PREDICTION OF ANTIMICROBIAL PEPTIDES BASED ON SEQUENCE ALIGNMENT AND SECONDARY STRUCTURE SEQUENCE AND SEGMENT SEQUENCE

SOH MENG WAH

UNIVERSITI SAINS MALAYSIA

2015

PREDICTION OF ANTIMICROBIAL PEPTIDES BASED ON SEQUENCE ALIGNMENT AND SECONDARY STRUCTURE SEQUENCE AND SEGMENT SEQUENCE

By

SOH MENG WAH

A Dissertation submitted for partial fulfilment of the requirement for the degree of Master of Science (Electronic Systems Design Engineering)

August 2015

ACKNOWLEDGEMENT

First of all, I would like to express my deepest sense of gratitude to my supervisor, Dr. Bakhtiar Affendi bin Rosdi for his advice with his extensive knowledge and his patience to guide throughout the project. As a part time student, this dissertation timeline is not been made easy for me. This research took place within one semester, i.e. 5 months of nights. Without Dr. Bakhtiar invaluable support and insightful suggestions, I wouldn't have reached the completion stage of this dissertation. Besides, I would like to thank my examiners, Dr Teoh Soo Siang and Dr Siti Noraini Sulaiman for their valuable feedback and suggestions on my research. Next, I would like to express my gratitude to my most wonderful friends and family. Their continuous love and encouragement are very strong support for me. Thanks to my company for supplying UNIX environment with powerful machines so that I can run multiple simultaneous simulations in shorter time. Finally, I would also like thank my employer and team mates for covering my job and not overloading me else I wouldn't have time to complete this dissertation.

TABLE OF CONTENTS

Acknowledgement ii

Table of Contents iii

List of tables vii

List of figures ix

Abbreviations x

Abstrak xi

Abstract xiii

CHAPTER 1 INTRODUCTION 1

1.1 Background 1

1.2 Problem Statement 2

1.3 Objectives 3

1.4 Project scope 3

1.5 Thesis Outline 4

CHAPTER 2 LITERATURE REVIEW 5

2.1 Introduction 5

2.2 Background of amino acid and AMPs 5

2.3 Previously proposed algorithm for prediction of AMPs 5

2.3.1 Sequence alignment method 6

2.3.2 Feature selection method by Wang et al. 2011 7

2.3.3 Machine pairwise algorithm 7

2.4 Feature extraction based on predicted secondary structure 8

2.4.1 Feature extraction method 8

2.4.2 Secondary structure sequence 9

2.4.3 Prediction of secondary structure of proteins 10

2.4.4 Segment sequence 10

2.5 Feature selection method 11

2.6 Classification by SVM (Support vector machines) 12

2.6.1 RBF kernel 13

2.6.2 Parameter selection 14

2.7 Statistical techniques for performance measurement 15

2.7.1 K-fold Cross validation 15

2.7.2 Jackknife test 15

2.7.3 AMPs accuracy Prediction 16

2.8 Summary 17

CHAPTER 3 METHODOLOGY 18

3.1 Introduction 18

3.2 Dataset 19

3.3 Platform 20

3.4 Proposed AMPs prediction algorithm 20

3.5 Phase 1: AMPs prediction using sequence alignment 22

3.6 Phase 2: AMPs prediction using secondary structure sequence and segment sequence feature 24

3.6.1 Secondary structure prediction 25

3.6.2 Feature extraction algorithm 25

3.7 Feature selection 30

3.8 Classification algorithm construction 31

3.9 Summary 33

CHAPTER 4 RESULT AND DISCUSSION 34

4.1 Introduction 34

4.2 Choosing the right scaling 34

4.3 Comparison between SSS and SS feature vectors 38

4.4 Influence of feature vectors 39

4.4.1 Comparison between 13 features and 18 features 39

4.4.2 Feature selection 40

4.5 Sequence alignment method (phase 1) 42

4.6 Classification of AMP using SSS and SS features (phase 2) 43

4.6.1 Jackknife test 43

4.6.2 Independent test 43

4.7 Comparison of proposed algorithm with previous researches 45

4.7.1 Sequence alignment method 45
4.7.2 Second phase algorithm only 46
4.7.3 Overall performance comparison 47
4.8 Analysis 49
4.9 Summary 52
CHAPTER 5 CONCLUSION 53
5.1 Conclusion 53
5.2 Areas of Future Work 54
REFERENCES 55
APPENDICES 59
Appendix A Script of sequence alignment using BLASTP 59
Appendix B Installation steps and scripts of secondary structure prediction 61
Appendix C Feature extraction script 63
Appendix D Feature selection script 68
Appendix E Jackknife test script 71

