., *et al.* The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research* 2000;28(1):33-36.
- Tenenbaum, J.B., de Silva, V. and Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. *Science* 2000;290(5500):2319-2323.

- Thomas, S., *et al.* CAMP: a useful resource for research on antimicrobial peptides. *Nucleic acids research* 2010;38(Database issue):D774-780.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* 1994;22(22):4673-4680.
- Torrent, M., *et al.* Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. *PloS one* 2011;6(2):e16968.
- UniProt, C. Activities at the Universal Protein Resource (UniProt). *Nucleic acids research* 2014;42(Database issue):D191-198.
- UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;43(Database issue):D204-212.
- Uzzell, T., *et al.* Hagfish cathelin-associated antimicrobial peptides and genes. In.; 2002.
- Valdar, W.S. Scoring residue conservation. *Proteins* 2002;48(2):227-241.
- Vaquerizas, J.M., *et al.* A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* 2009;10(4):252-263.
- Voet, D., Voet, J.G. and Pratt, C.W. Fundamentals of biochemistry. New York: Wiley; 1999.
- Vorontsov, I.E., Kulakovskiy, I.V. and Makeev, V.J. Jaccard index based similarity measure to compare transcription factor binding site models. *Algorithms Mol Biol* 2013;8(1):23.
- Waghu, F.H., *et al.* CAMP: Collection of sequences and structures of antimicrobial peptides. *Nucleic acids research* 2014;42(Database issue):D1154-1158.
- Wang, G., Li, X. and Wang, Z. APD2: the updated antimicrobial peptide database and its application in peptide design. *Nucleic acids research* 2009;37(Database issue):D933-937.
- Wang, J., *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* 2012;22(9):1798-1812.
- Wang, P., *et al.* Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. *PloS one* 2011;6(4):e18476.
- Wang, Z. and Wang, G. APD: the Antimicrobial Peptide Database. *Nucleic acids research* 2004;32(Database issue):D590-592.
- Weirauch, M.T., *et al.* Evaluation of methods for modeling transcription factor sequence specificity. *Nat Biotechnol* 2013;31(2):126-134.
- Wong, H., Bowie, J.H. and Carver, J.A. The solution structure and activity of caerin 1.1, an antimicrobial peptide from the Australian green tree frog, *Litoria splendida*. *European journal of biochemistry / FEBS* 1997;247(2):545-557.
- Yeaman, M.R. and Yount, N.Y. Mechanisms of antimicrobial peptide action and resistance. *Pharmacological reviews* 2003;55(1):27-55.
- Yuan, Y., Pei, J. and Lai, L. Binding site detection and druggability prediction of protein targets for structure-based drug design. *Current pharmaceutical design* 2013;19(12):2326-2333.
- Zanetti, M., Gennaro, R. and Romeo, D. Cathelicidins: a novel protein family with a common proregion and a variable C-terminal antimicrobial domain. *FEBS letters* 1995;374(1):1-5.

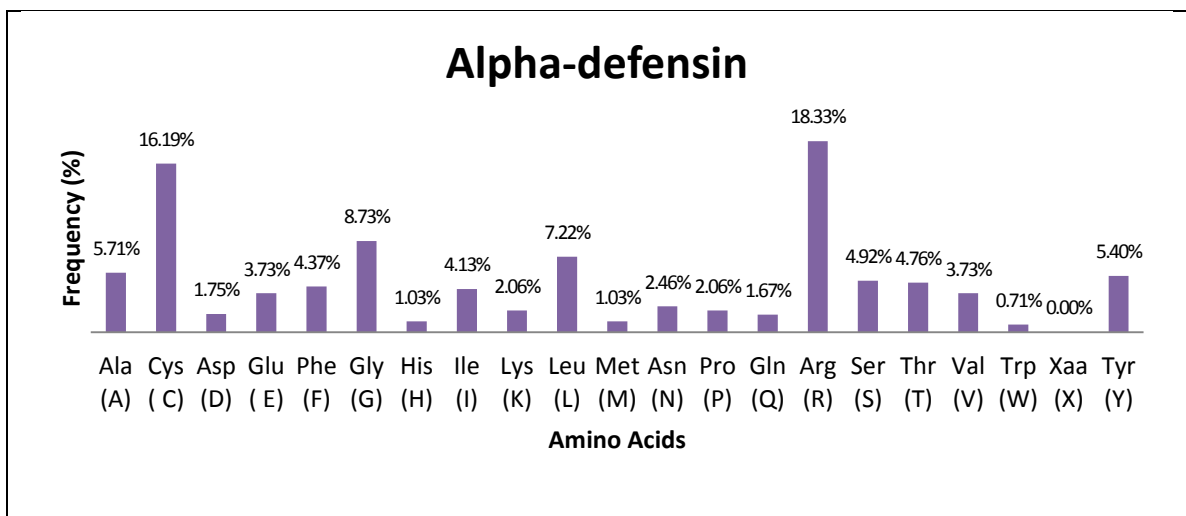
- Zangger, K., *et al.* Structures of the glycine-rich diastereomeric peptides bombinin H2 and H4. *Toxicon : official journal of the International Society on Toxinology* 2008;52(2):246-254.
- Zasloff, M. Antimicrobial peptides of multicellular organisms. *Nature* 2002;415(6870):389-395.
- Zheng, X., *et al.* Rational drug design: the search for Ras protein hydrolysis intermediate conformation inhibitors with both affinity and specificity. *Current pharmaceutical design* 2013;19(12):2246-2258.
- Zhou, T., *et al.* Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc Natl Acad Sci U S A* 2015;112(15):4654-4659.

APPENDICES

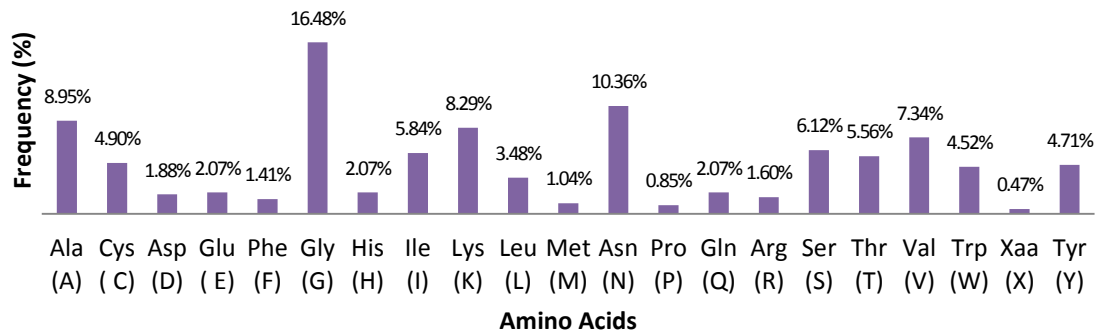
Appendix 1: Amino Acid Composition for 14 AMP Families

	Alpha-defensin	Bacteriocin	Beta-defensin	Bombinin	Cathelicidin	Cecropin	Cyclotide (Bracelet sub-family)	DEFL
Ala (A)	5.71%	8.95%	2.82%	9.96%	2.72%	16.42%	3.57%	6.95%
Cys (C)	16.19%	4.90%	12.37%	0.05%	1.85%	0.31%	19.78%	13.79%
Asp (D)	1.75%	1.88%	1.39%	1.36%	1.75%	1.86%	0.82%	2.52%
Glu (E)	3.73%	2.07%	1.39%	5.89%	2.43%	4.42%	4.12%	4.21%
Phe (F)	4.37%	1.41%	5.79%	1.36%	7.97%	3.80%	3.02%	4.98%
Gly (G)	8.73%	16.48%	10.09%	17.27%	7.87%	10.07%	8.24%	8.26%
His (H)	1.03%	2.07%	2.47%	1.51%	0.58%	0.54%	0.55%	3.01%
Ile (I)	4.13%	5.84%	5.49%	8.35%	7.09%	9.45%	8.79%	3.17%
Lys (K)	2.06%	8.29%	5.34%	10.75%	8.26%	11.39%	7.42%	8.21%
Leu (L)	7.22%	3.48%	8.61%	16.64%	8.07%	6.51%	3.02%	4.87%
Met (M)	1.03%	1.04%	1.78%	0.99%	0.29%	0.77%	0.27%	1.53%
Asn (N)	2.46%	10.36%	3.32%	2.61%	1.94%	3.49%	5.77%	6.62%
Pro (P)	2.06%	0.85%	5.94%	1.77%	17.20%	2.32%	5.49%	4.05%
Gln (Q)	1.67%	2.07%	3.22%	0.94%	2.53%	4.96%	0.00%	2.79%
Arg (R)	18.33%	1.60%	11.48%	4.54%	16.42%	6.35%	2.20%	5.36%
Ser (S)	4.92%	6.12%	6.28%	6.10%	2.43%	3.25%	9.62%	6.35%
Thr (T)	4.76%	5.56%	3.41%	3.23%	2.14%	3.18%	4.12%	5.20%
Val (V)	3.73%	7.34%	5.05%	4.96%	4.86%	8.52%	8.52%	3.94%
Trp (W)	0.71%	4.52%	1.63%	0.00%	1.94%	2.09%	1.10%	1.53%
Xaa (X)	0.00%	0.47%	0.00%	0.00%	0.00%	0.00%	0.00%	0.33%
Tyr (Y)	5.40%	4.71%	2.13%	1.72%	1.65%	0.31%	3.57%	2.35%
	FSAP (Brevinin sub-family)	FSAP (Caerin sub-family)	FSAP (Dermaseptin sub-family)	Invertebrate defensin (Type 1 sub-family)	Invertebrate defensin (Type 2 sub-family)	Type A lantibiotic	All Other AMP Families (114 Families/sub-families)	
Ala (A)	10.16%	9.74%	19.96%	11.05%	3.06%	3.00%	8.57%	
Cys (C)	5.94%	0.00%	0.00%	14.21%	14.59%	12.61%	3.49%	
Asp (D)	2.15%	0.75%	2.12%	3.38%	2.70%	0.60%	4.22%	

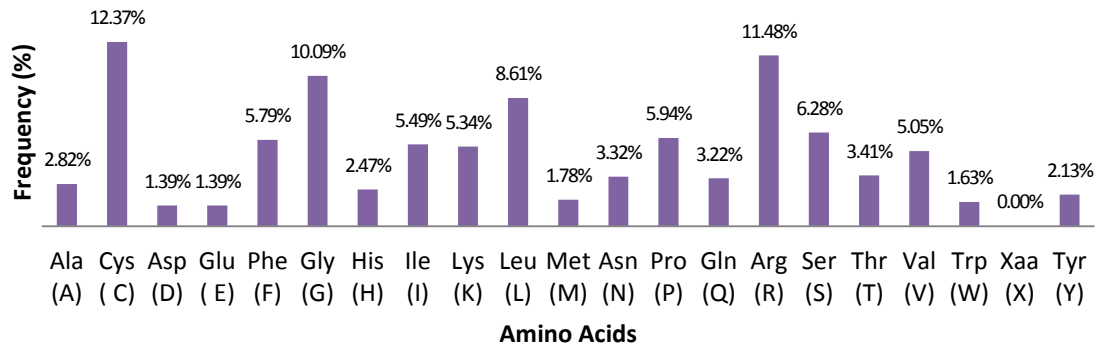
Glu (E)	1.13%	5.24%	2.01%	1.58%	1.80%	2.40%	4.43%	
Phe (F)	5.63%	2.25%	1.34%	2.59%	3.96%	4.20%	3.72%	
Gly (G)	11.03%	10.11%	12.93%	11.84%	15.32%	7.51%	8.81%	
His (H)	0.54%	6.37%	0.89%	3.95%	4.32%	3.00%	2.11%	
Ile (I)	7.65%	5.24%	4.35%	3.49%	3.06%	5.41%	4.97%	
Lys (K)	15.00%	10.11%	14.38%	5.86%	4.14%	8.71%	8.11%	
Leu (L)	14.52%	16.48%	12.37%	8.00%	4.14%	6.01%	7.75%	
Met (M)	1.36%	0.00%	2.68%	0.23%	1.44%	3.00%	1.82%	
Asn (N)	2.74%	0.37%	3.23%	5.75%	5.59%	4.80%	4.59%	
Pro (P)	3.28%	6.37%	0.33%	0.11%	3.42%	2.40%	4.95%	
Gln (Q)	1.33%	0.75%	3.68%	1.47%	2.70%	2.40%	3.82%	
Arg (R)	1.74%	0.37%	1.56%	6.99%	8.83%	1.80%	5.59%	
Ser (S)	5.40%	6.74%	4.01%	5.98%	5.41%	10.51%	6.83%	
Thr (T)	3.74%	0.00%	4.35%	4.28%	4.32%	12.61%	4.95%	
Val (V)	5.81%	18.35%	6.80%	6.54%	3.06%	5.71%	6.11%	
Trp (W)	0.26%	0.75%	2.90%	0.79%	2.16%	2.40%	1.53%	
Xaa (X)	0.00%	0.00%	0.00%	0.00%	0.54%	0.00%	0.07%	
Tyr (Y)	0.59%	0.00%	0.11%	1.92%	5.41%	0.90%	3.55%	



