ABSTRACT

Machine Learning-Assisted Prediction of Structure and Function of Cystine-Stabilized Peptides and Optimization of Expression in an *E. coli* System

S M Ashiqul Islam, Ph.D.

Co-Mentor: Christopher M. Kearney, Ph.D.
Co-Mentor: Erich J. Baker, Ph.D.

Cystine-stabilized peptides are promising prospects for the pharmaceutical industry as biologics. These peptides carry out a variety of useful functions which could be exploited to treat diseases and kill unwanted organisms. As well, an array of disulfide bonds makes the peptides highly stable against temperature, enzymatic degradation, pH and other adverse physiological conditions. There is a vast number of cystine-stabilized peptides serving as antimicrobial peptides, immunological modulators, ion channel blockers and other functions across a wide array of taxa, from fungi and bacteria to plants and humans. Practical access to these promising bioactive molecules could be greatly accelerated if it were possible to efficiently mine cystine-stabilized peptide sequences from genomic databases, determine the function and structure of each candidate from only the primary sequence, and then express the top candidates in *E. coli* for biological analysis. In this way, only the natural, presumably functional, variants of a particular family of cystine-stabilized peptides could be collected in large quantities. Going further,

it would be desirable to convert the nonspecific activity of antimicrobial peptides to a specific activity, targeting a specific pathogen and leaving the rest of the microbiome intact; in essence, developing a targeted antibiotic.

To contribute to developing this pipeline, I developed the machine learning-assisted algorithms PredSTP and CSPred to predict structural and functional characteristics, respectively, of cystine-stabilized peptides from primary sequence data. In addition, I developed an *E. coli*-based expression system for high yield production of recombinant antimicrobial peptides specifically targeted to *Staphylococcus aureus*. These techniques are now available to collect large libraries of cysteine-stabilized peptide sequences, to express top candidates in *E. coli*, and to target the peptides to specific pathogens.

Machine Learning-Assisted Prediction of Structure and Function of Cystine-Stabilized peptides and Optimization of Expression in an *E. coli* System

by

S M Ashiqul Islam, B.S., M.S.

A Dissertation

Approved by the Institute of Biomedical Studies

_____

Robert R. Kane, Ph.D., Director

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of

Doctor of Philosophy

Approved by the Dissertation Committee

_____

Christopher M. Kearney, Ph.D., Co-Chairperson

_____

Erich J. Baker, Ph.D., Co-Chairperson

_____

Bessie W. Kebaara, Ph.D.

_____

Sung Joon Kim, Ph.D.

_____

Kevin G. Pinney, Ph.D.

Accepted by the Graduate School

May 2018

_____

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF AlGORITHMS

LIST OF ABBREVIATIONS

A12C	Targeting domain from filamentous phage
AC	Auto correlation
ACRI	Acetylcholine receptor inhibitors
AMP	Antimicrobial peptides
ANN	Artificial Neural Network
ANOVA	Analysis of variance
APC	Amino acid composition
APD	Antimicrobial Peptide Database
AUC	Area under curve
BLAST	Basic Local Alignment Search Tool
BPP	Binary profile pattern
CaB	Calcium channel blocker
CTD	Composition Transition Distribution
Ctraid	Conjoint triad
DA	Discriminant Analysis
DPC	Dipeptide composition
ESI	Electrospray ionization
FN	False negative
FP	False positive
GPCR	G protein-coupled receptors
HCV	Hepatitis C virus
HIV	Human immunodeficiency virus
HLP	Hemolytic proteins
HMM	Hidden Markov Models
ICB	Ion channel blocker

ICK	Inhibitor cystine knot
IPTG	Isopropyl β-thiogalactopyranoside
KB	Potassium channel blocker
KNN	K-Nearest Neighbors
LB	Luria-Bertani
LC	Liquid chromatography
LR	Logistic regression
MCC	Matthews correlation coefficient
MIC	Minimum inhibitory concentration
M-NGSG	Modified n-grams and skip-grams
MS	Mass spectrometry
NaB	Sodium channel blocker
NLP	Natural Language Processing
NMR	Nuclear magnetic resonance
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate buffered saline
PDB	Protein data bank
PPI	Protein-protein interaction
PseAAC	Pseudo amino acid composition
PSI-BLAST	Position-Specific Iterative basic local alignment search tool
QSAR	Quantitative structure–activity relationship
QSO	Quasi-sequence-order
RF	Random Forest
ROC	receiver operating characteristic
SDS	Sodium dodecyl sulfate
SPI	serine protease inhibitors
STP	Sequential tri-disulfide peptide
SUMO	Small Ubiquitin-like Modifier

SVM	Support vector machine
TN	True negative
TP	True positive
TPC	Tripeptide composition
TSB	Tryptic-Soy Broth
ULP	Ubiquitin like protein

ACKNOWLEDGMENTS

At the very beginning, I want to be grateful to the God for bringing me to a successful end of my Ph.D. The long journey would not be possible without the help of the Almighty.

Next, I would like to express my gratitude to Dr. Christopher Michel Kearney. As my principal investigator, he gave enormous support to my research, thinking, planning and writing. He created all possible opportunities for me which helped to bring the best from me. Thank him for his faith in me and my work during the time. Beyond the academics, he also kept me under his support just like a parent.

I also want to give my special thanks to Dr. Erich J. Baker who played the most significant role in my published papers. I am quite grateful to him for regularly giving me time to teach scientific writing. Also, I received crucial directions from him to learn and apply computational biology on my research projects.

I want to give thanks to Dr. Bessie Kebaara and Dr. Sung Joon Kim who gave me essential guidance on my research throughout the time. I am also thankful to Dr. Kevin Pinney for being a member of my defense committee. I also gratefully appreciate the help from Sandra Harman for giving me her time on formatting the dissertation. I also give my thanks to all my lab members and the whole Baylor family who were exceptionally friendly and helpful to me throughout the journey.

I am quite grateful to my elder brother Nasimul Islam who continuously patronizes and encourages me towards success at each step in my life. Finally, I am quite

grateful to my wife Tarana Mou for being quite caring and romantic since we got married.

DEDICATION

Remembering
Azizul Islam and Khaleda Islam
who gifted me their priceless
chromosomes

CHAPTER ONE

Introduction

Cystine-Stabilized Peptides

Although proteins are the building blocks of life, sometimes they can be toxic to a living organism. However, this role of proteins can be made beneficial to mankind. Toxic peptides can be used as a biological toxin, particularly against insects and pathogenic microorganisms. The discovery stream of chemical insecticides and antibiotics has dried up even as resistance is building up in microbial pathogens (Hemingway and Ranson 2000; Brooke et al. 2002; Aloush et al. 2006; Hiramatsu et al. 1997). Moreover, in most cases, chemical leads that are toxic to insects and microbial pathogens are also harmful to humans. Under these circumstances, it is quite essential to find organic alternatives, namely, naturally occurring protein toxins, which have the additional property of typically not promoting pathogen resistance and often show low toxicity to humans (Whetstone and Hammock 2007).

Small cystine-stabilized peptides can be useful candidate protein toxins. The biological functions of cystine-stabilized peptides are varied, and they include ion channel blockers, antimicrobial peptides, acetylcholine receptor inhibitors and serine protease inhibitors. A substantial fraction of these proteins is also cytotoxic or hemolytic, indicating human toxicity. Such sequences need to be screened out before proceeding with any drug development of a group of cystine-stabilized peptides. Structurally, these cystine-stabilized peptides include inhibitory cystine knot (ICK) folds (Zhu et al. 2003),

cyclotides, anti-parallel beta-sheet folds, alpha helix-beta sheet folds, and β-hairpin folds stabilized by disulfide bonds. Generally, the length of these small peptides is around 50 amino acids (Jérôme Gracy and Chiche 2011). In most cases, the fold is stabilized by three disulfide bonds and maintains a sequential pattern of bonding: cys#1 bonding with cys#4, cys#2 with cys#5 and cys#3 with cys#6. We have named this highly stable structural group as sequential tri- disulfide peptides or STPs (Mishu 2015 ref). However, there are also functionally homologous cystine-stabilized peptides which do not follow the STP bonding pattern.

Structure-Based Subfamilies of Cystine-Stabilized Peptides

Knottins

Knottins or ICK (Inhibitor cystine knot) are the most known type of the cystine stabilized peptides. These peptides have the canonical STP bonding pattern (C1-C4, C2-C5, C3-C6). However, while folding, the third disulfide bond penetrates through the first two disulfide bonds creating a cystine knot. This group of the peptides is thus a subset of the STP group. ICK peptides have been commercially developed and have different functional characteristics and medical/agricultural applications, such as neurotransmitters, analgesics, anthelmintics, anti-erectile dysfunction, antimalarials, antimicrobials, antitumor agents, protease inhibitors, toxins and insecticides (J. Gracy et al. 2007). To offer a proper representation and documentation of the functional annotation and bibliographic data, a well-curated database KNOTTIN is dedicated to ICK peptides (Postic et al. 2018). Protein data bank and Uniprot also included “knottin” as a structural motif in their database. In addition, tools to predict knottins from the primary

and 3D structure of proteins are included in the database as Knotter 1D and 3D, respectively.