LIST OF TABLES

Page

Table 3.1 Dataset for this research 19

Table 3.2 Demonstration of features calculation for one protein sequence 29

Table 4.1 Grid search accuracy for 3 scaling of normal training set 36

Table 4.2 Grid search accuracy for 3 scaling of <0.7 similarity training set 37

Table 4.3 Best RBF kernel parameters 38

Table 4.4 Independent test accuracies based on 3 protein sets. 38

Table 4.5 Jackknife test accuracy for 13 features and 18 features 39

Table 4.6 (a) Feature selection test accuracies for FASTA normal training set 40

Table 4.7 Demonstration of BLAST accuracy on FASTA normal training set as independent test 42

Table 4.8 Jackknife test result of sequence alignment on FASTA <0.7 similarity training set 42

Table 4.9 Independent test sensitivity using sequence alignment method on 3 datasets 42

Table 4.10 Jackknife test comparison for before and after feature selection on 3 datasets 43

Table 4.11 Independent test prediction for before and after feature selection on full test sets 44

Table 4.12 Test prediction for before and after feature selection on 3 datasets on the remaining unpredicted sequences 44

Table 4.13 Performance comparison of independent test on sequence alignment method 45

Table 4.14 Performance comparison of jackknife test on sequence alignment method on <0.7 similarity training set 45

Table 4.15 Independent test accuracy based on second phase algorithm only 46

Table 4.16 Comparison between previous methods and this study on the test set based on Wang's normal training set (high similarity) 48

Table 4.17 Comparison between previous methods and this study on the test set based on Wang's <0.7 sequence similarity training set 48

Table 4.18 The performance comparison of CAMP methods with the proposed algorithm on independent test using CAMP test set. 48

Table 4.19 Difference of SSS prediction from different version of PSIPRED 50

LIST OF FIGURES

Page

Figure 2.1 An alpha helix structure (Griffiths et al, 2002) 9

Figure 2.2 A beta sheet (strand) structure (Griffiths et al, 2002) 9

Figure 2.3 The representation of E segments composing parallel beta-sheets or anti parallel beta sheets from the predicted secondary structural sequences. 11

Figure 2.4 Using a radial kernel to transform the input space to find the hyperplane in feature space with maximal marginal that separates the two classes. 13

Figure 2.5 An overfitting classifier and a better classifiers. 14

Figure 3.1 Flowchart of proposed algorithm for AMPs prediction 21

Figure 3.2 FASTA format representation of protein sequence 22

Figure 3.3 Flowchart of AMPs prediction with sequence alignment method 23

Figure 3.4 Flowchart of AMPs prediction algorithm with SSS and SS features 24

Figure 3.5 SVM input file format for one sequence 31

Figure 3.6 Flowchart of SVM training model generation and testing 33

Figure 4.1 Graph of highest test accuracy vs number of feature on normal training set 41

Figure 4.2 Reference chart of the optimal features selected by Wang et al. 2011. 50

ABBREVIATIONS

AC	Accuracy
AMP	antimicrobial peptides
BLAST	Basic Local Alignment Search Tool
BLASTP	Basic Local Alignment Search Tool for protein
FN	False negative
FP	False positive
HSP	high-scoring segment pairs
MCC	Matthews correlation coefficient
PSIPRED	Position Specific Iterated Prediction
SS	segment sequence
SSS	secondary structure sequence
SVM	Support vector machine
S_n	Sensitivity
S_p	Specificity
TN	True negative
TP	True positive