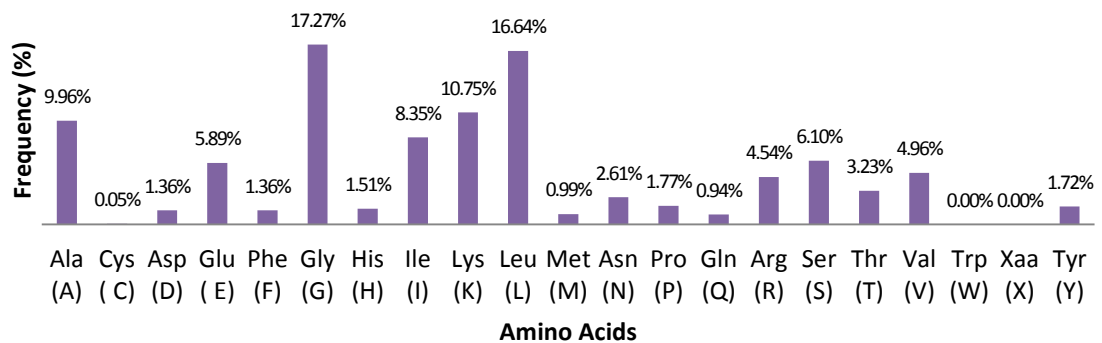
Bacteriocin



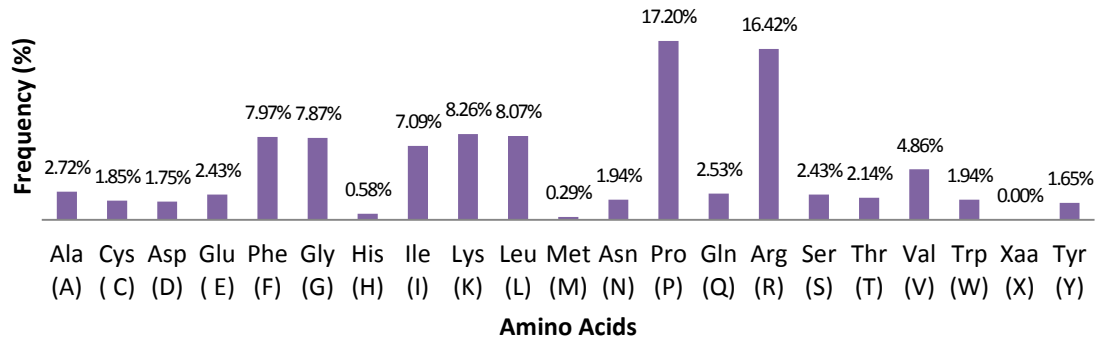
Beta-defensin



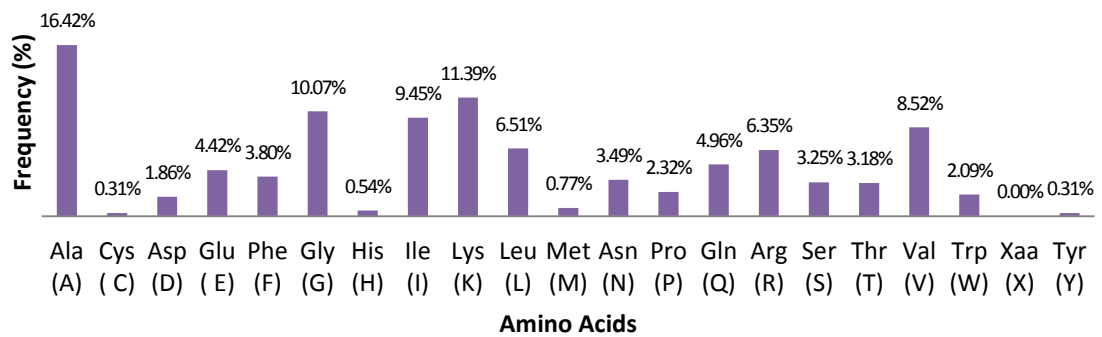
Bombinin



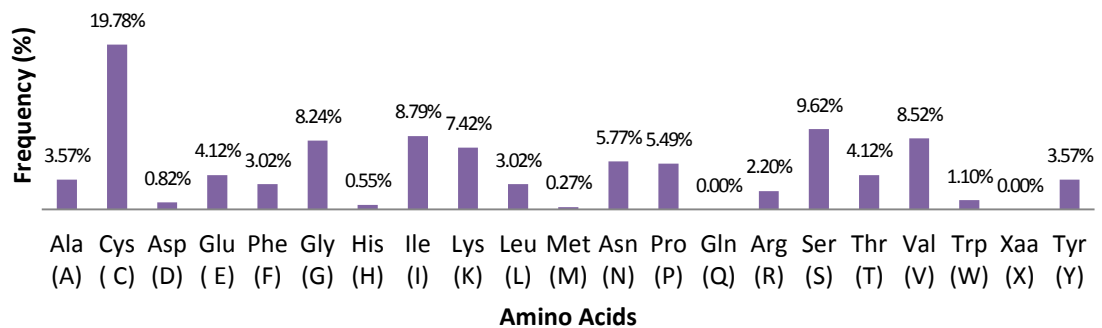
Cathelicidin

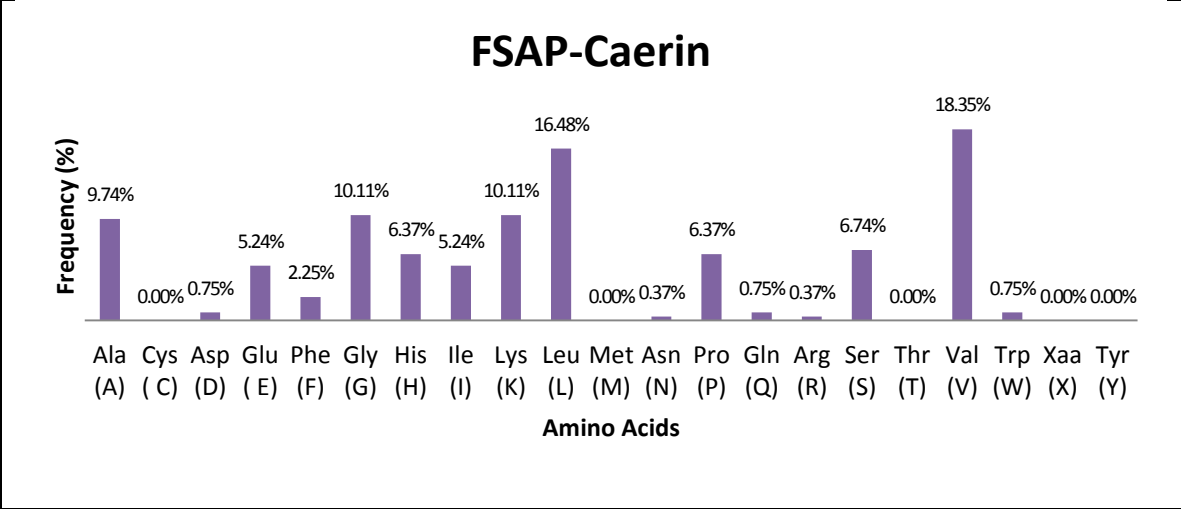
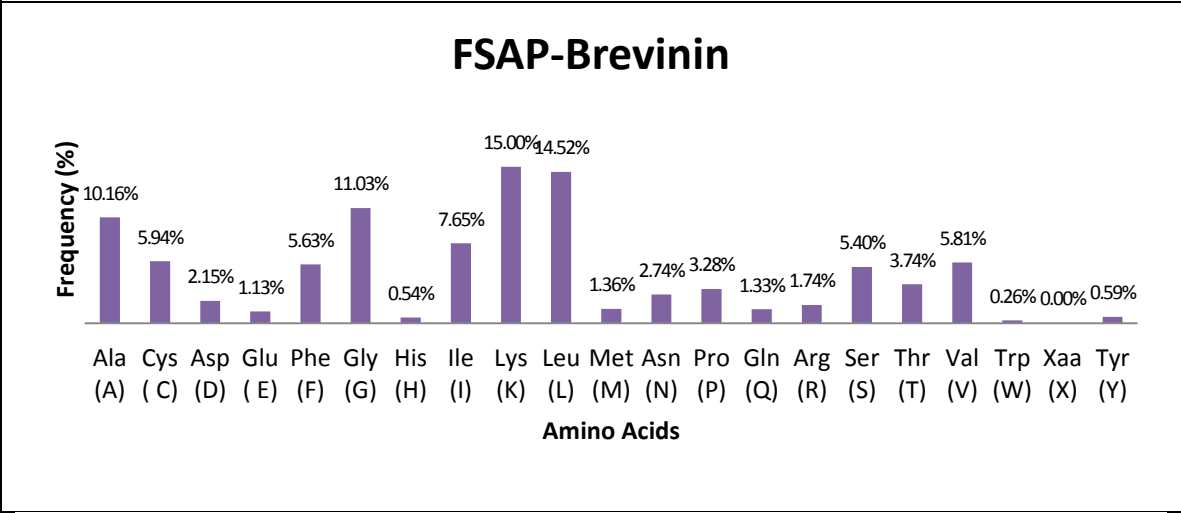
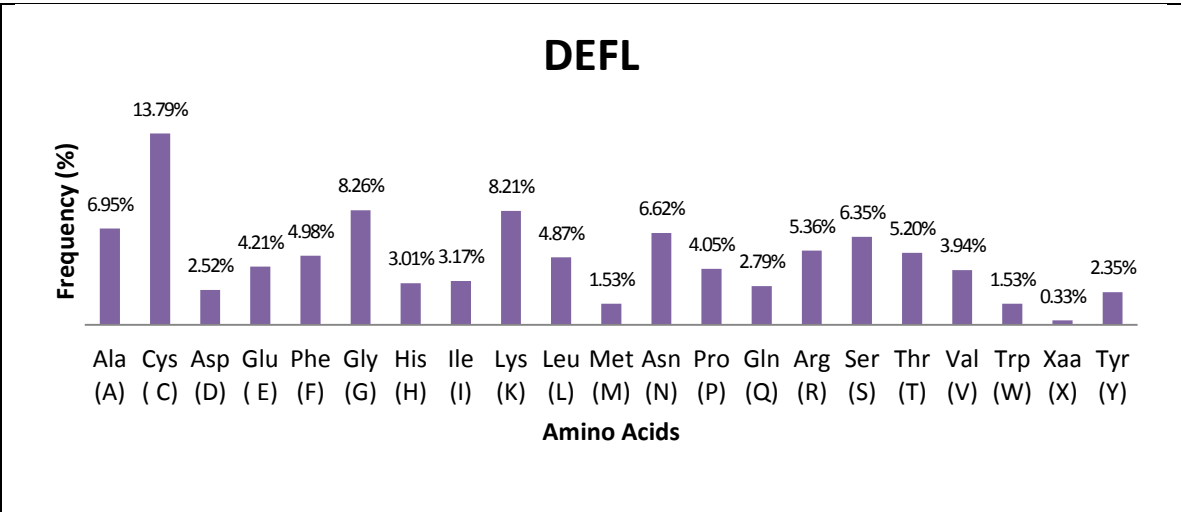