Cyclotides

Cyclotides are a subset of knottins which have a head to tail cyclic backbone. The combination of cystine knot and cyclic backbone is also known as a cyclic cystine knot. Kalata B1, from the plant *Oldenlandia affinis* (*Rubiaceae*), was the first cyclotide described and was experimentally confirmed as containing a cystine knot with the macrocyclic structure in 1995. However, the peptide Kalata B1 was discovered back in the early 1970s as an active ingredient of uterotonic, which was used as a boiled tea to accelerate childbirth (Saether et al. 1995). The harsh means of preparation of the uterotonic remedy revealed the stability of Kalata B1, which was later found to be highly resistant to high temperatures and digestive enzymes (Saether et al. 1995). Since then, cyclotides have been found to offer functionally diverse attributes as defense peptides such as insecticidal (C. Jennings et al. 2001; C. V. Jennings et al. 2005), nematicidal (Colgrave, Kotze, Huang, et al. 2008; Colgrave, Kotze, Ireland, et al. 2008; Colgrave et al. 2009; Malagón et al. 2013), and molluscicidal (Malagón et al. 2013) peptides being represented. Cyclotides can be formally divided into three groups: the Möbius, the bracelet, and trypsin inhibitor subfamilies. Kalata B1, Cycloviolacin O2, and the inhibitor MCoTI-II are the specific examples the group möbius, bracelet and trypsin inhibitor, respectively. The bracelet and Möbius cyclotide folds are similar to each other, with the main difference being in Loop 5, where a *cis* tryptophan–proline bond results in a 180º twist of the peptide backbone Ω-angle of the Möbius fold resulting from a *cis* tryptophan–proline bond (Rosengren et al. 2003) . On the other hand, trypsin inhibitors

have entirely different peptide sequences from the other groups with a more extended sequence in Loop 1. The coffee and violet plant families, *Rubiaceae* and *Violaceae* are the source of the Möbius and Bracelet type of cyclotides, respectively (Koehbach, Attah, et al. 2013; Koehbach, O'Brien, et al. 2013; Simonsen et al. 2005), while the third type, trypsin inhibitors, were isolated from the melon family, *Cucurbitaceae* (Avrutina et al. 2005; Mylne et al. 2012). A few cyclotides are also found in the bean, potato and grass families; *Fabaceae*, *Solanaceae*, and *Poaceae*, respectively (Nguyen et al. 2011, 2012, Poth et al. 2011, 2012). To facilitate the search and display of cyclic proteins for functional and structural analysis, a database named Cybase is dedicated to the cyclic proteins (Wang et al. 2007). A significant fraction of the documented proteins in Cybase are cyclotides, or other cystine-stabilized cyclic peptides such as theta defensin RTD-1(Conibear et al. 2012) and cyclic bacteriocin(González et al. 2000). As of January, 2018, A total of 314 cyclotides were documented in Cybase. However, it is conservatively accepted that a significant fraction of the cyclotides is undiscovered. CyPerl and CyExcel are two BLAST independent tools invented for the prediction of cyclotide analogs from plant genome and other protein databases (Zhang et al. 2015). A total of 202 novel cyclotide analog were harvested from seven different plant families using the CyPerl and CyExcel tools. Another machine leaning-based sequence alignment-independent predictor, named Cypred, is available for cyclic peptides. Utilizing a test set, Cypred offered a 98.7% percent accuracy which was better than the accuracy calculated using BLAST and pairwise sequence alignment on the same test set. CycloMod is another tool to predict the 3D structure of a putative cyclic protein as a PDB format.

Source-Based Subfamilies of Cystine-Stabilized Peptides

Spider and Scorpion Toxins

A large fraction of cystine-stabilized bioactive proteins consists of spider venom proteins. Spiders use their toxins to kill or paralyze their preys via interference with the neurotransmission process. Besides general neurotoxicity, spider toxins also display antiparasitic, hemolytic, analgesic, cytolytic, antimicrobial, antiarrhythmic, and enzyme inhibitory activity (Saez et al. 2010). The primary mode of actions of these toxins are interfering or binding with transporters, receptors, and carbohydrate of lectins, perturbing membrane and inhibiting different ion and other channels (Lewis and Garcia 2003; Liang and Pan 1995) . Arachnoserver (Herzig et al. 2011) is a spider toxin database which contains a repository of 1426 spider toxin (to date) those are categorized based on taxonomy, molecular targets, post-translational modifications, and phyletic specificity. It is assumed that only a small fraction of the spider toxins has been discovered, and a significant portion of the listed toxins are unexplored experimentally. While there is no computational tool to detect spider toxins from the primary sequence, SpiderP is an available tool to predict the subcellular location of spider toxins using support vector machine (Henrik Nielsen 2017). Another significant source of cystine-stabilized peptides are scorpion toxins. They are primarily divided into two categories: long-chain and short-chain toxins. Most of the scorpion toxins interact with voltage-gated sodium and potassium channels. While there is a database named SCORPION2 (Tan et al. 2006) was reported for scorpion toxins, the server has not been active recently. However, the same research group has developed a tool to predict functional properties of scorpion toxins from the primary sequences (Tan et al. 2005).

Conotoxins

Cone snails are another big source of the cystine-stabilized proteins, producing a wide array of toxin proteins in their venom gland (Gao et al. 2017). These are designated "conotoxins" and, as a group, they are primarily bioactive neurotoxins which are mainly divided into three subgroups: (1) voltage-gated ion channel blockers, (2) ligand-gated ion channel blockers, (3) other receptor blockers (Williams et al., 1992). Members of the third subgroup interact with neurotensin receptors, nicotinic acetylcholine, or G protein-coupled receptors (GPCRs) primarily. To facilitate the annotation procedure, signatures of different conotoxin families have been adopted at PROSITE which is a database for protein domain annotation (Sigrist et al. 2010). ConoServer (Kaas et al. 2012) is a database constructed with the sequence, structure, and functional characteristics of conotoxins. As of January 2018, the number of entries in ConoServer consisted of 2838 nucleotide sequences, 6255 protein sequences, and 176 protein structures. The Conus Biodiversity website (http://biology.burke.washington.edu/conus/) is a phylogeny database which keeps a record of pictures and videos of different *Conus* species, and this record is referenced by ConoServer. Although a huge number of conotoxins are recorded in the database, it commonly agreed that a significant fraction of conotoxins is still unexplored. To expedite the discovery, a number of the machine learning based algorithms have been reported which predict putative conotoxins from unknown primary protein sequences. In 2006, Mondal et al. proposed the first-ever conotoxin prediction model where features generated from PseAAC were used to train an SVM classifier with an 88.10% accuracy (Mondal et al. 2006). Later in 2007, an IDQD model that subclassifies conotoxins into superfamilies and families was developed with an accuracy

of 87.7% and 72%, respectively (Lin and Li 2007). Subsequently, the prediction and identification of conotoxins which function as ion-channel inhibitors have become quite popular, and a number of machine-learning based algorithm using diverse feature extraction techniques from the primary protein sequences have been proposed within this scope (Wu, Zheng, and Tang 2016; Xianfang et al. 2017; Yuan et al. 2013; Zhu et al. 2003). Presence of these prediction models accelerates the identification of undiscovered conotoxins and the rapid extension of the conotoxin sequence coverage in the related databases.

Plant Cysteine Rich Peptides

Plants serve as cystine-stabilized peptide factories, producing an enormous number of variations, which are primarily used for defensive purposes (Tam et al. 2015). These peptides include plant defensins (Stotz, Thomson, and Wang 2009), hevein-like peptides (Porto et al. 2012), crambins (Teeter, Mazer, and L'Italien 1981), lipid transfer proteins (García-Olmedo et al. 1995), knottin-like proteins (Rees and Lipscomb 1982) and snakins (Segura et al. 1999). Several of the plant pathogenesis-related proteins are cysteine rich peptides. Pathogenesis-related proteins were first discovered in the early 1970s from tobacco leaves. These proteins include diverse mechanisms, such as antiviral, antifungal, antibacterial, chitinase, anti-oxidative activity, and proteinase inhibitory activities (Sels et al. 2008; Sinha et al. 2014; Stintzi et al. 1993). For example, cyclotide-type proteins in plants inhibit the proteases, lipid transfer proteins bind to lipids and inhibits microbial infection into the cell membrane, and havein-like peptides bind to chitins to defend the source plant from fungal infections. Based on some common features extracted from thousands of plants genes, it is expected that the number of

cystine-stabilized proteins in plants is under-predicted (Silverstein et al. 2007). PhytAMP is a database of plant AMPs where a portion of the recorded proteins are plant cysteine rich proteins (Hammami et al. 2009). PhytAMP database contains information on the family, source organism, activity and target organisms for each of the total 273 entries. Although there is no dedicated computational tool to discover plant cysteine rich proteins from the primary sequence to date, an *in silico* method is available to predict hevein-like peptide precursors from the plant genome (Porto et al. 2012).