RAMALAN PEPTIDA ANTIMIKROBIAL BERDASARKAN PENJAJARAN URUTAN DAN URUTAN STRUKTUR SEKUNDER DAN URUTAN SEGMEN

ABSTRAK

Peptida antimicrobial (AMP) adalah sejenis peptide semula jadi yang penting untuk sistem imun. Penyelidik berminat untuk membuat ubat dengan AMP sebagai alternatif kerana bakteria semakin boleh menentang dengan antibiotik yang sedia ada. Walaubagaimanapun, eksperimen untuk mengekstrak AMP dari protein mahal dan mengambil masa. Oleh itu, alat pengiraan yang berkesan dan tepat meramalkan AMP baru amat dikehendaki untuk mengkaji ubat baru. Dalam projek ini, algoritma baru dicadangkan sebagai alat pengiraan dengan mengabungkan kaedah penjajaran urutan dan urutan struktur sekunder (SSS) dan urutan segmen (SS). Penjajaran urutan dilaksana berdasarkan HSPs maksimum skor yang diramalkan oleh BLASTP. Kaedah penjajaran urutan tidak dapat meramalkan semua urutan. Keputusan fasa penjajaran urutan adalah di 91.02 % bagi set data biasa, 80.88 % untuk urutan yang mempunyai persamaan <0.7, dan 96.02 % untuk CAMP set data. Bagi urutan yang tidak boleh diramalkan, ramalan diteruskan dengan menggunakan ciri-ciri SSS dan SS. Pengekstrakan ciri dan pilihan ciri dilakukan dan kemudian ciri-ciri tersebut digunakan untuk melatih pembelajaran mesin SVM bagi mengklasifikasikan urutan sama ada AMP atau bukan AMP. Keputusan ujian keseluruhan adalah 83.27% bagi set data biasa, 71.83% untuk urutan yang mempunyai persamaan <0.7, dan 91.49% untuk CAMP set data. Berbanding dengan fasa kedua kajian dulu yang menggabungkan dengan kaedah penjajaran jujukan, kajian ini mempunyai hasil yang rendah (<27%) dengan hanya menggunakan ramalan dengan SSS dan SS. Ini

menunjukkan bahawa algoritma baru yang dicadangkan tidak sesuai untuk digunakan sebagai peramal AMP.

PREDICTION OF ANTIMICROBIAL PEPTIDES

BASED ON SEQUENCE ALIGNMENT AND

SECONDARY STRUCTURE SEQUENCE AND SEGMENT SEQUENCE

ABSTRACT

Antimicrobial peptides (AMPs) are natural peptides that are important for immune system. Researchers are interested in designing alternative drugs with AMPs because more bacteria are becoming resistant to the available antibiotics. However, the experiments to extract AMP from protein sequences are time consuming and costly. Thus, a computational tool with more effective and accurately predicting novel AMPs is highly demanded to provide more candidates and useful insights for drug design. In this study, a new algorithm is proposed as a computational tool by integrating the sequence alignment method and the secondary structure sequence (SSS) and segment sequence (SS). The sequence alignment is accomplished by the classification of test sequences based on the maximum high-scoring segment pairs (HSPs) score predicted by Basic Local Alignment Search Tool for protein (BLASTP). The results of sequence alignment phase are in 91.02% for normal dataset, 80.88% on <0.7 sequence similarity train set and 96.02% for CAMP dataset. Sequence alignment method is not able to predict all sequences and the unpredicted sequences is then predicted by utilizing the SSS and SS features. Feature extraction and feature selection is performed to obtain the features. These features are used to train the SVM model which is then be used to classify the sequences to whether it is AMP or non-AMP. The overall results of independent test is 83.27% for normal dataset, 71.83% for sequence with <0.7 similarity dataset and 91.49% for CAMP dataset. In comparison of second phase with past research that combines with sequence alignment

method, this research has relatively low yield (<27%) contributed by the prediction utilizing SSS and SS features only. This indicates that the proposed algorithm is not suitable to be used as AMPs predictor.

CHAPTER 1

INTRODUCTION

1.1 Background

Mammals, reptiles, insects and plants, these organisms all produce antimicrobial peptides (AMPs) with broad spectrum antimicrobial activity to protect against microbial infection and survive in ever-changing environments. These microbial infection are mainly caused by bacteria, viruses and fungi. AMPs have been shown to be important in diverse functions as angiogenesis, wound healing and chemotaxis (Sang & Blecha, 2008). The AMPs are able to alter the host immune response through receptor-dependent interactions. Once these conserved AMP in a target microbial membrane, the peptide kills target cells through diverse mechanisms. Decreased levels of these peptides have been noted for patients diagnosed with atopic dermatitis and Kostmann's syndrome, a congenital neutropenia (Izadpanah, 2005).