Cecropin

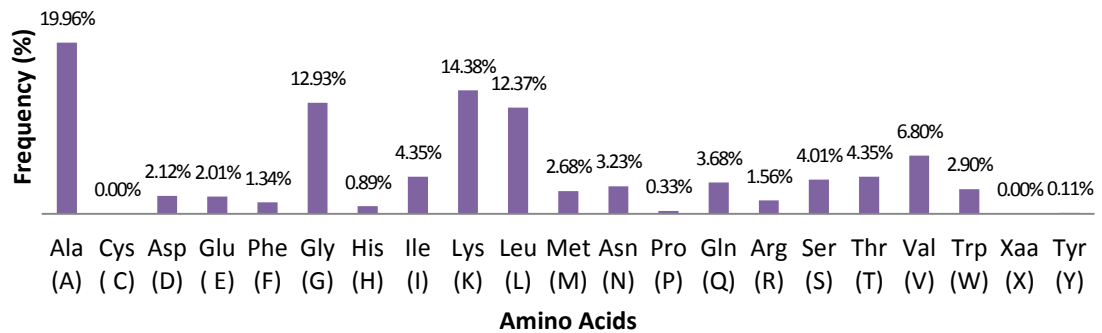


Cyclotide-Bracelet

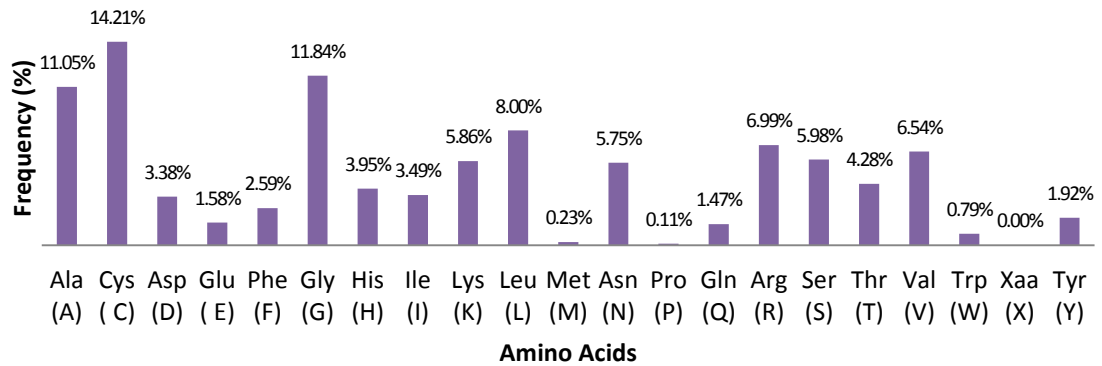




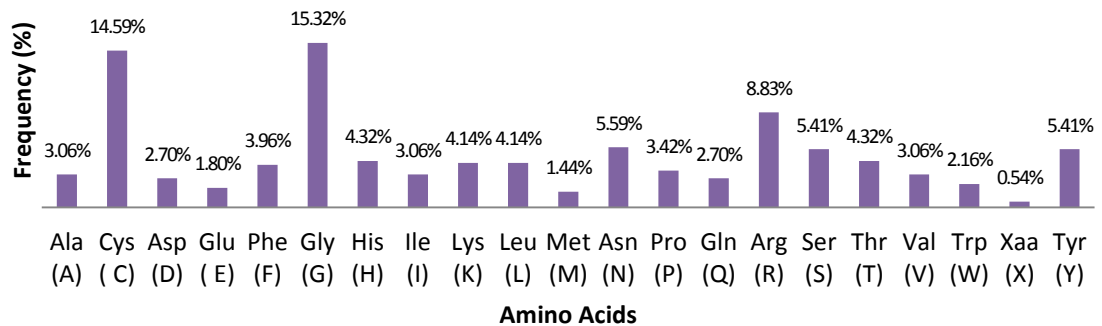
FSAP-Dermaseptin



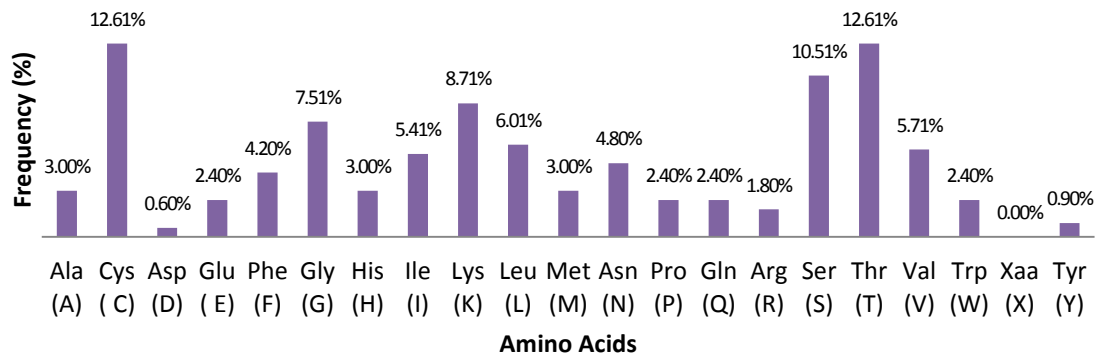
Invertebrate-Type1



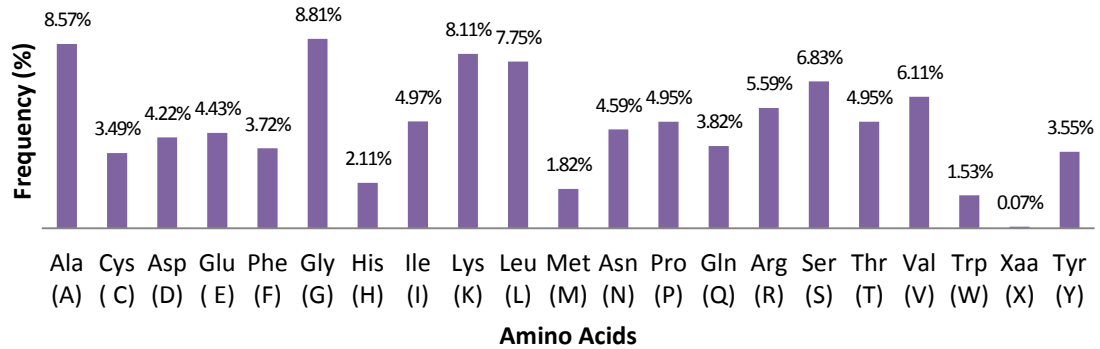
Invertebrate-Type2



Type A Lantibiotic



Other AMP Families



Appendix 2: Clustering Performance Using Different Values of Terminal Length

Parameters

The following tables show the clustering performance of selected properties using different values of terminal length parameters, dn and dc. Each row in a table refers to an AMP family of the 14 target AMP families in the same order as shown in table 3.1.

dn = 10, dc=8

Number of peptides	Original number of features	number of selected features	Number of clusters	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
34	322	19	11	99.60%	97.06%	99.72%	94.29%	91.67%	95.65%	0.0239	99.60%
24	216	15	9	99.73%	91.67%	100.00%	100.00%	91.67%	95.65%	0.0204	99.73%
41	258	30	14	99.34%	90.24%	99.86%	97.37%	88.10%	93.67%	0.0436	99.34%
31	1101	16	4	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
27	791	71	6	97.61%	66.67%	98.76%	66.67%	50.00%	66.67%	0.1216	97.61%
30	712	21	12	98.94%	100.00%	98.89%	78.95%	78.95%	88.24%	0.0375	98.94%
12	346	11	10	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
37	253	34	15	99.20%	86.49%	99.86%	96.97%	84.21%	91.43%	0.052	99.20%
143	1941	346	11	88.71%	76.22%	91.64%	68.12%	56.19%	71.95%	0.3398	89.11%
11	1942	54	15	99.73%	90.91%	99.87%	90.91%	83.33%	90.91%	0.0175	99.73%
30	294	21	10	99.47%	90.00%	99.86%	96.43%	87.10%	93.10%	0.0397	99.47%
21	299	20	9	99.87%	95.24%	100.00%	100.00%	95.24%	97.56%	0.0095	99.87%
13	1078	36	15	99.20%	69.23%	99.73%	81.82%	60.00%	75.00%	0.0408	99.20%
11	424	32	14	99.34%	72.73%	99.73%	80.00%	61.54%	76.19%	0.0372	99.34%

dn = 10, dc=10

Number of peptides	Original number of features	number of selected features	Number of clusters	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
34	299	14	14	99.73%	94.12%	100.00%	100.00%	94.12%	96.97%	0.0165	99.73%
24	228	19	13	99.87%	95.83%	100.00%	100.00%	95.83%	97.87%	0.0094	99.87%
41	261	36	14	99.60%	95.12%	99.86%	97.50%	92.86%	96.30%	0.0289	99.60%
31	1105	24	15	99.07%	77.42%	100.00%	100.00%	77.42%	87.27%	0.0592	99.07%
27	794	43	12	97.34%	44.44%	99.31%	70.59%	37.50%	54.55%	0.1416	97.34%
30	725	31	12	99.47%	86.67%	100.00%	100.00%	86.67%	92.86%	0.027	99.47%
12	350	10	13	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%

165

37	253	36	13	98.94%	91.89%	99.30%	87.18%	80.95%	89.47%	0.0595	98.94%
143	1945	126	11	95.48%	77.62%	99.67%	98.23%	76.55%	86.72%	0.1732	95.48%
11	1946	235	15	99.34%	81.82%	99.60%	75.00%	64.29%	78.26%	0.0295	99.34%
30	327	27	11	99.60%	96.67%	99.72%	93.55%	90.62%	95.08%	0.0253	99.60%
21	402	14	14	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
13	1088	35	11	99.07%	69.23%	99.59%	75.00%	56.25%	72.00%	0.0498	99.07%
11	194	17	15	98.54%	54.55%	99.19%	50.00%	35.29%	52.17%	0.0595	98.54%

dn = 12, dc=8

Number of peptides	Original number of features	number of selected features	Number of clusters	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
34	322	7	12	99.47%	88.24%	100.00%	100.00%	88.24%	93.75%	0.0319	99.47%
24	227	11	13	99.73%	91.67%	100.00%	100.00%	91.67%	95.65%	0.0195	99.73%
41	271	30	14	99.47%	90.24%	100.00%	100.00%	90.24%	94.87%	0.0367	99.47%
31	1092	16	13	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
27	789	81	15	97.74%	62.96%	99.04%	70.83%	50.00%	66.67%	0.1042	97.74%
30	835	33	11	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
12	347	9	15	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
37	252	26	14	99.47%	89.19%	100.00%	100.00%	89.19%	94.29%	0.0325	99.47%
143	1929	138	11	95.48%	80.42%	99.02%	95.04%	77.18%	87.12%	0.1738	95.88%
11	1943	28	8	99.73%	81.82%	100.00%	100.00%	81.82%	90.00%	0.0188	99.73%
30	295	19	12	99.47%	90.00%	99.86%	96.43%	87.10%	93.10%	0.0344	99.47%
21	208	14	12	99.87%	95.24%	100.00%	100.00%	95.24%	97.56%	0.0082	99.87%
13	795	15	12	99.34%	61.54%	100.00%	100.00%	61.54%	76.19%	0.0449	99.34%
11	192	35	15	98.80%	81.82%	99.06%	56.25%	50.00%	66.67%	0.0415	98.80%

dn = 12, dc=10

Number of peptides	Original number of features	number of selected features	Number of clusters	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
34	299	18	8	99.47%	88.24%	100.00%	100.00%	88.24%	93.75%	0.0411	99.47%
24	240	12	15	99.73%	91.67%	100.00%	100.00%	91.67%	95.65%	0.0193	99.73%
41	275	30	14	99.34%	87.80%	100.00%	100.00%	87.80%	93.51%	0.0362	99.34%
31	1095	24	13	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
27	791	33	14	98.01%	48.15%	99.86%	92.86%	46.43%	63.41%	0.1106	98.01%
30	781	32	14	99.73%	93.33%	100.00%	100.00%	93.33%	96.55%	0.017	99.73%
12	350	7	14	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
37	252	27	13	99.20%	83.78%	100.00%	100.00%	83.78%	91.18%	0.0532	99.20%
143	1945	165	13	95.62%	78.32%	99.67%	98.25%	77.24%	87.16%	0.1549	96.15%
11	1946	169	15	99.47%	72.73%	99.87%	88.89%	66.67%	80.00%	0.0293	99.47%

30	327	47	15	99.47%	96.67%	99.59%	90.62%	87.88%	93.55%	0.0293	99.47%
21	208	19	14	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
13	806	25	15	99.20%	69.23%	99.73%	81.82%	60.00%	75.00%	0.0427	99.20%
11	194	30	14	99.20%	100.00%	99.19%	64.71%	64.71%	78.57%	0.0211	99.20%

dn = 14, dc=8

Number of peptides	Original number of features	number of selected features	Number of clusters	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
34	321	12	15	99.47%	88.24%	100.00%	100.00%	88.24%	93.75%	0.0352	99.47%
24	210	13	10	99.73%	91.67%	100.00%	100.00%	91.67%	95.65%	0.0227	99.73%
41	348	53	15	99.20%	85.37%	100.00%	100.00%	85.37%	92.11%	0.046	99.20%
31	1091	37	15	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
27	790	68	14	98.41%	77.78%	99.17%	77.78%	63.64%	77.78%	0.0784	98.41%
30	894	53	13	99.73%	96.67%	99.86%	96.67%	93.55%	96.67%	0.018	99.73%
12	346	12	15	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
37	246	30	9	98.94%	81.08%	99.86%	96.77%	78.95%	88.24%	0.0706	98.94%
143	1941	179	14	91.24%	60.84%	98.36%	89.69%	56.86%	72.50%	0.221	93.63%
11	1942	44	14	99.87%	90.91%	100.00%	100.00%	90.91%	95.24%	0.0098	99.87%
30	294	26	12	99.60%	96.67%	99.72%	93.55%	90.62%	95.08%	0.026	99.60%
21	208	15	12	99.87%	95.24%	100.00%	100.00%	95.24%	97.56%	0.0083	99.87%
13	796	37	14	99.34%	69.23%	99.86%	90.00%	64.29%	78.26%	0.0385	99.34%
11	192	44	15	98.67%	100.00%	98.65%	52.38%	52.38%	68.75%	0.0278	98.67%

dn = 14, dc=10

Number of peptides	Original number of features	number of selected features	Number of clusters	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
34	299	13	13	99.47%	88.24%	100.00%	100.00%	88.24%	93.75%	0.0376	99.47%
24	225	9	12	99.87%	95.83%	100.00%	100.00%	95.83%	97.87%	0.0116	99.87%
41	354	58	14	98.27%	78.05%	99.44%	88.89%	71.11%	83.12%	0.1024	98.27%
31	1095	17	10	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
27	793	47	13	97.74%	48.15%	99.59%	81.25%	43.33%	60.47%	0.1018	97.74%
30	290	13	8	99.47%	86.67%	100.00%	100.00%	86.67%	92.86%	0.0302	99.47%
12	350	10	14	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
37	246	25	13	99.34%	86.49%	100.00%	100.00%	86.49%	92.75%	0.0424	99.34%
143	1945	118	12	95.88%	79.02%	99.84%	99.12%	78.47%	87.94%	0.1478	95.88%
11	1946	28	14	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
30	327	29	15	99.60%	96.67%	99.72%	93.55%	90.62%	95.08%	0.025	99.60%
21	208	19	15	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
13	807	27	15	99.07%	69.23%	99.59%	75.00%	56.25%	72.00%	0.0475	99.07%