Targeted Antimicrobial Peptides

In addition to previously discussed sources, cystine-stabilized peptides are also found in bacteria, fungi, sea anemones, jellyfish, centipedes, cephalopods, echinoderms, snakes, lizards, fish, platypus and arguably even fleas, mosquitoes, kissing bugs, leeches, ticks, and vampire bats (Fry et al. 2009). Bacteriocins are one group among the peptides which are produced by bacteria to kill other, competing bacterial genera. A substantial fraction of these consist of peptides is cystine-stabilized such as Laterosporulin (Singh et al. 2015) and thuricin CD (Sit et al. 2011). Bacteriocins are becoming popular as antimicrobial peptides (AMP) with growing infection caused by antibiotic-resistant bacteria. One advantage of the using bacteriocins over other AMPs is the bacteriocins are very specific to their target species, producing a minimal effect on commensal microbes.

Defensins are another group of cystine-stabilized peptides and are found in vertebrates, invertebrates, and plants. These are mainly antimicrobial peptides active against fungi, bacteria, protists, and viruses. “Defensin Knowledgebase” is a database which contains the record of the family, structure, target organism, and the strength of activity for the 566 defensins recorded to date. This database also includes the clinical

records of 255 defensins. Unlike bacteriocins, defensins demonstrate activity against a broader range of bacteria which often include commensal microbes. However, a few fungal defensin have been engineered to be targeted by attaching species-specific targeting domains. For example, plectasin (Mygind et al. 2005) is a defensin from a fungal species of the order *Pezizales* which was targeted to selectively kill methicillin-resistant *Staphylococcus aureus* (MRSA) when the peptide ArgD from the quorum sensing system was genetically fused to the N-terminal sequence of the defensin peptide. Islam et al. (unpublished) have produced a targeted plectasin (Mygind et al. 2005) and eurocin (Oeemig et al. 2012) which are selectively targeted against *Staphylococcus* species. In this case, the host binding protein from bacteriophage A12C was genetically fused to these defensins.

Importance of Prediction and Production of Cystine-Stabilized Peptides

The cystine-stabilized peptides have the potential to be used as peptide drugs due to the important functional and structural properties discussed above. It is also widely accepted that disulfide-rich cystine-stabilized peptides are naturally expressed by a broad range of organisms. Thus, a rich bank of natural, presumably fully functional, cystine-stabilized peptide variants is present, in hidden form, in the published genomic databanks. However, only a small fraction of these peptide sequences has been revealed and characterized because of the heterogeneity of their primary sequence, making sequence alignment algorithms such as BLAST effective only at expanding islands of sequence space surrounding previously characterized cystine-stabilized peptide sequences (Lu et al. 2006; Roth et al. 2002; Ryan and Sandell 1990). This phenomenon creates a high noise vs signal ratio which makes it difficult to predict cystine-stabilized

toxins using their primary sequences directly from the genome of an organism. In addition, there are challenges associated with physically producing peptides once the appropriate sequence is discovered from a databank set. Chemical synthesis generally cost effective for large scale screening only of smaller peptides. On the other hand, expression of cystine-stabilized peptide sequences in *E. coli* has been problematic historically, but for screening or commercially producing larger peptides, *E. coli* expression is crucial. Fortunately, new expression systems have allowed for routine expression of cystine-stabilized peptides in *E. coli*, as discussed later. This is especially important for the production of the larger, targeted versions of cystine-stabilized peptides and also for the exploration and testing of cystine-stabilized peptides that are designed to be expressed transgenically in plants or animals.

Recombinant Protein Expression Methods

Bacterial Expression Systems

E. *coli* expression is a well-studied and rapid production system. It is relatively cheap, easy to control, and has a high diversity of vectors and strains (Shatzman and Rosenberg 1987; Talmadge, Kaufman, and Gilbert 1980; Chevallier and Aigle 1979). It could be challenging to express cystine-stabilized peptides in an *E. coli* system because the expression system is not as optimal for proper folding of the disulfide-rich proteins as are eukaryotic systems. However, the use of small ubiquitin-like protein (SUMO) has provided a solution to this folding dilemma and SUMO has been routinely used as a fusion partner to express proteins which are difficult to produce in *E. coli* system (Butt et al. 2005). SUMO facilitates its payload partner to fold properly which makes the whole

fusion protein more soluble. The protein of interest can also be separated from the SUMO fusion partner quite conveniently using SUMO protease, which very specific to the C-terminal structure of the SUMO protein. The *E. coli* expression system with a SUMO fusion partner makes it routinely possible to express the cystine-stabilize peptides to study their functional and structural characteristics along with other downstream interests.

Yeast Expression Systems

Yeast is relatively cheap and easy to manipulate compared to other eukaryotic systems. This system allows recombinant proteins to be secreted to facilitate purification, provides high protein expression in bioreactors, and provides for eukaryotic posttranscriptional modifications like disulfide bond formation, glycosylations and protein maturation (Daly and Hearn 2006). Expression of a number of AMPs has already been reported in yeast expression systems (Cipáková and Hostinová 2005; Schoeman et al. 1999). As a fermentative yeast, *Saccharomyces cerevisiae* uses its carbon for ethanol production resulting in lower biomass and lower production protein (Mattanovich et al. 2012) . *Pichia pastoris* provides relatively better expression but retains problems with codon usage and undesired posttranslational modifications (Cereghino et al. 2002).

Plant Expression Systems

Plant expression systems are scalable and can handle the posttranscriptional modification complexity of heterogeneous proteins (Peters and Stoger 2011). Furthermore, the amount of peptide can be increased by proper promoter selection, non-target genomic insertion, transgene copy number, and target tissue (Delaunois et al. 2009;

Hernandez-Garcia et al. 2010). Persuaded by these conveniences, researchers have explored expressing peptides, including cysteine rich peptides, in plant systems. In most of the cases, cystine-stabilized antimicrobial peptides were expressed in plants for crop improvement (Parachin et al. 2012). Thus, expression for purification of cystine-stabilized for medical use is still rare using plant systems. In our lab, however, the expression of cystine-stabilized antimicrobial peptides at commercial levels in the tobacco *Nicotiana benthamiana* has been achieved (Ghidey, Islam and Kearney, unpublished data). The key to this success is the choice of anionic peptides over cationic peptides. The cationic antimicrobial peptides are relatively rarely represented in plant genomes compared to anionic antimicrobial peptides, which suggested to us that anionic antimicrobial peptides, though usually not chosen for commercial expression, would express well heterologously in plants.

Use of Machine Learning Algorithms as an Alignment-Free Prediction Method

Machine learning approaches would be especially suitable for the discovery of cystine-stabilized peptide from genome databanks as they exhibit a low sequence homology but have highly conserved structural features, which must be coded for by hidden sequence signatures. Machine learning has been used to classify the structure (Muggleton, King, and Sternberg 1992) and to determine interactions between proteins (Bock and Gough 2001) or to determine specific characteristics (H Nielsen, Brunak, and von Heijne 1999), though the prediction of 3D structure from primary sequence has not yet been achieved. However, prediction of cyclic proteins was made from the primary sequence through machine learning using support vector machine (SVM) learning (Kedarisetti et al. 2014). Furthermore, SVM was used very effectively to predict the 2D

structure of a protein (Sujun Hua and Sun 2001) and its subcellular localization (S. Hua and Sun 2001) from primary sequence. These studies consistently report that machine learning approaches are superior to alignment-based predictions when deriving protein characteristics from primary sequence and perform effectively in protein groups with low sequence similarity. However, the success of machine learning models depends heavily on training data, feature extraction, classifier algorithm selection and optimization. Therefore, it is imperative to select the optimal dataset, meaningful feature generation, and the selection of the perfect classifier with the optimal hyperparameters to construct a usable model.

Overview

In this work, I completed both wet lab and dry lab projects. For my algorithm development work, I focused on developing machine learning-assisted algorithms to predict the structural and functional characteristics of cystine-stabilized peptides. For my *E. coli* expression studies, I specifically focused on antimicrobial peptides (AMPs) because it was convenient to measure the functional characteristics of AMPs as opposed to the methods needed to obtain appropriate metrics pertinent to other functional classes of cystine-stabilized peptides.

In the next chapter, Chapter Two, I begin my algorithm work by noting that a commercially important group of the cystine-stabilized peptides has a common structural relationship that is defined by a disulfide bonding pattern of C1-C4, C2-C5 and C3-C6 where “C” represent the cysteine residues. We named these peptides as "sequential tri-disulfide peptides" (STPs). Next, I describe a species-agnostic machine learning-based classifier which is designed to nominate undefined STPs having low sequence identity

with currently described STPs. To make the predictor, I used a support vector machine as a classifier that was trained by eleven manually generated features which were extracted from the primary sequences of the peptides.