The usage of AMPs has motivated researchers to explore this alternative as substitute for conventional antibiotics. Researchers are interested in designing alternative drugs based on AMPs because they have found that a large number of bacterial strains have become resistant to the available antibiotics (Epand & Vogel, 1999). The AMPs drugs could be used for antibacterial, antifungi, antivirus, and even anticancer which are less likely to induce resistance. However, researchers have encountered obstacles in the AMPs designing process such that the experiments to extract AMPs from protein sequences are costly and require a long set-up time (Hadley & Hancock, 2010). Therefore, a computational tool for AMPs prediction is needed to resolve this problem by predicting the AMPs sequence.

1.2 Problem Statement

There are many computational tools have been introduced to predict AMPs used by past researches such as sequence alignment method and feature selection method (Wang et al., 2011). There are also focus on screening and in silico modeling novel AMPs (Pestana-Calsa, 2010; Hadley EB, 2010) as antimicrobial drug discovery and design can be accelerated with this computational approaches. Bioinformatics methods like APD method (Wang Z & Wang G, 2004) is developed to predict if the new peptides had the potential to be antimicrobial. Hidden Markov models are constructed to automatically discover AMP, known as AMPer method (Fjell, 2007). Other computational methods such as AntiBP (Lata & Sharma, 2007) and AntiBP2 (Lata & Sharma, 2010) which their peptides are limited to N and/or C terminus residue. These 2 computational methods use Artificial Neural Network (ANN), Quantitative Matrices, and Support Vector Machine (SVM) to predict the AMPs. CAMP method (Thomas et al., 2010) was developed based on Random Forest, SVM and Discriminant Analysis, trained on all classes of AMPs with full length of mature AMP sequences. Nonetheless, these methods do not have the capability to identify which features are optimal for accurately predicting AMPs and interpreting the biological implication meaningfully.

There are several algorithms have been developed by incorporating the secondary structure elements for bioinformatics applications such as protein structural class prediction. Promising results have been reported (Chou & Chai, 2004) with more than 90% overall jackknife success rate based on a low-similarity dataset designed by the authors for protein structural prediction into seven classes, with the consideration of the fact that proteins in same structural class is likely to have high similarity in their corresponding secondary structural elements. Feature extraction has been shown accurate prediction for protein structural classes (L. Kong et al., 2014) by utilizing

secondary structure sequence (SSS) and segment sequence (SS) information. However, this feature extraction method has not been applied in the AMPs prediction study. Since AMP is a family of protein peptides, and previous research has shown successful case in protein class structural prediction, hence it might have great potential for AMPs prediction using the SSS and SS features. New AMPs usage in drugs, but with costly and time consuming experiment, yet lacking of study in predicting AMP using secondary structure in computational tool lead to the interest of the study in this project.

1.3 Objectives

The main objective of this research is:

1) To study the weightage of SSS and SS influence in AMPs prediction.

The supporting objectives are:

1) To develop an algorithm to predict if a peptide is an AMP or non-AMP from the pool of protein peptides based on the combination of sequence alignment method and secondary structure sequence (SSS) and segment sequence (SS) using feature extraction and feature selection methodology,
2) To compare the performance of this research with previous researches using standard statistical analysis method.

1.4 Project scope

This research focuses on the prediction of AMPs sequences based on primary sequences, secondary structure sequences and segment sequences. To make comparison meaningful, the database used is the same as previous research (Wang et al., 2011, Xin Yi et al., 2014). In implementation stage of this research, the sequence alignment method and the feature extraction method are used to predict the AMPs. The sequence alignment method needs primary sequence as input and it is not able to

predict all the AMPs sequences due to having 0 score in HSP score. The remaining unpredictable primary sequences will then be converted into secondary structure sequence from the prediction of BLAST. The secondary protein sequences will be used for the AMP feature extraction to form a feature vector and been analyzed for classification by support vector machine (SVM). The AMP prediction process including sequence alignment stage and feature extraction process until classification stage is carried out in Perl programming language in UNIX environment. Finally, the analysis of the prediction accuracy was done based on a statistical analysis technique named jackknife cross validation. The performance of the proposed algorithm is evaluated and is compared with previous research.