11	194	33	15	99.34%	100.00%	99.33%	68.75%	68.75%	81.48%	0.019	99.34%
-----------	-----	----	----	--------	---------	--------	--------	--------	--------	-------	--------

dn = 16, dc=8

Number of peptides	Original number of features	number of selected features	Number of clusters	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
34	489	11	9	99.47%	88.24%	100.00%	100.00%	88.24%	93.75%	0.0403	99.47%
24	244	17	13	99.60%	87.50%	100.00%	100.00%	87.50%	93.33%	0.0306	99.60%
41	377	45	15	98.67%	80.49%	99.72%	94.29%	76.74%	86.84%	0.0808	98.67%
31	1091	22	14	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
27	521	36	11	98.80%	88.89%	99.17%	80.00%	72.73%	84.21%	0.058	98.80%
30	290	20	12	99.34%	83.33%	100.00%	100.00%	83.33%	90.91%	0.0417	99.34%
12	346	11	10	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
37	234	27	15	99.20%	83.78%	100.00%	100.00%	83.78%	91.18%	0.0491	99.20%
143	1941	222	14	89.64%	58.04%	97.05%	82.18%	51.55%	68.03%	0.2679	91.37%
11	1942	29	15	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
30	294	31	10	99.47%	96.67%	99.59%	90.62%	87.88%	93.55%	0.0297	99.47%
21	208	18	12	99.87%	95.24%	100.00%	100.00%	95.24%	97.56%	0.0094	99.87%
13	510	9	15	99.34%	61.54%	100.00%	100.00%	61.54%	76.19%	0.0426	99.34%
11	192	38	14	98.94%	100.00%	98.92%	57.89%	57.89%	73.33%	0.0248	98.94%

dn = 16, dc=10

Number of peptides	Original number of features	number of selected features	Number of clusters	Accuracy	Sensitivity	Specificity	Precision	Jaccard Index	F-Measure	Entropy	Purity
34	482	24	14	99.60%	91.18%	100.00%	100.00%	91.18%	95.38%	0.0233	99.60%
24	313	12	13	99.60%	87.50%	100.00%	100.00%	87.50%	93.33%	0.0246	99.60%
41	383	49	14	98.41%	75.61%	99.72%	93.94%	72.09%	83.78%	0.099	98.41%
31	1095	13	9	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
27	524	27	9	97.34%	48.15%	99.17%	68.42%	39.39%	56.52%	0.1284	97.34%
30	290	13	10	99.47%	86.67%	100.00%	100.00%	86.67%	92.86%	0.0325	99.47%
12	350	10	11	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0	100.00%
37	269	27	15	98.80%	78.38%	99.86%	96.67%	76.32%	86.57%	0.0744	98.80%
143	1945	125	12	94.16%	77.62%	98.03%	90.24%	71.61%	83.46%	0.2072	94.82%
11	1946	101	14	99.87%	100.00%	99.87%	91.67%	91.67%	95.65%	0.0066	99.87%
30	327	25	15	99.60%	96.67%	99.72%	93.55%	90.62%	95.08%	0.0254	99.60%
21	208	18	11	99.87%	95.24%	100.00%	100.00%	95.24%	97.56%	0.0102	99.87%
13	541	10	11	99.07%	53.85%	99.86%	87.50%	50.00%	66.67%	0.049	99.07%
11	194	26	14	99.47%	100.00%	99.46%	73.33%	73.33%	84.62%	0.0167	99.47%

Appendix 3: Selected Properties Using GA to Discriminate 14 AMP Families

The following table shows the entire set of the selected compositional and physicochemical properties by the GA. Different properties were selected for each AMP family.

AMP Family Name	Number of Selected Properties	Selected Properties
Alpha-defensin	14	<p>Basic Properties Related To Composition and Distance Frequency:</p> <ul style="list-style-type: none"> Frequency of Amino Acid (Q) in n2 region Frequency of Amino Acid (G) in n2 region Frequency of Amino Acid (F) in n2 region Frequency of Amino Acid (C) in n3 region Frequency of Amino Acid (F) in n3 region Frequency of Amino Acid (Y) in n3 region Frequency of Amino Acid (N) in n4 region Frequency of Amino Acid (C) in n4 region Frequency of pairs Amino Acid (R) in M region Frequency of Amino Acid (R) in C region Frequency of Amino Acid (K) in C region Frequency of Amino Acid (Y) in C region <p>Properties Extracted from n1 Sub-Region:</p> <ul style="list-style-type: none"> Short and medium range non-bonded energy per atom (Oobatake-Ooi, 1977) <p>Properties Extracted from C Region:</p> <ul style="list-style-type: none"> Weights for beta-sheet at the window position of 2 (Qian-Sejnowski, 1988)
Bacteriocin	9	<p>Basic Properties Related To Composition and Distance Frequency:</p> <ul style="list-style-type: none"> Frequency of Amino Acid (G) in n2 region Frequency of Amino Acid (V) in n2 region Frequency of Amino Acid (A) in n3 region Frequency of Amino Acid (N) in n3 region Frequency of Amino Acid (Y) in n3 region Frequency of (>16 aa and <=21 aa) Distances of basic amino

		<p>acids (RHK) in N region</p> <p>Frequency of Amino Acid (C) in the entire sequence region</p> <p>Properties Extracted from n4 Sub-Region:</p> <p>The number of atoms in the side chain labelled 2+1 (Charton-Charton, 1983)</p> <p>Properties Extracted from M Region:</p> <p>van der Waals parameter epsilon (Levitt, 1976)</p>
Beta-defensin	36	<p>Basic Properties Related To Composition and Distance Frequency:</p> <p>Frequency of Amino Acid (R) in n1 region</p> <p>Frequency of Amino Acid (H) in n1 region</p> <p>Frequency of Amino Acid (N) in n2 region</p> <p>Frequency of Amino Acid (I) in n2 region</p> <p>Frequency of Amino Acid (L) in n2 region</p> <p>Frequency of Amino Acid (V) in n2 region</p> <p>Frequency of Amino Acid (L) in n3 region</p> <p>Frequency of Amino Acid (M) in n3 region</p> <p>Frequency of Amino Acid (F) in n3 region</p> <p>Frequency of Amino Acid (P) in n3 region</p> <p>Frequency of Amino Acid (T) in n3 region</p> <p>Frequency of Amino Acid (W) in n3 region</p> <p>Frequency of Amino Acid (Y) in n3 region</p> <p>Frequency of Amino Acid (N) in n4 region</p> <p>Frequency of Amino Acid (C) in n4 region</p> <p>Frequency of Amino Acid (G) in n4 region</p> <p>Frequency of Amino Acid (H) in n4 region</p> <p>Frequency of Amino Acid (L) in n4 region</p> <p>Frequency of Amino Acid (M) in n4 region</p> <p>Frequency of Amino Acid (T) in n4 region</p> <p>Frequency of Amino Acid (C) in M region</p> <p>Frequency of Amino Acid (R) in C region</p> <p>Frequency of Amino Acid (C) in C region</p> <p>Frequency of Amino Acid (M) in C region</p> <p>Frequency of Amino Acid (W) in C region</p> <p>Frequency of Amino Acid (Y) in C region</p> <p>Frequency of (1 aa) Distances of basic amino acids (RHK) in N region</p> <p>Frequency of (>11 aa and <=16 aa) Distances of other non-hydrophobic amino acids (DNEQYSTC) in M region</p> <p>Frequency of Amino Acid (W) in the entire sequence region</p> <p>Properties Extracted from n1 Sub-Region:</p> <p>The number of atoms in the side chain labelled 2+1 (Charton-Charton, 1983)</p>

		<p>Properties Extracted from n2 Sub-Region: A parameter of charge transfer donor capability (Charton-Charton, 1983)</p> <p>Properties Extracted from n3 Sub-Region: Average relative fractional occurrence in AL(i-1) (Rackovsky-Scheraga, 1982) Normalized positional residue frequency at helix termini N" (Aurora-Rose,</p> <p>Properties Extracted from n4 Sub-Region: Average relative fractional occurrence in AL(i-1) (Rackovsky-Scheraga, 1982) A parameter of charge transfer donor capability (Charton-Charton, 1983) Linker propensity from all dataset (George-Heringa, 2003)</p>
Bombinin	13	<p>Basic Properties Related To Composition and Distance Frequency: Frequency of Amino Acid (G) in n1 region Frequency of Amino Acid (T) in n4 region Frequency of Amino Acid (G) in C region Frequency of Amino Acid (L) in C region</p> <p>Properties Extracted from n1 Sub-Region: Normalized frequency of beta-sheet (Crawford et al., 1973) Weights for alpha-helix at the window position of -4 (Qian-Sejnowski, 1988)</p> <p>Properties Extracted from n4 Sub-Region: AA composition of membrane proteins (Nakashima et al., 1990) Slope in regression analysis $\times 1.0E1$ (Prabhakaran-Ponnuswamy, 1982) Distribution of amino acid residues in the alpha-helices in thermophilic Averaged turn propensities in a transmembrane helix (Monne et al., 1999)</p> <p>Properties Extracted from C Region: Normalized relative frequency of extended structure (Isogai et al., 1980) Surface composition of amino acids in extracellular proteins of mesophiles Weights for alpha-helix at the window position of -4 (Qian-Sejnowski, 1988)</p>
Cathelicidin	36	<p>Basic Properties Related To Composition and Distance Frequency: Frequency of Amino Acid (R) in n1 region Frequency of Amino Acid (H) in n1 region</p>

		<p> Frequency of Amino Acid (Y) in n1 region Frequency of Amino Acid (R) in n2 region Frequency of Amino Acid (Q) in n2 region Frequency of Amino Acid (T) in n2 region Frequency of Amino Acid (Y) in n2 region Frequency of Amino Acid (C) in n3 region Frequency of Amino Acid (H) in n3 region Frequency of Amino Acid (Y) in n3 region Frequency of Amino Acid (C) in n4 region Frequency of pairs Amino Acid (A) in M region Frequency of Amino Acid (A) in C region Frequency of Amino Acid (K) in C region Frequency of Amino Acid (F) in C region Frequency of Amino Acid (P) in C region Frequency of Amino Acid (Y) in C region </p> <p>Properties Extracted from n3 Sub-Region:</p> <p> Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986) N.m.r. chemical shift of alpha-carbon (Fauchere et al., 1988) Relative mutability (Jones et al., 1992) Normalized frequency of beta-sheet, with weights (Levitt, 1978) Frequency of occurrence in beta-bends (Lewis et al., 1971) Normalized frequency of alpha region (Maxfield-Scheraga, 1976) AA composition of membrane proteins (Nakashima et al., 1990) Transmembrane regions of non-mt-proteins (Nakashima et al., 1990) Weights for coil at the window position of -5 (Qian-Sejnowski, 1988) Weights for coil at the window position of 5 (Qian-Sejnowski, 1988) Relative preference value at N2 (Richardson-Richardson, 1988) Principal property value z3 (Wold et al., 1987) Normalized positional residue frequency at helix termini N4'(Aurora-Rose, Normalized positional residue frequency at helix termini N3 (Aurora-Rose, Alpha-helix propensity derived from designed sequences (Koehl-Levitt, 1999) Linker propensity from 1-linker dataset (George-Heringa, 2003) </p> <p>Properties Extracted from C Region:</p> <p> Helix initiation parameter at position i,i+1,i+2 (Finkelstein et al., 1991) Helix termination parameter at position j+1 (Finkelstein et al., </p>
--	--	---