In Chapter Three, I developed an algorithm that is broadly applicable in the analysis of protein structure and function. This algorithm in fact can be used to generate a variety of new algorithms customized to find whatever function or structure a research scientist is interested in studying. Briefly, this comprises a novel supervised protein classification procedure with an automated feature generation model without the requirement of expert intervention for optimal feature selection. The model was defined as m-NGSG (modified *n-grams* and *skip-grams*). Here, I used modified (optimized for protein sequence) *n-grams* (Cavnar and John 1994) and *skip-grams* (Guthrie et al. 2006) to extract features in a protein-family agnostic fashion which is integrated with a logistic regression classifier. Further, I have performed a meta-comparison between our generalized classification model with several other published specialized protein classification models using the corresponding benchmark datasets and cross-validation methods to validate our new model.

In Chapter Four, I applied the m-NGSG system to build five individual models to predict ion channel blockers (ICB), antimicrobial peptides (AMP), acetylcholine receptor inhibitors (ACRI), serine protease inhibitors (SPI), and hemolytic proteins (HLP) from disulfide stabilized protein primary sequence. Identification of hemolytic characteristics will allow the researcher to eliminate from consideration proteins cytotoxic to humans. The results demonstrate the superiority of m-NGSG-based models to PSI-BLAST (Altschul 1997) with different E-values, HMMER (Finn, Clements, and Eddy 2011) and

other available models. Finally, I constructed CSPred model which combines the results of the five different models and gives a probability score for the five crucial functional characteristics of cystine-stabilized proteins. Since the ion channel blockers consist of three significant subclasses (sodium, potassium and calcium channel blockers), I also constructed three classifiers to determine the ability of m-NGSG to classify the ion channel blockers into those three subclasses.

In Chapter Five, I demonstrate a high level of production of the cystine-stabilized antimicrobial peptides plectasin (Mygind et al. 2005) and eurocin (Oeemig et al. 2012) targeted by fusion to bacteriophage A12C coat protein display peptide with specificity for *Staphylococcus aureus* (Yacoby et al. 2006). The SUMO E. *coli* expression vector was used (Butt et al. 2005) for expression of the fusion peptides. This is the first reported study which demonstrates the use of viral-based targeting with antimicrobial peptides.

Division of Work

In Chapter Two, S M Ashiqul Islam conducted the primary investigation, including data aggregation and computation and drafted the manuscript, S M Ashiqul Islam and Tanvir Sajed developed the machine learning approach and on-line analysis tools, Christopher Kearney designed the study and participated in the manuscript, Erich Baker aided in the design of the study and developed the manuscript. All authors have read and approved the manuscript.

In Chapter Three, S M Ashiqul Islam and Christopher Kearney designed the research. S M Ashiqul Islam designed the feature generation algorithm. S M Ashiqul Islam, Benjamin Heil and Erich Baker constructed the optimization algorithm. S M Ashiqul Islam ran the analysis. S M Ashiqul Islam and Erich Baker prepared the manuscript.

In Chapter Four, S M Ashiqul Islam, Christopher Kearney and Erich Baker designed the experiment. S M Ashiqul Islam constructed the model and ran the analysis. S M Ashiqul Islam, Christopher Kearney and Erich Baker prepared the manuscript.

In Chapter Five, S M Ashiqul Islam, Meron Ghidey, and Christopher Kearney designed the experiment. S M Ashiqul Islam constructed the clones, expressed, purified and characterized the recombinant proteins. Ankan Choudhury conducted the high-volume protein production and standardized the method of antimicrobial assay. Ankan Choudhury and Meron Ghidey performed the antimicrobial assay. S M Ashiqul Islam, Ankan Choudhury, Meron Ghidey and Christopher Kearney prepared the manuscript.

References

Aloush, Valerie, Shiri Navon-Venezia, Yardena Seigman-Igra, Shaltiel Cabili, and Yehuda Carmeli. 2006. "Multidrug-Resistant Pseudomonas Aeruginosa: Risk Factors and Clinical Impact." *Antimicrobial Agents and Chemotherapy* 50 (1): 43–48. https://doi.org/10.1128/AAC.50.1.43-48.2006.

Altschul, S. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25 (17): 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

Avrutina, Olga, Hans-Ulrich Schmoldt, Dusica Gabrijelcic-Geiger, Dung Le Nguyen, Christian P. Sommerhoff, Ulf Diederichsen, and Harald Kolmar. 2005. "Trypsin Inhibition by Macrocyclic and Open-Chain Variants of the Squash Inhibitor MCoTI-II." *Biological Chemistry* 386 (12): 1301–6. https://doi.org/10.1515/BC.2005.148.

Bock, J R, and D A Gough. 2001. "Predicting Protein--Protein Interactions from Primary Structure." *Bioinformatics (Oxford, England)* 17 (5): 455–60.

Brooke, B D, R H Hunt, F Chandre, P Carnevale, and M Coetzee. 2002. "Stable Chromosomal Inversion Polymorphisms and Insecticide Resistance in the Malaria Vector Mosquito Anopheles Gambiae (Diptera: Culicidae)." *Journal of Medical Entomology* 39 (4): 568–73.

Butt, Tauseef R., Suzanne C. Edavettal, John P. Hall, and Michael R. Mattern. 2005. "SUMO Fusion Technology for Difficult-to-Express Proteins." *Protein Expression and Purification* 43 (1): 1–9. https://doi.org/10.1016/j.pep.2005.03.016.

Cavnar, William B., and M. Trenkle John. 1994. "N-Gram-Based Text Categorization." *Ann Arbor Mi* 48113 (2): 161–75.

Cereghino, Geoff P Lin, Joan Lin Cereghino, Christine Ilgen, and James M Cregg. 2002. "Production of Recombinant Proteins in Fermenter Cultures of the Yeast Pichia Pastoris." *Current Opinion in Biotechnology* 13 (4): 329–32.

Chevallier, M R, and M Aigle. 1979. "Qualitative Detection of Penicillinase Produced by Yeast Strains Carrying Chimeric Yeast-Coli Plasmids." *FEBS Letters* 108 (1): 179–80.

Cipáková, Ingrid, and Eva Hostinová. 2005. "Production of the Human-Beta-Defensin Using Saccharomyces Cerevisiae as a Host." *Protein and Peptide Letters* 12 (6): 551–54.

Colgrave, Michelle L., Andrew C. Kotze, Yen-Hua Huang, John O'Grady, Shane M. Simonsen, and David J. Craik. 2008. "Cyclotides: Natural, Circular Plant Peptides That Possess Significant Activity against Gastrointestinal Nematode Parasites of Sheep." *Biochemistry* 47 (20): 5581–89. https://doi.org/10.1021/bi800223y.

Colgrave, Michelle L., Andrew C. Kotze, David C. Ireland, Conan K. Wang, and David J. Craik. 2008. "The Anthelmintic Activity of the Cyclotides: Natural Variants with Enhanced Activity." *Chembiochem: A European Journal of Chemical Biology* 9 (12): 1939–45. https://doi.org/10.1002/cbic.200800174.

Colgrave, Michelle L., Andrew C. Kotze, Steven Kopp, James S. McCarthy, Glen T. Coleman, and David J. Craik. 2009. "Anthelmintic Activity of Cyclotides: In Vitro Studies with Canine and Human Hookworms." *Acta Tropica* 109 (2): 163–66. https://doi.org/10.1016/j.actatropica.2008.11.003.

Conibear, Anne C., K. Johan Rosengren, Peta J. Harvey, and David J. Craik. 2012. "Structural Characterization of the Cyclic Cystine Ladder Motif of θ-Defensins." *Biochemistry* 51 (48): 9718–26. https://doi.org/10.1021/bi301363a.

Daly, Rachel, and Milton T W Hearn. 2006. "Expression of the Human Activin Type I and II Receptor Extracellular Domains in Pichia Pastoris." *Protein Expression and Purification* 46 (2): 456–67. https://doi.org/10.1016/j.pep.2005.10.001.

Delaunois, Bertrand, Sylvain Cordelier, Alexandra Conreux, Christophe Clément, and Philippe Jeandet. 2009. "Molecular Engineering of Resveratrol in Plants." *Plant Biotechnology Journal* 7 (1): 2–12. https://doi.org/10.1111/j.1467-7652.2008.00377.x.

Finn, R. D., J. Clements, and S. R. Eddy. 2011. “HMMER Web Server: Interactive Sequence Similarity Searching.” *Nucleic Acids Research* 39 (suppl): W29–37. https://doi.org/10.1093/nar/gkr367.

Fry, Bryan G., Kim Roelants, Donald E. Champagne, Holger Scheib, Joel D. A. Tyndall, Glenn F. King, Timo J. Nevalainen, et al. 2009. “The Toxicogenomic Multiverse: Convergent Recruitment of Proteins into Animal Venoms.” *Annual Review of Genomics and Human Genetics* 10: 483–511. https://doi.org/10.1146/annurev.genom.9.081307.164356.