1.5 Thesis Outline

This dissertation is divided into 5 chapters. Chapter 1 gives an overview and motive of this research. Chapter 2 gives an overview of the concept of AMP primary protein structure in terms of amino acid, secondary structure and discusses previous researches methodologies. Chapter 3 describes how this research was conducted with the proposed algorithms. Chapter 4 discusses the results obtained and analysis of it. Chapter 5 would conclude this research. Possible future work is suggested in the final chapter.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter gives an overview of AMP, discuss the past researches algorithm and how the past researches methodology can be modified and adopted in this research. The standard performance measurement is mentioned in this chapter as well.

2.2 Background of amino acid and AMPs

There are 20 major out of 500 types of amino acids used in known biological life. Of the 20, each of the amino acid element is represented by letter code A-Y except B, J, O, U, X, Z which is the non-standard residues (Biochemical Compounds Declarative Database). These elements of amino acid in a structural sequence will then form a primary protein sequence or peptides. AMP is the antimicrobial peptide which essentially contains the sequence of amino acids. Amino acid composition is closely related with its attributes, such as subcellular location (Chou & Elrod, 1999), folding type (Nakashima et al., 1986), domain (Dumontier et al., 2005) and secondary structure content (Lee S et al., 2006).

2.3 Previously proposed algorithm for prediction of AMPs

The primary protein structure attribute analysis has been used to predict AMP in the past by using sequence alignment method and feature selection method (Wang et al., 2011) or machine – pairwise algorithm (Xin Yi et al., 2014) which will be reviewed in section 2.3.1, 2.3.2and 2.3.3.

2.3.1 Sequence alignment method

Sequence alignment method predicts the protein sequences by assigning it to the category which has highest sequence similarity. There are many software have been developed to categorize the nucleotide or protein sequences which some of them are based on either database search only (BLAST (Lipman & Pearson,1985), FASTA (Altschul et al., 1990), HMMER (Eddy, 1998) and Smith-Waterman algorithm (Smith, 1981), pairwise alignment (SWIFT suit (Rasmussen et al., 2006), CUDAlign (Sandes et al., 2013)), or multiple sequence alignment (eg. PROMALS3D (Jimin ,2008)).

BLAST (Basic Local Alignment Search Tool) is a popular tool to compare the query proteins or nucleotide sequence with the target database to identify the regions of local alignment and report out the alignments that score above threshold score. The hit of 1 or more high-scoring segment pairs (HSPs), between the training sequence set and the test sequence, is the BLAST threshold score. It would return a zero if there is no hits of pairing segments between the test sequence and the training set and hence cannot be predicted by BLAST.

BLASTP sequence alignment method was implemented in past researches (Xin Yi et al., 2014, Wang et al. 2011) to predict AMPs. The AMP sequences used in their researches are in protein based form, hence BLASTP was used (BLASTN is nucleotide BLAST; BLASTP is protein BLAST). As some of the test result of BLAST returns zero with no prediction, the remaining unpredictable sequences are further predicted by with machine pairwise algorithm (Xin Yi et al., 2014) or with feature selection method (Wang et al., 2011).

2.3.2 Feature selection method by Wang et al. 2011

Feature selection has been introduced by Wang et al to predict the unpredicted remaining sequences from sequence alignment method. In their research of using feature selection method for AMPs prediction, a peptide sequence is defined in a vector space which is coded by amino acid composition and pseudo amino acid composition method. Maximum Relevance Minimum Redundancy (mRMR) method (Peng et al., 2005) is used to prioritize the features in the vector space. The Nearest Neighbor Algorithm (NNA) (Friedman et al., 1975; Chou and Shen, 2007) is used to construct the prediction model based on the optimal feature subset that is selected by Incremental Feature Selection (IFS) method (Huang et al., 2009; Huang et al., 2010).