		1991) Relative preference value at C2 (Richardson-Richardson, 1988)
Cecropin	33	<p>Basic Properties Related To Composition and Distance Frequency:</p> <p>Frequency of Amino Acid (A) in n1 region Frequency of Amino Acid (D) in n1 region Frequency of Amino Acid (D) in n2 region Frequency of Amino Acid (K) in n3 region Frequency of Amino Acid (E) in n4 region Frequency of Amino Acid (A) in C region Frequency of (>16 aa and <=21 aa) Distances of other non-hydrophobic amino acids (DNEQYSTC) in M region Frequency of Amino Acid (Y) in the entire sequence region</p> <p>Properties Extracted from n1 Sub-Region:</p> <p>Positive charge (Fauchere et al., 1988) Conformational parameter of beta-turn (Beghin-Dirkx, 1975) Bitterness (Venanzi, 1984)</p> <p>Properties Extracted from n2 Sub-Region:</p> <p>Residue volume (Bigelow, 1967) Normalized frequency of beta-sheet (Crawford et al., 1973) Entropy of formation (Hutchens, 1970)</p> <p>Properties Extracted from n3 Sub-Region:</p> <p>Information measure for extended without H-bond (Robson-Suzuki, 1976) Weights for alpha-helix at the window position of 6 (Qian-Sejnowski, 1988) Weights for coil at the window position of 3 (Qian-Sejnowski, 1988) Free energy in beta-strand conformation (Munoz-Serrano, 1994) Distribution of amino acid residues in the 18 non-redundant families of</p> <p>Properties Extracted from n4 Sub-Region:</p> <p>Hydration number (Hopfinger, 1971), Cited by Charton-Charton (1982) Normalized positional residue frequency at helix termini N2 (Aurora-Rose, Relative preference value at N3 (Richardson-Richardson, 1988) Hydrophobicity coefficient in RP-HPLC, C4 with 0.1%TFA/MeCN/H2O (Wilce et al. AA composition of mt-proteins (Nakashima et al., 1990)</p> <p>Properties Extracted from M Region:</p> <p>Averaged turn propensities in a transmembrane helix (Monne</p>

		<p>et al., 1999)</p> <p>Propensity of amino acids within pi-helices (Fodje-Al-Karadaghi, 2002)</p> <p>Linker propensity from long dataset (linker length is greater than 14)</p> <p>Optimized beta-structure-coil equilibrium constant (Oobatake et al., 1985)</p> <p>Weights for alpha-helix at the window position of 5 (Qian-Sejnowski, 1988)</p> <p>Weights for coil at the window position of 5 (Qian-Sejnowski, 1988)</p> <p>Relative preference value at N" (Richardson-Richardson, 1988)</p> <p>Properties Extracted from C Region:</p> <p>Hydration number (Hopfinger, 1971), Cited by Charton-Charton (1982)</p> <p>van der Waals parameter epsilon (Levitt, 1976)</p>
Cyclotide (Bracelet sub-family)	7	<p>Basic Properties Related To Composition and Distance Frequency:</p> <p>Frequency of Amino Acid (E) in n4 region</p> <p>Frequency of Amino Acid (G) in C region</p> <p>Properties Extracted from n1 Sub-Region:</p> <p>Normalized positional residue frequency at helix termini Cc (Aurora-Rose,</p> <p>Properties Extracted from n2 Sub-Region:</p> <p>The number of atoms in the side chain labelled 2+1 (Charton-Charton, 1983)</p> <p>Properties Extracted from C Region:</p> <p>The Kerr-constant increments (Khanarian-Moore, 1980)</p> <p>Normalized frequency of turn in alpha+beta class (Palau et al., 1981)</p> <p>Hydrostatic pressure asymmetry index, PAI (Di Giulio, 2005)</p>
DEFL	26	<p>Basic Properties Related To Composition and Distance Frequency:</p> <p>Frequency of Amino Acid (E) in n1 region</p> <p>Frequency of Amino Acid (K) in n1 region</p> <p>Frequency of Amino Acid (S) in n1 region</p> <p>Frequency of Amino Acid (T) in n1 region</p> <p>Frequency of Amino Acid (C) in n2 region</p> <p>Frequency of Amino Acid (C) in n3 region</p> <p>Frequency of Amino Acid (E) in n3 region</p> <p>Frequency of Amino Acid (L) in n3 region</p> <p>Frequency of Amino Acid (Y) in n3 region</p> <p>Frequency of Amino Acid (V) in n3 region</p>

		<p>Frequency of Amino Acid (A) in n4 region Frequency of Amino Acid (C) in n4 region Frequency of Amino Acid (G) in n4 region Frequency of Amino Acid (K) in n4 region Frequency of Amino Acid (S) in n4 region Frequency of pairs Amino Acid (C) in M region Frequency of Amino Acid (C) in C region Frequency of Amino Acid (W) in C region Frequency of Amino Acid (V) in C region Frequency of Amino Acid (K) in the entire sequence region</p> <p>Properties Extracted from n3 Sub-Region: The number of atoms in the side chain labelled 2+1 (Charton-Charton, 1983)</p> <p>Properties Extracted from n4 Sub-Region: A parameter of charge transfer donor capability (Charton-Charton, 1983) Conformational parameter of inner helix (Beghin-Dirkx, 1975) Residue accessible surface area in folded protein (Chothia, 1976) Ratio of average and computed composition (Nakashima et al., 1990)</p> <p>Properties Extracted from M Region: The number of atoms in the side chain labelled 2+1 (Charton-Charton, 1983)</p>
FSAP (Brevinin sub-family)	118	<p>Basic Properties Related To Composition and Distance Frequency:</p> <p>Frequency of Amino Acid (R) in n1 region Frequency of Amino Acid (E) in n1 region Frequency of Amino Acid (F) in n1 region Frequency of Amino Acid (V) in n1 region Frequency of Amino Acid (D) in n2 region Frequency of Amino Acid (Q) in n2 region Frequency of Amino Acid (P) in n2 region Frequency of Amino Acid (W) in n2 region Frequency of Amino Acid (R) in n3 region Frequency of Amino Acid (C) in n3 region Frequency of Amino Acid (S) in n4 region Frequency of Amino Acid (W) in n4 region Frequency of Amino Acid (V) in n4 region Frequency of pairs Amino Acid (M) in M region Frequency of Amino Acid (C) in C region Frequency of Amino Acid (I) in C region Frequency of Amino Acid (S) in C region</p>

		<p>Frequency of Amino Acid (Y) in C region</p> <p>Properties Extracted from n1 Sub-Region:</p> <p>Normalized flexibility parameters (B-values) for each residue surrounded by</p> <p>Weights for beta-sheet at the window position of 3 (Qian-Sejnowski, 1988)</p> <p>Weights for coil at the window position of 3 (Qian-Sejnowski, 1988)</p> <p>Average relative fractional occurrence in EL(i-1) (Rackovsky-Scheraga, 1982)</p> <p>Normalized positional residue frequency at helix termini N4'(Aurora-Rose,</p> <p>Average relative fractional occurrence in E0(i-1) (Rackovsky-Scheraga, 1982)</p> <p>A parameter defined from the residuals obtained from the best correlation of</p> <p>Free energies of transfer of AcWI-X-LL peptides from bilayer interface to</p> <p>N.m.r. chemical shift of alpha-carbon (Fauchere et al., 1988)</p> <p>Information measure for C-terminal turn (Robson-Suzuki, 1976)</p> <p>Slopes tripeptide FDPB PARSE neutral (Avbelj, 2000)</p> <p>Screening coefficients gamma, local (Avbelj, 2000)</p> <p>Properties Extracted from n2 Sub-Region:</p> <p>Weights for alpha-helix at the window position of -6 (Qian-Sejnowski, 1988)</p> <p>Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)</p> <p>Direction of hydrophobic moment (Eisenberg-McLachlan, 1986)</p> <p>Side chain angle theta(AAR) (Levitt, 1976)</p> <p>Localized electrical effect (Fauchere et al., 1988)</p> <p>Normalized composition from fungi and plant (Nakashima et al., 1990)</p> <p>Frequency of the 4th residue in turn (Chou-Fasman, 1978b)</p> <p>Normalized positional residue frequency at helix termini N'(Aurora-Rose,</p> <p>Normalized frequency of extended structure (Burgess et al., 1974)</p> <p>Properties Extracted from n3 Sub-Region:</p> <p>Helix termination parameter at position j-2,j-1,j (Finkelstein et al., 1991)</p> <p>Normalized composition of mt-proteins (Nakashima et al., 1990)</p> <p>Residue accessible surface area in folded protein (Chothia, 1976)</p> <p>The number of atoms in the side chain labelled 1+1 (Charton-</p>
--	--	--

		<p>Charton, 1983)</p> <p>Normalized frequency of alpha region (Maxfield-Scheraga, 1976)</p> <p>pK (-COOH) (Jones, 1975)</p> <p>Normalized frequency of beta-sheet in alpha+beta class (Palau et al., 1981)</p> <p>Information measure for extended without H-bond (Robson-Suzuki, 1976)</p> <p>Weights for alpha-helix at the window position of 6 (Qian-Sejnowski, 1988)</p> <p>Weights for coil at the window position of 4 (Qian-Sejnowski, 1988)</p> <p>Optimized beta-structure-coil equilibrium constant (Oobatake et al., 1985)</p> <p>Linker propensity from helical (annotated by DSSP) dataset (George-Heringa,</p> <p>Optimized side chain interaction parameter (Oobatake et al., 1985)</p> <p>Free energies of transfer of AcWI-X-LL peptides from bilayer interface to</p> <p>Normalized frequency of turn (Crawford et al., 1973)</p> <p>Information measure for C-terminal turn (Robson-Suzuki, 1976)</p> <p>Normalized frequency of extended structure (Burgess et al., 1974)</p> <p>Properties Extracted from n4 Sub-Region:</p> <p>van der Waals parameter epsilon (Levitt, 1976)</p> <p>Normalized composition of membrane proteins (Nakashima et al., 1990)</p> <p>Electron-ion interaction potential (Veljkovic et al., 1985)</p> <p>Relative preference value at C1 (Richardson-Richardson, 1988)</p> <p>Relative population of conformational state E (Vasquez et al., 1983)</p> <p>Frequency of the 3rd residue in turn (Chou-Fasman, 1978b)</p> <p>Free energy in beta-strand conformation (Munoz-Serrano, 1994)</p> <p>Properties Extracted from M Region:</p> <p>Weights for beta-sheet at the window position of -3 (Qian-Sejnowski, 1988)</p> <p>Bitterness (Venanzi, 1984)</p> <p>Information measure for extended without H-bond (Robson-Suzuki, 1976)</p> <p>Entropy of formation (Hutchens, 1970)</p> <p>Relative preference value at C2 (Richardson-Richardson, 1988)</p>
--	--	--

		<p>Normalized positional residue frequency at helix termini N'(Aurora-Rose, Residue accessible surface area in folded protein (Chothia, 1976)</p> <p>Normalized frequency of beta-sheet (Chou-Fasman, 1978b) Relative mutability (Dayhoff et al., 1978b) Localized electrical effect (Fauchere et al., 1988)</p> <p>Weights for alpha-helix at the window position of 1 (Qian-Sejnowski, 1988)</p> <p>Smoothed epsilon steric parameter (Fauchere et al., 1988) The Kerr-constant increments (Khanarian-Moore, 1980)</p> <p>Weights for alpha-helix at the window position of -4 (Qian-Sejnowski, 1988)</p> <p>Average relative fractional occurrence in A0(i-1) (Rackovsky-Scheraga, 1982)</p> <p>Relative preference value at C4 (Richardson-Richardson, 1988)</p> <p>Normalized positional residue frequency at helix termini N" (Aurora-Rose, Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/MeCN/H2O (Wilce et Information value for accessibility; average fraction 35% (Biou et al., 1988)</p> <p>The Chou-Fasman parameter of the coil conformation (Charton-Charton, 1983)</p> <p>Weights for alpha-helix at the window position of 6 (Qian-Sejnowski, 1988)</p> <p>Weights for beta-sheet at the window position of 6 (Qian-Sejnowski, 1988)</p> <p>Weights for coil at the window position of 4 (Qian-Sejnowski, 1988)</p> <p>Average relative fractional occurrence in E0(i) (Rackovsky-Scheraga, 1982)</p> <p>Free energies of transfer of AcWI-X-LL peptides from bilayer interface to</p> <p>Weights for coil at the window position of -3 (Qian-Sejnowski, 1988)</p> <p>Thermodynamic beta sheet propensity (Kim-Berg, 1993) Linker propensity from long dataset (linker length is greater than 14</p> <p>Optimized average non-bonded energy per atom (Oobatake et al., 1985)</p> <p>Hydropathy scale based on self-information values in the two-state model (50%</p>
--	--	--