Gao, Bingmiao, Chao Peng, Jiaan Yang, Yunhai Yi, Junqing Zhang, and Qiong Shi. 2017. “Cone Snails: A Big Store of Conotoxins for Novel Drug Discovery.” *Toxins* 9 (12). https://doi.org/10.3390/toxins9120397.

García-Olmedo, F., A. Molina, A. Segura, and M. Moreno. 1995. “The Defensive Role of Nonspecific Lipid-Transfer Proteins in Plants.” *Trends in Microbiology* 3 (2): 72–74.

González, C., G. M. Langdon, M. Bruix, A. Gálvez, E. Valdivia, M. Maqueda, and M. Rico. 2000. “Bacteriocin AS-48, a Microbial Cyclic Polypeptide Structurally and Functionally Related to Mammalian NK-Lysin.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (21): 11221–26. https://doi.org/10.1073/pnas.210301097.

Gracy, J., D. Le-Nguyen, J.-C. Gelly, Q. Kaas, A. Heitz, and L. Chiche. 2007. “KNOTTIN: The Knottin or Inhibitor Cystine Knot Scaffold in 2007.” *Nucleic Acids Research* 36 (Database): D314–19. https://doi.org/10.1093/nar/gkm939.

Gracy, Jérôme, and Laurent Chiche. 2011. “Structure and Modeling of Knottins, a Promising Molecular Scaffold for Drug Discovery.” *Current Pharmaceutical Design* 17 (38): 4337–50.

Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. 2006. “A Closer Look at Skip-Gram Modelling.” In , 1–4. sn.

Hammami, R., J. Ben Hamida, G. Vergoten, and I. Fliss. 2009. “PhytAMP: A Database Dedicated to Antimicrobial Plant Peptides.” *Nucleic Acids Research* 37 (Database): D963–68. https://doi.org/10.1093/nar/gkn655.

Hemingway, J, and H Ranson. 2000. “Insecticide Resistance in Insect Vectors of Human Disease.” *Annual Review of Entomology* 45: 371–91. https://doi.org/10.1146/annurev.ento.45.1.371.

Hernandez-Garcia, Carlos M, Robert A Bouchard, Paul J Rushton, Michelle L Jones, Xianfeng Chen, Michael P Timko, and John J Finer. 2010. “High Level Transgenic Expression of Soybean (Glycine Max) GmERF and Gmubi Gene Promoters Isolated by a Novel Promoter Analysis Pipeline.” *BMC Plant Biology* 10 (1): 237. https://doi.org/10.1186/1471-2229-10-237.

Herzig, V., D. L. A. Wood, F. Newell, P.-A. Chaumeil, Q. Kaas, G. J. Binford, G. M. Nicholson, D. Gorse, and G. F. King. 2011. “ArachnoServer 2.0, an Updated Online Resource for Spider Toxin Sequences and Structures.” *Nucleic Acids Research* 39 (Database): D653–57. https://doi.org/10.1093/nar/gkq1058.

Hiramatsu, K, N Aritaka, H Hanaki, S Kawasaki, Y Hosoda, S Hori, Y Fukuchi, and I Kobayashi. 1997. “Dissemination in Japanese Hospitals of Strains of Staphylococcus Aureus Heterogeneously Resistant to Vancomycin.” *Lancet* 350 (9092): 1670–73. https://doi.org/10.1016/S0140-6736(97)07324-8.

Hua, S., and Z. Sun. 2001. “Support Vector Machine Approach for Protein Subcellular Localization Prediction.” *Bioinformatics* 17 (8): 721–28. https://doi.org/10.1093/bioinformatics/17.8.721.

Hua, Sujun, and Zhirong Sun. 2001. “A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach.” *Journal of Molecular Biology* 308 (2): 397–407. https://doi.org/10.1006/jmbi.2001.4580.

Jennings, C., J. West, C. Waine, D. Craik, and M. Anderson. 2001. “Biosynthesis and Insecticidal Properties of Plant Cyclotides: The Cyclic Knotted Proteins from Oldenlandia Affinis.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (19): 10614–19. https://doi.org/10.1073/pnas.191366898.

Jennings, Cameron V., K. Johan Rosengren, Norelle L. Daly, Manuel Plan, Jackie Stevens, Martin J. Scanlon, Clement Waine, David G. Norman, Marilyn A. Anderson, and David J. Craik. 2005. “Isolation, Solution Structure, and Insecticidal Activity of Kalata B2, a Circular Protein with a Twist: Do Möbius Strips Exist in Nature?” *Biochemistry* 44 (3): 851–60. https://doi.org/10.1021/bi047837h.

Kaas, Q., R. Yu, A.-H. Jin, S. Dutertre, and D. J. Craik. 2012. “ConoServer: Updated Content, Knowledge, and Discovery Tools in the Conopeptide Database.” *Nucleic Acids Research* 40 (D1): D325–30. https://doi.org/10.1093/nar/gkr886.

Kedarisetti, Pradyumna, Marcin J. Mizianty, Quentin Kaas, David J. Craik, and Lukasz Kurgan. 2014. “Prediction and Characterization of Cyclic Proteins from Sequences in Three Domains of Life.” *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1844 (1): 181–90. https://doi.org/10.1016/j.bbapap.2013.05.002.

Koehbach, Johannes, Alfred F. Attah, Andreas Berger, Roland Hellinger, Toni M. Kutchan, Eric J. Carpenter, Megan Rolf, et al. 2013. "Cyclotide Discovery in Gentianales Revisited-Identification and Characterization of Cyclic Cystine-Knot Peptides and Their Phylogenetic Distribution in Rubiaceae Plants: Cyclotide Discovery in Gentianales Revisited." *Biopolymers* 100 (5): 438–52. https://doi.org/10.1002/bip.22328.

Koehbach, Johannes, Margaret O'Brien, Markus Muttenthaler, Marion Miazzo, Muharrem Akcan, Alysha G. Elliott, Norelle L. Daly, et al. 2013. "Oxytocic 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* 110 (52): 21183–88. https://doi.org/10.1073/pnas.1311183110.

Lewis, Richard J., and Maria L. Garcia. 2003. "Therapeutic Potential of Venom Peptides." *Nature Reviews Drug Discovery* 2 (10): 790–802. https://doi.org/10.1038/nrd1197.

Liang, Song-Ping, and Xin Pan. 1995. "A Lectin-like Peptide Isolated from the Venom of the Chinese Bird Spider Selenocosmia Huwena." *Toxicon* 33 (7): 875–82. https://doi.org/10.1016/0041-0101(95)00033-I.

Lin, Hao, and Qian-Zhong Li. 2007. "Predicting Conotoxin Superfamily and Family by Using Pseudo Amino Acid Composition and Modified Mahalanobis Discriminant." *Biochemical and Biophysical Research Communications* 354 (2): 548–51. https://doi.org/10.1016/j.bbrc.2007.01.011.

Lu, S., J. Van Eck, X. Zhou, A. B. Lopez, D. M. O'Halloran, K. M. Cosman, B. J. Conlin, et al. 2006. "The Cauliflower Or Gene Encodes a DnaJ Cysteine-Rich Domain-Containing Protein That Mediates High Levels of -Carotene Accumulation." *THE PLANT CELL ONLINE* 18 (12): 3594–3605. https://doi.org/10.1105/tpc.106.046417.

Malagón, David, Bonnie Botterill, Darren J. Gray, Erica Lovas, Mary Duke, Christian Gray, Steven R. Kopp, et al. 2013. "Anthelminthic Activity of the Cyclotides (Kalata B1 and B2) against Schistosome Parasites." *Biopolymers* 100 (5): 461–70. https://doi.org/10.1002/bip.22229.

Mattanovich, Diethard, Paola Branduardi, Laura Dato, Brigitte Gasser, Michael Sauer, and Danilo Porro. 2012. "Recombinant Protein Production in Yeasts." In *Recombinant Gene Expression*, edited by Argelia Lorence, 824:329–58. Totowa, NJ: Humana Press. http://link.springer.com/10.1007/978-1-61779-433-9_17.

Mondal, Sukanta, Rajasekaran Bhavna, Rajasekaran Mohan Babu, and Suryanarayanarao Ramakumar. 2006. "Pseudo Amino Acid Composition and Multi-Class Support Vector Machines Approach for Conotoxin Superfamily Classification." *Journal of Theoretical Biology* 243 (2): 252–60. https://doi.org/10.1016/j.jtbi.2006.06.014.

Muggleton, S, R D King, and M J Sternberg. 1992. “Protein Secondary Structure Prediction Using Logic-Based Machine Learning.” *Protein Engineering* 5 (7): 647–57.

Mygind, Per H., Rikke L. Fischer, Kirk M. Schnorr, Mogens T. Hansen, Carsten P. Sönksen, Svend Ludvigsen, Dorotea Raventós, et al. 2005. “Plectasin Is a Peptide Antibiotic with Therapeutic Potential from a Saprophytic Fungus.” *Nature* 437 (7061): 975–80. https://doi.org/10.1038/nature04051.