2.3.3 Machine pairwise algorithm

This concept of machine pairwise algorithm was introduced with the aim of detecting remote protein evolutionary and protein structural relationships (L. Liao, & Noble 2003). The machine pairwise similarity score has been reported produced from BLAST for protein sequence (L. Liao, & Noble 2003, Muh et al., 2009) or Lempel-Ziv (LZ) complexity algorithm for AMPs (Xin Yi et al., 2014). BLAST is used to generate the pairwise similarity scores of protein test sequence against training set. LZ complexity algorithm generates the similarity score by computing the complexity-based distance measure of the AMP sequences and often related to the steps that are required to build a sequence.

2.4 Feature extraction based on predicted secondary structure

L.Kong 2014 and J. Wang 2014 has shown successful prediction rate in bioinformatics application on protein structural class prediction. They extracted information such as spatial alignment, maximum segment length, secondary order composition moment, degree of segregation, etc from the secondary structure to form SSS and SS. The prediction accuracy is >80%. Seeing the high prediction rate on the application of protein structural class prediction and never been applied in AMP prediction, the method is proposed as the second phase of prediction in this project. The basic concept of their method is reviewed in the following sub chapters, and the details of the algorithm is discussed in Chapter 3 during the implementation process.

2.4.1 Feature extraction method

In machine learning, feature extraction starts from an initial set of measured data and build derived values so that the feature is informative, non-redundant, reflects the component of the dataset so that it can facilitate the machine learning. This method extract features in a format supported by machine learning algorithms such as SVM from datasets consisting of formats such as text and image. A predictor variable is called an attribute, and a transformed attribute that is used to define the machine language hyperplane is called a feature. A set of features that describes one case (i.e., a row of predictor values) is called a vector.

In this AMP prediction research context, the dataset is the strings of text consisting E (beta strands), H (alpha helix) and C (coil) that form the secondary sequence of the protein sequence. Then the string of secondary structure sequence is then been analyzed to get their features which is hope to be important features that can help to reflect the characteristic of the protein and categorize it into either AMP or non AMP.

2.4.2 Secondary structure sequence

Two main types of secondary structure of proteins are alpha helix (Figure 2.1) and beta strand (Figure 2.2) (Pauling et al., 1951). These structures are defined by patterns of hydrogen bonds between the main-chain peptide groups and have regular geometry. An unfolded polypeptide chain that lacks of 3D structure is a random coil (C). 4 traditional secondary AMP structures are alpha helices (H), beta strands (E), loop structures and extended structures (H. Jenssen et al., 2006; L.T.Nguyen et al., 2011) and random coil. AMPs can be classified to 4 secondary structures, but to the best of our knowledge, there's no published research on the prediction of AMPs from its secondary structure.

Secondary structure sequence was used to predict the protein structural classes but not in AMP yet. Protein structural classes prediction has been demonstrated by combining with PSI-BLAST profile (Shuyan Ding et al., 2014) or pseudo amino acid PseAA structural properties (J. Wang et al., 2014).The content and spatial arrangements of the secondary structural elements of a given protein sequence were used as the feature to perform the prediction.

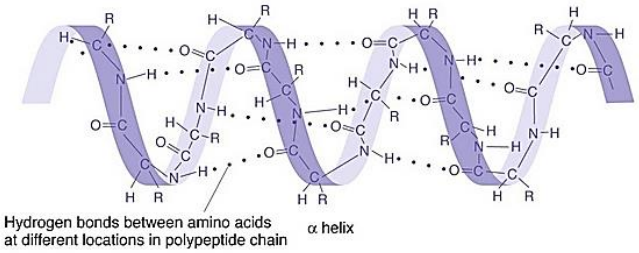

Figure 2.1 An alpha helix structure (Griffiths et al, 2002)

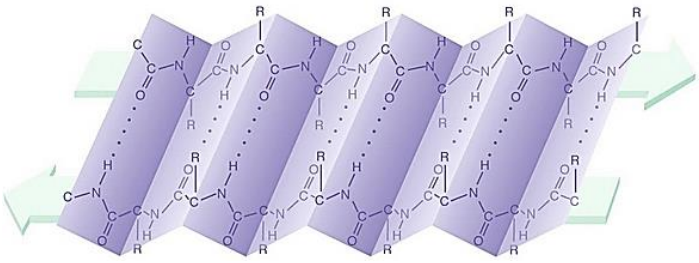

Figure 2.2 A beta sheet (strand) structure (Griffiths et al, 2002)