		<p>Properties Extracted from C Region:</p> <ul style="list-style-type: none"> Weights for alpha-helix at the window position of -6 (Qian-Sejnowski, 1988) Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986) Relative preference value at C3 (Richardson-Richardson, 1988) Proportion of residues 95% buried (Chothia, 1976) Weights for alpha-helix at the window position of -3 (Qian-Sejnowski, 1988) Weights for coil at the window position of -4 (Qian-Sejnowski, 1988) Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga, 1982) Slopes tripeptide FDPB PARSE neutral (Avbelj, 2000) Hydrophobicity index (Argos et al., 1982) Short and medium range non-bonded energy per atom (Oobatake-Ooi, 1977) Value of theta(i-1) (Rackovsky-Scheraga, 1982) Relative preference value at N3 (Richardson-Richardson, 1988) Information measure for C-terminal turn (Robson-Suzuki, 1976) Normalized positional residue frequency at helix termini C" (Aurora-Rose, p-Values of thermophilic proteins based on the distributions of B values Weights for alpha-helix at the window position of -4 (Qian-Sejnowski, 1988) Weights for beta-sheet at the window position of 2 (Qian-Sejnowski, 1988) Normalized positional residue frequency at helix termini C5 (Aurora-Rose, Hydropathy scale based on self-information values in the two-state model (50% Normalized frequency of N-terminal helix (Chou-Fasman, 1978b) Relative population of conformational state A (Vasquez et al., 1983) Normalized positional residue frequency at helix termini N2 (Aurora-Rose, Normalized frequency of beta-sheet in alpha+beta class (Palau et al., 1981) Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/2-PrOH/MeCN/H2O AA composition of EXT of multi-spanning proteins (Nakashima-Nishikawa, 1992)
--	--	--

FSAP (Caerin sub-family)	28	<p>Basic Properties Related To Composition and Distance Frequency: Frequency of Amino Acid (V) in C region</p> <p>Properties Extracted from n1 Sub-Region: Normalized hydrophobicity scales for alpha-proteins (Cid et al., 1992) Weights for alpha-helix at the window position of 5 (Qian-Sejnowski, 1988) van der Waals parameter R0 (Levitt, 1976) Surrounding hydrophobicity in turn (Ponnuswamy et al., 1980) Free energies of transfer of AcWI-X-LL peptides from bilayer interface to</p> <p>Properties Extracted from n2 Sub-Region: The Chou-Fasman parameter of the coil conformation (Charton-Charton, 1983) Relative preference value at C5 (Richardson-Richardson, 1988) Normalized frequency of isolated helix (Tanaka-Scheraga, 1977)</p> <p>Properties Extracted from n3 Sub-Region: Information measure for N-terminal turn (Robson-Suzuki, 1976) Electron-ion interaction potential (Veljkovic et al., 1985)</p> <p>Properties Extracted from n4 Sub-Region: Transmembrane regions of non-mt-proteins (Nakashima et al., 1990) Weights for alpha-helix at the window position of -6 (Qian-Sejnowski, 1988) Average relative fractional occurrence in EL(i-1) (Rackovsky-Scheraga, 1982) Relative preference value at C" (Richardson-Richardson, 1988)</p> <p>Properties Extracted from M Region: Helix initiation parameter at position i-1 (Finkelstein et al., 1991) Normalized relative frequency of helix end (Isogai et al., 1980) Weights for beta-sheet at the window position of -5 (Qian-Sejnowski, 1988) Relative preference value at C2 (Richardson-Richardson, 1988) Slopes dekapeptide, FDPB VFF neutral (Avbelj, 2000) Weights for coil at the window position of -6 (Qian-Sejnowski, 1988) Weights for coil at the window position of 6 (Qian-Sejnowski, 1988) Information measure for N-terminal turn (Robson-Suzuki, 1976)</p> <p>Properties Extracted from C Region: Optical rotation (Fasman, 1976)</p>
--------------------------	----	--

		<p>Beta-strand indices (Geisow-Roberts, 1980) Composition (Grantham, 1974) Normalized frequency of beta-sheet in alpha+beta class (Palau et al., 1981) Normalized positional residue frequency at helix termini N" (Aurora-Rose,</p>
<p>FSAP (Dermaseptin sub-family)</p>	25	<p>Basic Properties Related To Composition and Distance Frequency: Frequency of Amino Acid (D) in n1 region Frequency of Amino Acid (F) in n1 region Frequency of Amino Acid (W) in n1 region Frequency of Amino Acid (D) in n2 region Frequency of Amino Acid (V) in n2 region Frequency of Amino Acid (D) in n3 region Frequency of Amino Acid (M) in n3 region Frequency of Amino Acid (R) in n4 region Frequency of Amino Acid (W) in n4 region Frequency of Amino Acid (V) in M region Frequency of pairs Amino Acid (A) in M region Frequency of Amino Acid (C) in C region Frequency of Amino Acid (Q) in C region Frequency of Amino Acid (L) in C region Frequency of Amino Acid (K) in C region Frequency of Amino Acid (M) in C region</p> <p>Properties Extracted from n1 Sub-Region: Optimized average non-bonded energy per atom (Oobatake et al., 1985) Bitterness (Venanzi, 1984)</p> <p>Properties Extracted from n2 Sub-Region: Helix termination parameter at position $j-2, j-1, j$ (Finkelstein et al., 1991)</p> <p>Properties Extracted from n4 Sub-Region: Weights for alpha-helix at the window position of -6 (Qian-Sejnowski, 1988)</p> <p>Properties Extracted from M Region: van der Waals parameter epsilon (Levitt, 1976) Average interactions per side chain atom (Warme-Morgan, 1978)</p> <p>Properties Extracted from C Region: Entropy of formation (Hutchens, 1970) AA composition of EXT of multi-spanning proteins (Nakashima-Nishikawa, 1992) Information measure for extended without H-bond (Robson-</p>

		Suzuki, 1976)
Invertebrate defensin (Type 1 sub-family)	14	<p>Basic Properties Related To Composition and Distance Frequency:</p> <p>Frequency of Amino Acid (D) in n1 region Frequency of Amino Acid (A) in n2 region Frequency of Amino Acid (C) in n2 region Frequency of Amino Acid (C) in n4 region Frequency of Amino Acid (L) in n4 region Frequency of Amino Acid (C) in C region Frequency of Amino Acid (S) in C region Frequency of Amino Acid (V) in C region</p> <p>Properties Extracted from n4 Sub-Region: Linker propensity from 3-linker dataset (George-Heringa, 2003) Conformational parameter of inner helix (Beghin-Dirkx, 1975) The Chou-Fasman parameter of the coil conformation (Charton-Charton, 1983) Frequency of occurrence in beta-bends (Lewis et al., 1971) Retention coefficient in HPLC, pH2.1 (Meek, 1980) Principal component III (Sneath, 1966)</p>
Invertebrate defensin (Type 2 sub-family)	9	<p>Basic Properties Related To Composition and Distance Frequency:</p> <p>Frequency of Amino Acid (H) in n1 region Frequency of Amino Acid (C) in n2 region Frequency of Amino Acid (C) in n3 region Frequency of Amino Acid (T) in C region</p> <p>Properties Extracted from n1 Sub-Region: AA composition of EXT of multi-spanning proteins (Nakashima-Nishikawa, 1992) Value of theta(i) (Rackovsky-Scheraga, 1982) Loss of Side chain hydrophathy by helix formation (Roseman, 1988) Principal component IV (Sneath, 1966)</p> <p>Properties Extracted from n2 Sub-Region: Loss of Side chain hydrophathy by helix formation (Roseman, 1988)</p>
Type A lantibiotic	26	<p>Basic Properties Related To Composition and Distance Frequency:</p> <p>Frequency of Amino Acid (N) in n1 region Frequency of Amino Acid (D) in n1 region Frequency of Amino Acid (C) in n1 region Frequency of Amino Acid (S) in n1 region Frequency of Amino Acid (A) in n2 region Frequency of Amino Acid (C) in n2 region</p>

		<p> Frequency of Amino Acid (G) in n2 region Frequency of Amino Acid (V) in n2 region Frequency of Amino Acid (N) in n3 region Frequency of Amino Acid (T) in n4 region Frequency of Amino Acid (M) in M region Frequency of pairs Amino Acid (C) in M region Frequency of pairs Amino Acid (M) in M region Frequency of pairs Amino Acid (F) in M region Frequency of Amino Acid (N) in C region Frequency of Amino Acid (C) in C region Frequency of Amino Acid (Q) in C region Frequency of Amino Acid (K) in C region Frequency of Amino Acid (F) in C region Frequency of Amino Acid (S) in C region Frequency of Amino Acid (T) in C region Frequency of Amino Acid (Y) in C region Frequency of (>1 aa and <=6 aa) Distances of basic amino acids (RHK) in N region Frequency of Amino Acid (R) in the entire sequence region Frequency of Amino Acid (C) in the entire sequence region Frequency of Amino Acid (M) in the entire sequence region </p>
--	--	---

Appendix 4: DRAF Models Performance and Comparison of Results

Table A4.1. The prediction results on the test data by the RF models for TFBS lengths (L=7,8,9...20) using the holdout method. 70% of the data were used for training, while 30% were used for testing. The results below were obtained using the thresholds giving the highest **accuracy** on the training data

TFBS Length	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Average
threshold	0.3875	0.413	0.51	0.441	0.5105	0.568	0.5605	0.533	0.574	0.5745	0.5085	0.535	0.5215	0.314	
Total Number of Samples	1433	2460	19787	17583	39233	62890	78933	63680	58043	37997	22813	34110	45240	17143	
Total Training Samples	1003	1722	13851	12308	27463	44023	55253	44576	40630	26598	15969	23877	31668	12000	
Total Testing Samples	430	738	5936	5275	11770	18867	23680	19104	17413	11399	6844	10233	13572	5143	
True Positive (TP)	36	64	502	471	1049	1626	2032	1599	1401	914	591	815	1200	467	
False Positive (FP)	0	4	7	8	16	17	22	16	8	12	4	15	8	7	
True Negative (TN)	382	666	5386	4778	10647	17121	21504	17402	15861	10387	6202	9322	12308	4661	
False Negative (FN)	12	4	41	18	58	103	122	87	143	86	47	81	56	8	
Accuracy	0.9721	0.9892	0.9919	0.9951	0.9937	0.9936	0.9939	0.9946	0.9913	0.9914	0.9925	0.9906	0.9953	0.9971	99.16%
Sensitivity	0.7500	0.9412	0.9245	0.9632	0.9476	0.9404	0.9434	0.9484	0.9074	0.9140	0.9263	0.9096	0.9554	0.9832	92.53%

Specificity	1.0000	0.9940	0.9987	0.9983	0.9985	0.9990	0.9990	0.9991	0.9995	0.9988	0.9994	0.9984	0.9994	0.9985	99.86%
Precision	1.0000	0.9412	0.9862	0.9833	0.9850	0.9897	0.9893	0.9901	0.9943	0.9870	0.9933	0.9819	0.9934	0.9852	98.57%
F-Measure	0.8571	0.9412	0.9544	0.9731	0.9659	0.9644	0.9658	0.9688	0.9489	0.9491	0.9586	0.9444	0.9740	0.9842	95.36%
MCC	0.8527	0.9352	0.9505	0.9705	0.9627	0.9613	0.9628	0.9661	0.9453	0.9453	0.9552	0.9401	0.9717	0.9826	95.01%

Table A4.2. The prediction results on the test data by the RF models for TFBS lengths (L=7,8,9...20) using the holdout method. 70% of the data were used for training, while 30% were used for testing. The results below were obtained using the thresholds giving the highest **specificity** on the training data

TFBS Length	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Average
threshold	0.3875	0.5625	0.608	0.6265	0.651	0.793	0.823	0.6645	0.6585	0.7185	0.552	0.535	0.5215	0.314	
Total Number of Samples	1433	2460	19787	17583	39233	62890	78933	63680	58043	37997	22813	34110	45240	17143	
Total Training Samples	1003	1722	13851	12308	27463	44023	55253	44576	40630	26598	15969	23877	31668	12000	
Total Testing Samples	430	738	5936	5275	11770	18867	23680	19104	17413	11399	6844	10233	13572	5143	
True Positive (TP)	36	56	454	414	915	1261	1534	1501	1298	779	579	815	1200	467	
False Positive (FP)	0	1	3	4	5	3	2	2	4	7	3	15	8	7	

True Negative (TN)	382	669	5390	4782	10658	17135	21524	17416	15865	10392	6203	9322	12308	4661	
False Negative (FN)	12	12	89	75	192	468	620	185	246	221	59	81	56	8	
Accuracy	0.9721	0.9824	0.9845	0.9850	0.9833	0.9750	0.9737	0.9902	0.9856	0.9800	0.9909	0.9906	0.9953	0.9971	98.47%
Sensitivity	0.7500	0.8235	0.8361	0.8466	0.8266	0.7293	0.7122	0.8903	0.8407	0.7790	0.9075	0.9096	0.9554	0.9832	84.21%
Specificity	1.0000	0.9985	0.9994	0.9992	0.9995	0.9998	0.9999	0.9999	0.9997	0.9993	0.9995	0.9984	0.9994	0.9985	99.94%
Precision	1.0000	0.9825	0.9934	0.9904	0.9946	0.9976	0.9987	0.9987	0.9969	0.9911	0.9948	0.9819	0.9934	0.9852	99.28%
F-Measure	0.8571	0.8960	0.9080	0.9129	0.9028	0.8426	0.8314	0.9414	0.9122	0.8723	0.9492	0.9444	0.9740	0.9842	90.92%
MCC	0.8527	0.8906	0.9036	0.9081	0.8983	0.8414	0.8314	0.9379	0.9083	0.8690	0.9454	0.9401	0.9717	0.9826	90.58%