Mylne, Joshua S., Lai Yue Chan, Aurelie H. Chanson, Norelle L. Daly, Hanno Schaefer, Timothy L. Bailey, Philip Nguyencong, Laura Cascales, and David J. Craik. 2012. “Cyclic Peptides Arising by Evolutionary Parallelism via Asparaginyl-Endopeptidase-Mediated Biosynthesis.” *The Plant Cell* 24 (7): 2765–78. https://doi.org/10.1105/tpc.112.099085.

Nguyen, Giang Kien Truc, Wei Han Lim, Phuong Quoc Thuc Nguyen, and James P. Tam. 2012. “Novel Cyclotides and Uncyclotides with Highly Shortened Precursors from Chassalia Chartacea and Effects of Methionine Oxidation on Bioactivities.” *The Journal of Biological Chemistry* 287 (21): 17598–607. https://doi.org/10.1074/jbc.M111.338970.

Nguyen, Giang Kien Truc, Sen Zhang, Ngan Thi Kim Nguyen, Phuong Quoc Thuc Nguyen, Ming Sheau Chiu, Antony Hardjojo, and James P. Tam. 2011. “Discovery and Characterization of Novel Cyclotides Originated from Chimeric Precursors Consisting of Albumin-1 Chain a and Cyclotide Domains in the Fabaceae Family.” *The Journal of Biological Chemistry* 286 (27): 24275–87. https://doi.org/10.1074/jbc.M111.229922.

Nielsen, H, S Brunak, and G von Heijne. 1999. “Machine Learning Approaches for the Prediction of Signal Peptides and Other Protein Sorting Signals.” *Protein Engineering* 12 (1): 3–9.

Nielsen, Henrik. 2017. “Predicting Secretory Proteins with SignalP.” *Methods in Molecular Biology (Clifton, N.J.)* 1611: 59–73. https://doi.org/10.1007/978-1-4939-7015-5_6.

Oeemig, Jesper S., Carina Lynggaard, Daniel H. Knudsen, Frederik T. Hansen, Kent D. Nørgaard, Tanja Schneider, Brian S. Vad, et al. 2012. “Eurocin, a New Fungal Defensin: Structure, Lipid Binding, and Its Mode of Action.” *The Journal of Biological Chemistry* 287 (50): 42361–72. https://doi.org/10.1074/jbc.M112.382028.

Parachin, Nádia Skorupa, Kelly Cristina Mulder, Antônio Américo Barbosa Viana, Simoni Campos Dias, and Octávio Luiz Franco. 2012. “Expression Systems for Heterologous Production of Antimicrobial Peptides.” *Peptides* 38 (2): 446–56. https://doi.org/10.1016/j.peptides.2012.09.020.

Peters, Jenny, and Eva Stoger. 2011. “Transgenic Crops for the Production of Recombinant Vaccines and Anti-Microbial Antibodies.” *Human Vaccines* 7 (3): 367–74.

Porto, William F., Valéria A. Souza, Diego O. Nolasco, and Octávio L. Franco. 2012. “In Silico Identification of Novel Hevein-like Peptide Precursors.” *Peptides* 38 (1): 127–36. https://doi.org/10.1016/j.peptides.2012.07.025.

Postic, Guillaume, Jérôme Gracy, Charlotte Périn, Laurent Chiche, and Jean-Christophe Gelly. 2018. “KNOTTIN: The Database of Inhibitor Cystine Knot Scaffold after 10 Years, toward a Systematic Structure Modeling.” *Nucleic Acids Research* 46 (D1): D454–58. https://doi.org/10.1093/nar/gkx1084.

Poth, Aaron G., Michelle L. Colgrave, Russell E. Lyons, Norelle L. Daly, and David J. Craik. 2011. “Discovery of an Unusual Biosynthetic Origin for Circular Proteins in Legumes.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (25): 10127–32. https://doi.org/10.1073/pnas.1103660108.

Poth, Aaron G., Joshua S. Mylne, Julia Grassl, Russell E. Lyons, A. Harvey Millar, Michelle L. Colgrave, and David J. Craik. 2012. “Cyclotides Associate with Leaf Vasculature and Are the Products of a Novel Precursor in Petunia (Solanaceae).” *The Journal of Biological Chemistry* 287 (32): 27033–46. https://doi.org/10.1074/jbc.M112.370841.

Rees, D. C., and W. N. Lipscomb. 1982. “Refined Crystal Structure of the Potato Inhibitor Complex of Carboxypeptidase A at 2.5 A Resolution.” *Journal of Molecular Biology* 160 (3): 475–98.

Rosengren, K. Johan, Norelle L. Daly, Manuel R. Plan, Clement Waine, and David J. Craik. 2003. “Twists, Knots, and Rings in Proteins. Structural Definition of the Cyclotide Framework.” *The Journal of Biological Chemistry* 278 (10): 8606–16. https://doi.org/10.1074/jbc.M211147200.

Roth, Amy F., Ying Feng, Linyi Chen, and Nicholas G. Davis. 2002. “The Yeast DHHC Cysteine-Rich Domain Protein Akr1p Is a Palmitoyl Transferase.” *The Journal of Cell Biology* 159 (1): 23–28. https://doi.org/10.1083/jcb.200206120.

Ryan, M. C., and L. J. Sandell. 1990. “Differential Expression of a Cysteine-Rich Domain in the Amino-Terminal Propeptide of Type II (Cartilage) Procollagen by Alternative Splicing of MRNA.” *The Journal of Biological Chemistry* 265 (18): 10334–39.

Saether, Olav, David J. Craik, Iain D. Campbell, Knut Sletten, Jessie Juul, and David G. Norman. 1995. “Elucidation of the Primary and Three-Dimensional Structure of the Uterotonic Polypeptide Kalata B1.” *Biochemistry* 34 (13): 4147–58. https://doi.org/10.1021/bi00013a002.

Saez, Natalie J., Sebastian Senff, Jonas E. Jensen, Sing Yan Er, Volker Herzig, Lachlan D. Rash, and Glenn F. King. 2010. “Spider-Venom Peptides as Therapeutics.” *Toxins* 2 (12): 2851–71. https://doi.org/10.3390/toxins2122851.

Schoeman, H, M A Vivier, M Du Toit, L M Dicks, and I S Pretorius. 1999. “The Development of Bactericidal Yeast Strains by Expressing the Pediococcus Acidilactici Pediocin Gene (PedA) in Saccharomyces Cerevisiae.” *Yeast (Chichester, England)* 15 (8): 647–56. https://doi.org/10.1002/(SICI)1097-0061(19990615)15:8<647::AID-YEA409>3.0.CO;2-5.

Segura, Ana, Manuel Moreno, Francisco Madueño, Antonio Molina, and Francisco García-Olmedo. 1999. “Snakin-1, a Peptide from Potato That Is Active Against Plant Pathogens.” *Molecular Plant-Microbe Interactions* 12 (1): 16–23. https://doi.org/10.1094/MPMI.1999.12.1.16.

Sels, Jan, Janick Mathys, Barbara M. A. De Coninck, Bruno P. A. Cammue, and Miguel F. C. De Bolle. 2008. “Plant Pathogenesis-Related (PR) Proteins: A Focus on PR Peptides.” *Plant Physiology and Biochemistry: PPB* 46 (11): 941–50. https://doi.org/10.1016/j.plaphy.2008.06.011.

Shatzman, A R, and M Rosenberg. 1987. “Expression, Identification, and Characterization of Recombinant Gene Products in Escherichia Coli.” *Methods in Enzymology* 152: 661–73.

Sigrist, Christian J. A., Lorenzo Cerutti, Edouard de Castro, Petra S. Langendijk-Genevaux, Virginie Bulliard, Amos Bairoch, and Nicolas Hulo. 2010. “PROSITE, a Protein Domain Database for Functional Characterization and Annotation.” *Nucleic Acids Research* 38 (suppl_1): D161–66. https://doi.org/10.1093/nar/gkp885.

Silverstein, Kevin A. T., William A. Moskal, Hank C. Wu, Beverly A. Underwood, Michelle A. Graham, Christopher D. Town, and Kathryn A. VandenBosch. 2007. “Small Cysteine-Rich Peptides Resembling Antimicrobial Peptides Have Been under-Predicted in Plants.” *The Plant Journal: For Cell and Molecular Biology* 51 (2): 262–80. https://doi.org/10.1111/j.1365-313X.2007.03136.x.

Simonsen, Shane M., Lillian Sando, David C. Ireland, Michelle L. Colgrave, Rekha Bharathi, Ulf Göransson, and David J. Craik. 2005. “A Continent of Plant Defense Peptide Diversity: Cyclotides in Australian Hybanthus (Violaceae).” *The Plant Cell* 17 (11): 3176–89. https://doi.org/10.1105/tpc.105.034678.

Singh, Pradip Kumar, Vipul Solanki, Shalley Sharma, Krishan Gopal Thakur, Beena Krishnan, and Suresh Korpole. 2015. “The Intramolecular Disulfide-Stapled Structure of Laterosporulin, a Class IId Bacteriocin, Conceals a Human Defensin-like Structural Module.” *The FEBS Journal* 282 (2): 203–14. https://doi.org/10.1111/febs.13129.