Table A4.3. The prediction results on the test data by the RF models for TFBS lengths (L=7,8,9...20) using the holdout method. 70% of the data were used for training, while 30% were used for testing. The results below were obtained using the thresholds giving the highest **sensitivity** on the training data

TFBS Length	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Average
threshold	0.4025	0.549	0.415	0.48	0.4855	0.488	0.474	0.504	0.5755	0.577	0.5505	0.578	0.5955	0.6295	
Total Number of Samples	1433	2460	19787	17583	39233	62890	78933	63680	58043	37997	22813	34110	45240	17143	
Total Training	1003	1722	13851	12308	27463	44023	55253	44576	40630	26598	15969	23877	31668	12000	

Samples															
Total Testing Samples	430	738	5936	5275	11770	18867	23680	19104	17413	11399	6844	10233	13572	5143	
True Positive (TP)	33	58	531	465	1062	1681	2088	1614	1395	913	580	777	1143	426	
False Positive (FP)	0	1	18	6	19	24	42	18	8	12	3	10	5	0	
True Negative (TN)	382	669	5375	4780	10644	17114	21484	17400	15861	10387	6203	9327	12311	4668	
False Negative (FN)	15	10	12	24	45	48	66	72	149	87	58	119	113	49	
Accuracy	0.9651	0.9851	0.9949	0.9943	0.9946	0.9962	0.9954	0.9953	0.9910	0.9913	0.9911	0.9874	0.9913	0.9905	99.03%
Sensitivity	0.6875	0.8529	0.9779	0.9509	0.9593	0.9722	0.9694	0.9573	0.9035	0.9130	0.9091	0.8672	0.9100	0.8968	90.91%
Specificity	1.0000	0.9985	0.9967	0.9987	0.9982	0.9986	0.9980	0.9990	0.9995	0.9988	0.9995	0.9989	0.9996	1.0000	99.89%
Precision	1.0000	0.9831	0.9672	0.9873	0.9824	0.9859	0.9803	0.9890	0.9943	0.9870	0.9949	0.9873	0.9956	1.0000	98.82%
F-Measure	0.8148	0.9134	0.9725	0.9688	0.9707	0.9790	0.9748	0.9729	0.9467	0.9486	0.9500	0.9234	0.9509	0.9456	94.52%
MCC	0.8133	0.9080	0.9698	0.9658	0.9678	0.9770	0.9723	0.9705	0.9431	0.9447	0.9463	0.9188	0.9473	0.9421	94.19%

Table A4.4. The prediction results on the test data by the RF models for TFBS lengths (L=7, 8, 9...20) using 10-fold cross-validation. The

results below were obtained using the thresholds giving the highest accuracy on the training data

TFBS Length	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Average
-------------	---	---	---	----	----	----	----	----	----	----	----	----	----	----	---------

threshold	0.2695	0.3575	0.421	0.4555	0.3975	0.41	0.466	0.3715	0.373	0.3775	0.349	0.3485	0.3925	0.3515	
Total Number of Samples	3283	5720	46073	40947	91463	146710	184100	148540	135407	88637	53223	79577	105550	39993	
Total Samples for Cross Valdiation	985	1716	13822	12284	27439	44013	55230	44562	40622	26591	15967	23873	31665	11998	
True Positive (TP)	85	152	1225	1104	2498	4022	4856	4006	3661	2396	1410	2144	2862	1074	
False Positive (FP)	16	9	45	23	75	94	100	94	118	118	31	129	65	22	
True Negative (TN)	878	1551	12519	11116	24830	39845	50126	40421	36802	24046	14483	21576	28698	10878	
False Negative (FN)	6	4	33	41	36	52	148	41	41	31	43	24	40	24	
Accuracy	0.9777	0.9924	0.9944	0.9948	0.9960	0.9967	0.9955	0.9970	0.9961	0.9944	0.9954	0.9936	0.9967	0.9962	0.9940
Sensitivity	0.9341	0.9744	0.9738	0.9642	0.9858	0.9872	0.9704	0.9899	0.9889	0.9872	0.9704	0.9889	0.9862	0.9781	0.9771
Specificity	0.9821	0.9942	0.9964	0.9979	0.9970	0.9976	0.9980	0.9977	0.9968	0.9951	0.9979	0.9941	0.9977	0.9980	0.9958
Precision	0.8416	0.9441	0.9646	0.9796	0.9709	0.9772	0.9798	0.9771	0.9688	0.9531	0.9785	0.9432	0.9778	0.9799	0.9597
F-Measure	0.8854	0.9590	0.9691	0.9718	0.9783	0.9822	0.9751	0.9834	0.9787	0.9698	0.9744	0.9655	0.9820	0.9790	0.9681
MCC	0.8745	0.9550	0.9661	0.9690	0.9761	0.9804	0.9726	0.9818	0.9767	0.9669	0.9719	0.9623	0.9802	0.9769	0.9650

Table A4.5. The prediction results on the test data by the RF models for TFBS lengths (L=7, 8, 9...20) using 10-fold cross-validation. The results below were obtained using the thresholds giving the highest specificity on the training data

TFBS Length	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Average
threshold	0.431	0.4195	0.75	0.7945	0.8525	0.9315	0.884	0.953	0.8305	0.8625	0.8275	0.9285	0.7765	0.4015	
Total Number of Samples	3283	5720	46073	40947	91463	146710	184100	148540	135407	88637	53223	79577	105550	39993	
Total Samples for Cross Valdiation	985	1716	13822	12284	27439	44013	55230	44562	40622	26591	15967	23873	31665	11998	
True Positive (TP)	70	142	635	609	1023	1131	2729	1404	2404	1216	1119	460	2096	1061	
False Positive (FP)	3	6	1	3	1	0	4	1	1	3	0	1	2	11	
True Negative (TN)	891	1554	12563	11136	24904	39939	50222	40514	36919	24161	14514	21704	28761	10889	
False Negative (FN)	21	14	623	536	1511	2943	2275	2643	1298	1211	334	1708	806	37	
Accuracy	0.9756	0.9883	0.9549	0.9561	0.9449	0.9331	0.9587	0.9407	0.9680	0.9543	0.9791	0.9284	0.9745	0.9960	0.9609
Sensitivity	0.7692	0.9103	0.5048	0.5319	0.4037	0.2776	0.5454	0.3469	0.6494	0.5010	0.7701	0.2122	0.7223	0.9663	0.5794
Specificity	0.9966	0.9962	0.9999	0.9997	1.0000	1.0000	0.9999	1.0000	1.0000	0.9999	1.0000	1.0000	0.9999	0.9990	0.9994
Precision	0.9589	0.9595	0.9984	0.9951	0.9990	1.0000	0.9985	0.9993	0.9996	0.9975	1.0000	0.9978	0.9990	0.9897	0.9923
F-Measure	0.8537	0.9342	0.6705	0.6932	0.5750	0.4346	0.7054	0.5150	0.7873	0.6670	0.8701	0.3499	0.8384	0.9779	0.7052
MCC	0.8466	0.9282	0.6929	0.7103	0.6166	0.5085	0.7217	0.5704	0.7919	0.6898	0.8676	0.4429	0.8377	0.9758	0.7286

Table A4.6. The prediction results on the test data by the RF models for TFBS lengths (L=7, 8, 9...20) using 10-fold cross-validation. The results below were obtained using the thresholds giving the highest sensitivity on the training data

TFBS Length	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Average
threshold	0.2715	0.36	0.222	0.2825	0.2865	0.147	0.228	0.2	0.231	0.163	0.164	0.154	0.131	0.1495	
Total Number of Samples	3283	5720	46073	40947	91463	146710	184100	148540	135407	88637	53223	79577	105550	39993	
Total Samples for Cross Valdiation	985	1716	13822	12284	27439	44013	55230	44562	40622	26591	15967	23873	31665	11998	
True Positive (TP)	85	151	1258	1142	2530	4073	4997	4046	3700	2427	1449	2168	2901	1098	
False Positive (FP)	14	8	255	108	197	944	530	464	345	588	225	710	895	180	
True Negative (TN)	880	1552	12309	11031	24708	38995	49696	40051	36575	23576	14289	20995	27868	10720	
False Negative (FN)	6	5	0	3	4	1	7	1	2	0	4	0	1	0	
Accuracy	0.9797	0.9924	0.9816	0.9910	0.9927	0.9785	0.9903	0.9896	0.9915	0.9779	0.9857	0.9703	0.9717	0.9850	0.9841
Sensitivity	0.9341	0.9679	1.0000	0.9974	0.9984	0.9998	0.9986	0.9998	0.9995	1.0000	0.9972	1.0000	0.9997	1.0000	0.9923
Specificity	0.9843	0.9949	0.9797	0.9903	0.9921	0.9764	0.9894	0.9885	0.9907	0.9757	0.9845	0.9673	0.9689	0.9835	0.9833
Precision	0.8586	0.9497	0.8315	0.9136	0.9278	0.8118	0.9041	0.8971	0.9147	0.8050	0.8656	0.7533	0.7642	0.8592	0.8612
F-Measure	0.8947	0.9587	0.9080	0.9537	0.9618	0.8961	0.9490	0.9457	0.9552	0.8920	0.9268	0.8593	0.8662	0.9242	0.9208
MCC	0.8845	0.9546	0.9025	0.9498	0.9585	0.8902	0.9451	0.9416	0.9516	0.8862	0.9217	0.8536	0.8603	0.9192	0.9157

Table A4.7. The prediction results on the test data by the RF models for TFBS lengths (L=7, 8, 9...20) using 5-fold cross-validation. The results below were obtained using the thresholds giving the highest accuracy on the training data

TFBS Length	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Average
threshold	0.3165	0.4345	0.4125	0.404	0.3965	0.3995	0.3885	0.3815	0.3615	0.428	0.3145	0.404	0.361	0.2815	
Total Number of Samples	3283	5720	46073	40947	91463	146710	184100	148540	135407	88637	53223	79577	105550	39993	
Total Samples for Cross Valdiation	985	1716	13822	12284	27439	44013	55230	44562	40622	26591	15967	23873	31665	11998	
True Positive (TP)	81	141	1217	1119	2492	4022	4927	4002	3653	2370	1423	2115	2877	1085	
False Positive (FP)	12	8	51	34	77	109	173	92	123	89	43	80	82	36	
True Negative (TN)	882	1552	12513	11105	24828	39830	50053	40423	36797	24075	14471	21625	28681	10864	
False Negative (FN)	10	15	41	26	42	52	77	45	49	57	30	53	25	13	
Accuracy	0.9777	0.9866	0.9933	0.9951	0.9957	0.9963	0.9955	0.9969	0.9958	0.9945	0.9954	0.9944	0.9966	0.9959	0.9936
Sensitivity	0.8901	0.9038	0.9674	0.9773	0.9834	0.9872	0.9846	0.9889	0.9868	0.9765	0.9794	0.9756	0.9914	0.9882	0.9700
Specificity	0.9866	0.9949	0.9959	0.9969	0.9969	0.9973	0.9966	0.9977	0.9967	0.9963	0.9970	0.9963	0.9971	0.9967	0.9959
Precision	0.8710	0.9463	0.9598	0.9705	0.9700	0.9736	0.9661	0.9775	0.9674	0.9638	0.9707	0.9636	0.9723	0.9679	0.9600
F-Measure	0.8804	0.9246	0.9636	0.9739	0.9767	0.9804	0.9753	0.9832	0.9770	0.9701	0.9750	0.9695	0.9817	0.9779	0.9649
MCC	0.8682	0.9175	0.9599	0.9712	0.9743	0.9784	0.9728	0.9815	0.9747	0.9671	0.9725	0.9665	0.9799	0.9757	0.9615

Table A4.8. The prediction results on the test data by the RF models for TFBS lengths (L=7, 8, 9...20) using 5-fold cross-validation. The results below were obtained using the thresholds giving the highest specificity on the training data































TFBS Length	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Average
threshold	0.3165	0.4345	0.688	0.715	0.7595	0.896	0.884	0.958	0.812	0.8995	0.773	0.934	0.808	0.383	
Total Number of Samples	3283	5720	46073	40947	91463	146710	184100	148540	135407	88637	53223	79577	105550	39993	
Total Samples for Cross Valdiation	985	1716	13822	12284	27439	44013	55230	44562	40622	26591	15967	23873	31665	11998	
True Positive (TP)	81	141	815	816	1617	1848	2672	1275	2474	922	1158	408	1908	1068	
False Positive (FP)	12	8	2	4	6	0	5	1	2	1	0	1	2	13	
True Negative (TN)	882	1552	12562	11135	24899	39939	50221	40514	36918	24163	14514	21704	28761	10887	
False Negative (FN)	10	15	443	329	917	2226	2332	2772	1228	1505	295	1760	994	30	
Accuracy	0.9777	0.9866	0.9678	0.9729	0.9664	0.9494	0.9577	0.9378	0.9697	0.9434	0.9815	0.9262	0.9685	0.9964	0.9644
Sensitivity	0.8901	0.9038	0.6479	0.7127	0.6381	0.4536	0.5340	0.3150	0.6683	0.3799	0.7970	0.1882	0.6575	0.9727	0.6256
Specificity	0.9866	0.9949	0.9998	0.9996	0.9998	1.0000	0.9999	1.0000	0.9999	1.0000	1.0000	1.0000	0.9999	0.9988	0.9985
Precision	0.8710	0.9463	0.9976	0.9951	0.9963	1.0000	0.9981	0.9992	0.9992	0.9989	1.0000	0.9976	0.9990	0.9880	0.9847
F-Measure	0.8804	0.9246	0.7855	0.8305	0.7780	0.6241	0.6957	0.4791	0.8009	0.5504	0.8870	0.3166	0.7930	0.9803	0.7376
MCC	0.8682	0.9175	0.7900	0.8297	0.7829	0.6555	0.7136	0.5428	0.8039	0.5976	0.8838	0.4166	0.7967	0.9783	0.7555

Table A4.9. The prediction results on the test data by the RF models for TFBS lengths (L=7, 8, 9...20) using 5-fold cross-validation. The results below were obtained using the thresholds giving the highest sensitivity on the training data

TFBS Length	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Average
threshold	0.321	0.3745	0.256	0.2525	0.2005	0.158	0.199	0.12	0.228	0.1595	0.2435	0.19	0.0995	0.2125	
Total Number of Samples	3283	5720	46073	40947	91463	146710	184100	148540	135407	88637	53223	79577	105550	39993	
Total Samples for Cross Valdiation	985	1716	13822	12284	27439	44013	55230	44562	40622	26591	15967	23873	31665	11998	
True Positive (TP)	80	151	1254	1143	2534	4074	4999	4047	3698	2427	1440	2168	2902	1094	
False Positive (FP)	11	11	188	132	495	917	752	1281	376	617	93	498	1434	84	
True Negative (TN)	883	1549	12376	11007	24410	39022	49474	39234	36544	23547	14421	21207	27329	10816	
False Negative (FN)	11	5	4	2	0	0	5	0	4	0	13	0	0	4	
Accuracy	0.9777	0.9907	0.9861	0.9891	0.9820	0.9792	0.9863	0.9713	0.9906	0.9768	0.9934	0.9791	0.9547	0.9927	0.9821
Sensitivity	0.8791	0.9679	0.9968	0.9983	1.0000	1.0000	0.9990	1.0000	0.9989	1.0000	0.9911	1.0000	1.0000	0.9964	0.9877
Specificity	0.9877	0.9929	0.9850	0.9881	0.9801	0.9770	0.9850	0.9684	0.9898	0.9745	0.9936	0.9771	0.9501	0.9923	0.9816
Precision	0.8791	0.9321	0.8696	0.8965	0.8366	0.8163	0.8692	0.7596	0.9077	0.7973	0.9393	0.8132	0.6693	0.9287	0.8510
F-Measure	0.8791	0.9497	0.9289	0.9446	0.9110	0.8988	0.9296	0.8634	0.9511	0.8872	0.9645	0.8970	0.8019	0.9613	0.9120
MCC	0.8668	0.9448	0.9239	0.9403	0.9055	0.8930	0.9248	0.8576	0.9473	0.8814	0.9613	0.8914	0.7974	0.9580	0.9067

Appendix 5: Sequence Logos for 321 ChIP-seq Datasets

Table A5.1. The sequence logos for the predicted TFBS sequences by the DRAF models for 321 ChIP-seq datasets at different sensitivity thresholds.

TF	Cell Type	Original TFBS Sequence Logo	Sensitivity 60%	Sensitivity 70%	Sensitivity 80%	Sensitivity 90%
ATF1	K562					
ATF3	A549					
ATF3	GM12878					
ATF3	H1-hESC					
ATF3	HepG2					
ATF3	K562					

ATF3	K562					
BATF	GM12878					
BRCA1	GM12878					
BRCA1	H1-hESC					
BRCA1	HeLa-S3					
BRCA1	HepG2					
CEBPB	A549					
CEBPB	GM12878					
CEBPB	H1-hESC					

CEBPB	HeLa-S3					
CEBPB	HepG2					
CEBPB	HepG2					
CEBPB	HepG2					
CEBPB	IMR90					
CEBPB	K562					
CEBPB	K562					
CEBPD	HepG2					
CREB1	A549					

CTCF	A549					
CTCF	A549					
CTCF	A549					
CTCF	A549					
CTCF	AG04449					
CTCF	AG04450					
CTCF	AG09309					
CTCF	AG09319					
CTCF	AG10803					

CTCF	AoAF					
CTCF	BJ					
CTCF	Caco-2					
CTCF	C					
CTCF	Dnd41					
CTCF	ECC-1					
CTCF	Fibrobl					
CTCF	Gliobla					
CTCF	GM06990					

CTCF	GM12801					
CTCF	GM12864					
CTCF	GM12865					
CTCF	GM12872					
CTCF	GM12873					
CTCF	GM12874					
CTCF	GM12875					
CTCF	GM12878					
CTCF	GM12878					

CTCF	GM12878					
CTCF	GM12878					
CTCF	GM12891					
CTCF	GM12892					
CTCF	GM19238					
CTCF	GM19239					
CTCF	GM19240					
CTCF	H1-hESC					
CTCF	H1-hESC					

CTCF	H1-hESC					
CTCF	HAc					
CTCF	HA-sp					
CTCF	HBMEC					
CTCF	HCFAa					
CTCF	HCM					
CTCF	HCPEpiC					
CTCF	HCT-116					
CTCF	HEEpiC					

CTCF	HEK293					
CTCF	HeLa-S3					
CTCF	HeLa-S3					
CTCF	HeLa-S3					
CTCF	HepG2					
CTCF	HepG2					
CTCF	HepG2					
CTCF	HepG2					
CTCF	HFF-Myc					

CTCF	HFF					
CTCF	HL-60					
CTCF	HMEC					
CTCF	HMEC					
CTCF	HMF					
CTCF	HPAF					
CTCF	HPF					
CTCF	HRE					
CTCF	HRPEpiC					


























CTCF	HSMMtube					
CTCF	HSMM					
CTCF	HUVEC					
CTCF	HUVEC					
CTCF	HUVEC					
CTCF	HVMF					
CTCF	IMR90					
CTCF	K562					
CTCF	K562					














































CTCF	K562					
CTCF	K562					
CTCF	K562					
CTCF	MCF-7					
CTCF	MCF-7					
CTCF	MCF-7					
CTCF	MCF-7					
CTCF	MCF-7					
CTCF	MCF-7					

CTCF	NB4					
CTCF	NH-A					
CTCF	NHDF-Ad					
CTCF	NHDF-neo					
CTCF	NHEK					
CTCF	NHEK					
CTCF	NHEK					
CTCF	NHLF					
CTCF	NHLF					

CTCF	Osteobl					
CTCF	ProgFib					
CTCF	RA					
CTCF	RA					
CTCF	RPTEC					
CTCF	SAEC					
CTCF	T-47D					
CTCF	WERI-Rb-1					
CTCF	WI-38					














































E2F1	HeLa-S3					
E2F4	GM12878					
E2F4	HeLa-S3					
E2F4	K562					
E2F4	MCF10A-Er-Src					
E2F6	HeLa-S3					
E2F6	K562					
E2F6	K562					
ELF1	A549					

ELF1	GM12878					
ELF1	HepG2					
ELF1	K562					
ELK1	GM12878					
ELK1	HeLa-S3					
ELK1	K562					
ELK4	HEK293					
ELK4	HeLa-S3					
ESR1	ECC-1					

ESR1	ECC-1					
ESR1	ECC-1					
ESR1	T-47D					
ESR1	T-47D					
ESR1	T-47D					
ETS1	A549					
ETS1	GM12878					
ETS1	K562					
FOS	GM12878					

FOS	HeLa-S3					
FOS	HUVEC					
FOS	K562					
FOSL1	H1-hESC					
FOSL1	K562					
FOSL2	A549					
FOSL2	HepG2					
FOS	MCF10A-Er- Src					
FOS	MCF10A-Er- Src					

FOS	MCF10A-Er- Src					
FOS	MCF10A-Er- Src					
FOXA1	A549					
FOXA1	ECC-1					
FOXA1	HepG2					
FOXA1	HepG2					
FOXA1	T-47D					
FOXA2	HepG2					
FOXP2	PFSK-1					

FOXP2	SK-N-MC					
GABPA	A549					
GABPA	GM12878					
GABPA	H1-hESC					
GABPA	HeLa-S3					
GABPA	HepG2					
GABPA	K562					
GATA1	K562					
GATA1	PBDEFetal					

GATA1	PBDE					
GATA2	HUVEC					
GATA2	K562					
GATA2	K562					
GATA2	SH-SY5Y					
GATA3	MCF-7					
GATA3	MCF-7					
GATA3	SH-SY5Y					
GATA3	T-47D					

HNF4A	HepG2					
HNF4A	HepG2					
HSF1	HepG2					
IRF1	K562					
IRF1	K562					
IRF1	K562					
IRF1	K562					
IRF3	HeLa-S3					
IRF3	HepG2					














































IRF4	GM12878					
JUND	GM12878					
JUND	H1-hESC					
JUND	H1-hESC					
JUND	HeLa-S3					
JUND	HepG2					
JUND	HepG2					
JUND	K562					
JUN	H1-hESC					

JUN	HeLa-S3					
JUN	HepG2					
JUN	HUVEC					
JUN	K562					
JUN	K562					
JUN	K562					
JUN	K562					
JUN	K562					
MAFK	H1-hESC					

MAFK	HeLa-S3					
MAFK	HepG2					
MAFK	HepG2					
MAFK	IMR90					
MAFK	K562					
MAX	A549					
MAX	GM12878					
MAX	H1-hESC					
MAX	HeLa-S3					

MAX	HepG2					
MAX	HUVEC					
MAX	K562					
MAX	K562					
MAX	NB4					
MAZ	GM12878					
MAZ	HeLa-S3					
MAZ	HepG2					
MAZ	K562					













































MYC	HUVEC					
MYC	K562					
MYC	K562					
MYC	K562					
MYC	K562					
MYC	K562					
MYC	K562					
MYC	K562					
MYC	MCF10A-Er- Src					

MYC	MCF10A-Er- Src					
MYC	MCF-7					
MYC	MCF-7					
MYC	MCF-7					
MYC	MCF-7					
MYC	NB4					
NFE2	GM12878					
NFE2	K562					
NFYB	GM12878					

NFYB	HeLa-S3					
NFYB	K562					
NR2C2	GM12878					
NR2C2	HeLa-S3					
NR2C2	HepG2					
NR2C2	K562					
NRF1	GM12878					
NRF1	H1-hESC					
NRF1	HeLa-S3					

NRF1	HepG2					
NRF1	K562					
PAX5	GM12878					
PAX5	GM12878					
PAX5	GM12891					
PAX5	GM12892					
PBX3	GM12878					
PRDM1	HeLa-S3					
RUNX3	GM12878					

RXRA	GM12878					
RXRA	H1-hESC					
RXRA	HepG2					
SPI1	GM12878					
SPI1	GM12891					
SPI1	K562					
SRF	GM12878					
SRF	H1-hESC					
SRF	HepG2					

SRF	K562					
STAT1	GM12878					
STAT1	HeLa-S3					
STAT1	K562					
STAT1	K562					
STAT1	K562					
STAT1	K562					
STAT2	K562					
STAT2	K562					

STAT3	GM12878					
STAT3	HeLa-S3					
STAT3	MCF10A-Er- Src					
STAT3	MCF10A-Er- Src					
STAT3	MCF10A-Er- Src					
STAT3	MCF10A-Er- Src					
STAT3	MCF10A-Er- Src					
TAL1	K562					
TBP	GM12878					

TBP	H1-hESC					
TBP	HeLa-S3					
TBP	HepG2					
TBP	K562					
USF1	A549					
USF1	A549					
USF1	A549					
USF1	GM12878					
USF1	H1-hESC					

USF1	HepG2								
USF1	K562								
USF1	RA								
USF2	GM12878								
USF2	H1-hESC								
USF2	HeLa-S3								
USF2	HepG2								
USF2	K562								
ZEB1	GM12878								