Sinha, Mau, Rashmi Prabha Singh, Gajraj Singh Kushwaha, Naseer Iqbal, Avinash Singh, Sanket Kaushik, Punit Kaur, Sujata Sharma, and Tej P. Singh. 2014. "Current Overview of Allergens of Plant Pathogenesis Related Protein Families." *TheScientificWorldJournal* 2014: 543195. https://doi.org/10.1155/2014/543195.

Sit, Clarissa S., Ryan T. McKay, Colin Hill, R. Paul Ross, and John C. Vederas. 2011. "The 3D Structure of Thuricin CD, a Two-Component Bacteriocin with Cysteine Sulfur to α-Carbon Cross-Links." *Journal of the American Chemical Society* 133 (20): 7680–83. https://doi.org/10.1021/ja201802f.

Stintzi, A., T. Heitz, V. Prasad, S. Wiedemann-Merdinoglu, S. Kauffmann, P. Geoffroy, M. Legrand, and B. Fritig. 1993. "Plant 'pathogenesis-Related' Proteins and Their Role in Defense against Pathogens." *Biochimie* 75 (8): 687–706.

Stotz, Henrik U., James Thomson, and Yueju Wang. 2009. "Plant Defensins: Defense, Development and Application." *Plant Signaling & Behavior* 4 (11): 1010–12. https://doi.org/10.4161/psb.4.11.9755.

Talmadge, K, J Kaufman, and W Gilbert. 1980. "Bacteria Mature Preproinsulin to Proinsulin." *Proceedings of the National Academy of Sciences of the United States of America* 77 (7): 3988–92.

Tam, James, Shujing Wang, Ka Wong, and Wei Tan. 2015. "Antimicrobial Peptides from Plants." *Pharmaceuticals* 8 (4): 711–57. https://doi.org/10.3390/ph8040711.

Tan, Paul T. J., K. N. Srinivasan, Seng Hong Seah, Judice L. Y. Koh, Tin Wee Tan, Shoba Ranganathan, and Vladimir Brusic. 2005. "Accurate Prediction of Scorpion Toxin Functional Properties from Primary Structures." *Journal of Molecular Graphics & Modelling* 24 (1): 17–24. https://doi.org/10.1016/j.jmgm.2005.01.003.

Tan, Paul T. J., Anitha Veeramani, Kellathur N. Srinivasan, Shoba Ranganathan, and Vladimir Brusic. 2006. "SCORPION2: A Database for Structure-Function Analysis of Scorpion Toxins." *Toxicon: Official Journal of the International Society on Toxinology* 47 (3): 356–63. https://doi.org/10.1016/j.toxicon.2005.12.001.

Teeter, Martha M., Jonathan A. Mazer, and James J. L'Italien. 1981. "Primary Structure of the Hydrophobic Plant Protein Crambin." *Biochemistry* 20 (19): 5437–43. https://doi.org/10.1021/bi00522a013.

Wang, C. K. L., Q. Kaas, L. Chiche, and D. J. Craik. 2007. "CyBase: A Database of Cyclic Protein Sequences and Structures, with Applications in Protein Discovery and Engineering." *Nucleic Acids Research* 36 (Database): D206–10. https://doi.org/10.1093/nar/gkm953.

Whetstone, Paul A, and Bruce D Hammock. 2007. “Delivery Methods for Peptide and Protein Toxins in Insect Control.” *Toxicon: Official Journal of the International Society on Toxinology* 49 (4): 576–96. https://doi.org/10.1016/j.toxicon.2006.11.009.

Wu, Yun, Yufei Zheng, and Hua Tang. 2016. “Identifying the Types of Ion Channel-Targeted Conotoxins by Incorporating New Properties of Residues into Pseudo Amino Acid Composition.” *BioMed Research International* 2016: 3981478. https://doi.org/10.1155/2016/3981478.

Xianfang, Wang, Wang Junmei, Wang Xiaolei, and Zhang Yue. 2017. “Predicting the Types of Ion Channel-Targeted Conotoxins Based on AVC-SVM Model.” *BioMed Research International* 2017: 2929807. https://doi.org/10.1155/2017/2929807.

Yacoby, I., M. Shamis, H. Bar, D. Shabat, and I. Benhar. 2006. “Targeting Antibacterial Agents by Using Drug-Carrying Filamentous Bacteriophages.” *Antimicrobial Agents and Chemotherapy* 50 (6): 2087–97. https://doi.org/10.1128/AAC.00169-06.

Yuan, Lu-Feng, Chen Ding, Shou-Hui Guo, Hui Ding, Wei Chen, and Hao Lin. 2013. “Prediction of the Types of Ion Channel-Targeted Conotoxins Based on Radial Basis Function Network.” *Toxicology in Vitro: An International Journal Published in Association with BIBRA* 27 (2): 852–56. https://doi.org/10.1016/j.tiv.2012.12.024.

Zhang, Jun, Zhengshuang Hua, Zebo Huang, QiZhu Chen, Qingyun Long, David J. Craik, Alan J. M. Baker, Wensheng Shu, and Bin Liao. 2015. “Two Blast-Independent Tools, CyPerl and CyExcel, for Harvesting Hundreds of Novel Cyclotides and Analogues from Plant Genomes and Protein Databases.” *Planta* 241 (4): 929–40. https://doi.org/10.1007/s00425-014-2229-5.

Zhu, Shunyi, Herve Darbon, Karin Dyason, Fons Verdonck, and Jan Tytgat. 2003. “Evolutionary Origin of Inhibitor Cystine Knot Peptides.” *The FASEB Journal* 17 (12): 1765–67. https://doi.org/10.1096/fj.02-1044fje.

CHAPTER TWO

PredSTP: A Highly Accurate SVM Based Model to Predict Sequential Cystine Stabilized Peptides

This chapter is published as: Islam, SM Ashiqul, Tanvir Sajed, Christopher Michel Kearney, and Erich J. Baker. "PredSTP: a highly accurate SVM based model to predict sequential cystine stabilized peptides." BMC bioinformatics 16, no. 1 (2015): 210.

Abstract

Numerous organisms have evolved a wide range of toxic peptides for self-defense and predation. Their effective interstitial and macro-environmental use requires energetic and structural stability. One successful group of these peptides includes a tri-disulfide domain arrangement that offers toxicity and high stability. Sequential tri-disulfide connectivity variants create highly compact disulfide folds capable of withstanding a variety of environmental stresses. Their combination of toxicity and stability make these peptides remarkably valuable for their potential as bio-insecticides, antimicrobial peptides and peptide drug candidates. However, the wide sequence variation, sources and modalities of group members impose serious limitations on our ability to rapidly identify potential members. As a result, there is a need for automated high-throughput member classification approaches that leverage their demonstrated tertiary and functional homology. We developed an SVM-based model to predict sequential tri-disulfide peptide (STP) toxins from peptide sequences. One optimized model, called PredSTP, predicted STPs from training set with sensitivity, specificity, precision, accuracy and a Matthews correlation coefficient of 94.86%, 94.11%, 84.31%, 94.30% and 0.86, respectively , using 200 fold cross validation. The same model outperforms existing prediction

approaches in three independent out of sample testsets derived from PDB. PredSTP can accurately identify a wide range of cystine stabilized peptide toxins directly from sequences in a species-agnostic fashion. The ability to rapidly filter sequences for potential bioactive peptides can greatly compress the time between peptide identification and testing structural and functional properties for possible antimicrobial and insecticidal candidates. A web interface is freely available to predict STP toxins from http://crick.ecs.baylor.edu/.

Introduction

Certain proteins are known to be toxic to living organisms(Carlini and Grossi-de-Sá 2002; Gordon, Romanowski, and McDermott 2005; Lehrer, Lichtenstein, and Ganz 1993) and this toxicity can serve to provide defense for the host organism against opportunistic insects and microorganisms. In medicine and agriculture, naturally occurring toxic proteins provide an alternative to the rapidly dwindling supply of effective synthetic chemical insecticides, antimicrobials and antifungals (Hemingway and Ranson 2000; Brooke et al. 2002; Aloush et al. 2006; Hiramatsu et al. 1997).

Structural stability is critical to the success of these toxic peptides (Marr, Gooderham, and Hancock 2006). For example, the physiological environment of an organism contains proteases and highly variable pH which can greatly impact peptide integrity. While a number of approaches can increase the stability of peptides under adverse environments (Monroc et al. 2006; Braunstein, Papo, and Shai 2004), the inclusion of disulfide bonds is one natural way to increase stability (Matsumura, Signor, and Matthews 1989; Tugyi et al. 2005). Conversely, in several cases, disulfide bonds may hinder the potent activity of a peptide (Schroeder et al. 2011; Circo et al. 2002), much

work is being undertaken to elucidate disulfide rich stable toxic peptides as insecticides(Jennings et al. 2005; Bende et al. 2014), antimicrobial peptides (Reddy, Yedery, and Aranha 2004) and therapeutic potentials (Henriques and Craik 2010; Lewis and Garcia 2003).

Despite a wide range of diversity based on their sources and modes of actions, all cystine stabilized toxins contain a fold with multiple disulfide connectivity (Lewis and Garcia 2003). A sequential array of tri-disulfide connectivity is regarded as the most stable (Góngora-Benítez, Tulla-Puche, and Albericio 2014). It has a compact cystine trio, where the first cysteine participating in the fold makes a disulfide bond with the fourth cysteine, the second one with the fifth cysteine and the third one with the sixth cysteine (C1-C4, C2-C5, C3-C6). There may be other cysteines in the primary sequence of these peptides, but they do not participate in that sequential tri-disulfide connectivity. This class of proteins includes several large protein families such as knottins (Gracy et al. 2008a), scorpion toxin-like superfamily (Zhu et al. 2011a), cyclotides (Gould et al. 2011), and a substantial proportion of diverse peptides comprising antimicrobial peptides and defensins (Bulet et al. 1999). For clarity, toxic peptides containing this particular stable disulfide connectivity can be referred to as sequential tri-disulfide peptide toxins (STP toxins). Cystine stabilized toxins which do not contain the exact STP bonding array may also offer stability and toxicity (Conibear et al. 2013, 2014; Ovchinnikova et al. 2006; Ye et al. 2012) and can be denoted as nonsequential tri-disulfide peptides (NTPs) (Figure 2.1). While STP toxins imply a compact tri-disulfide tertiary confirmation, NTPs toxins may contain both compact or non-compact tri-disulfide folds (Figure 2.2)

STP toxins can be further divided into three major groups based on their canonical 3D definitions: Cyclotides [29, 30], inhibitor cystine knots (ICKs) [31] and nonknotted STPs [32–34]. Cyclotides form cyclization through N-C terminus adherence and are renowned as stable peptides containing the sequential tri-disulfide array [35]. In this type of peptide, the third disulfide bond penetrates through the other two disulfide bonds participating in the array and forms a knotted macrocycle of disulfide bonds. ICKs, also known as knottins, are a second type of STPs [36]. They contain the same knotted macrocycle as cyclotides but do not necessarily take the cyclic form. The third type has three sequentially paired disulfide bonds but the third bond does not penetrate the macrocycle, preventing the formation of a 'knot'. This group may actually contain as many toxins as the first two subgroups combined and includes scorpion toxin-like peptides [32, 34], insect peptides [33], plant peptides [38], and a variety of other peptides. All three STP subgroups are characterized by high stability and toxicity [33, 39–42]. Although STP toxins show similarity in their function and highly constrained folds, they share little sequence identity (Gracy et al. 2008b; Zhu et al. 2011b). As a consequence, discovery of new STPs has traditionally been slow and almost exclusively based on functional properties. In the case of ICKs, an automated discovery process based on sequence similarity using BLAST has previously been paired with sequence and structural algorithms (Knoter 1D and 3D, respectively) to precisely verify knottin candidates (Gracy et al. 2008b; Gelly et al. 2004). The discovery of knottins via sequence similarity has produced an extensive and well-organized database, despite a scope limited to sequence similarity [25]. Cypred (Kedarisetti et al. 2014) is another relevant software that can predict cyclic proteins and a significant subset of these cyclic

peptides have STP like connectivity. While there is no known software to predict non-knotted STPs, there are databases focusing on limited specific families, such as CyBase for cyclotides (Mulvenna, Wang, and Craik 2006; Wang et al. 2008), Conoserver for conotoxins (Kaas et al. 2012) and Arachnoserver for spider toxins (Herzig et al. 2011), but these have little broad application.

Machine learning approaches offer one possible solution for the broad discovery of STP toxins through the use of soft or fuzzy classification schemas, based on salient STP features that extend beyond a reliance on primary sequence similarity. Logic-based machine learning has been used previously to classify the 2D structure of α/α domain type proteins (Muggleton, King, and Sternberg 1992), protein-protein interactions (Bock and Gough 2001) or functional classifications of proteins from primary sequence. In particular, Support Vector Machines (SVM), a robust class of machine learning approaches (C. Z. Cai et al. 2003), have been successfully used to predict cyclic proteins (Kedarisetti et al. 2014), 2D and 3D protein structures (Sujun Hua and Sun 2001; Y. D. Cai et al. 2001) and subcellular localization (S. Hua and Sun 2001) from primary sequence.

Here, we illustrate a species-agnostic machine learning methodology, called PredSTP (http://crick.ecs.baylor.edu), which is designed to nominate undefined STPs having low sequence identity with currently described STPs. Efficient discovery of new functional members of this class of proteins will enhance our repertoire of potentially stable insecticidal and antimicrobial proteins.

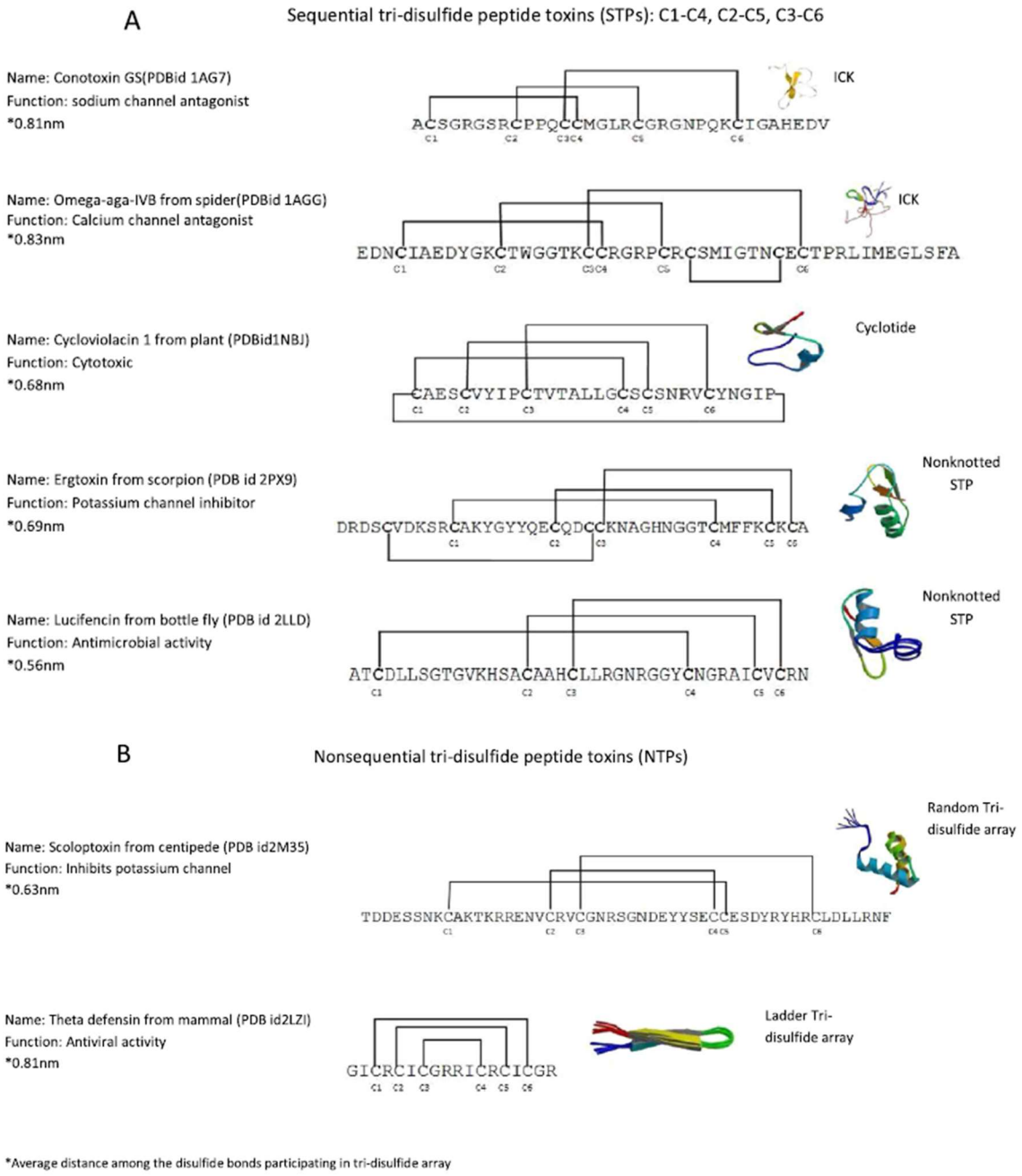

Figure 2.1: Diagrams of the disulfide connectivity of different cystine stabilized toxic peptides. (A) This figure illustrates the pattern of disulfide connectivity of different types of STP toxins (knotted and non-knotted). Each type is annotated with its name, PDB id, function and jmol estimated average 3D structural distance between disulfide bonds. (B) Illustrates the pattern of disulfide connectivity of NTP toxins with the same type of information.

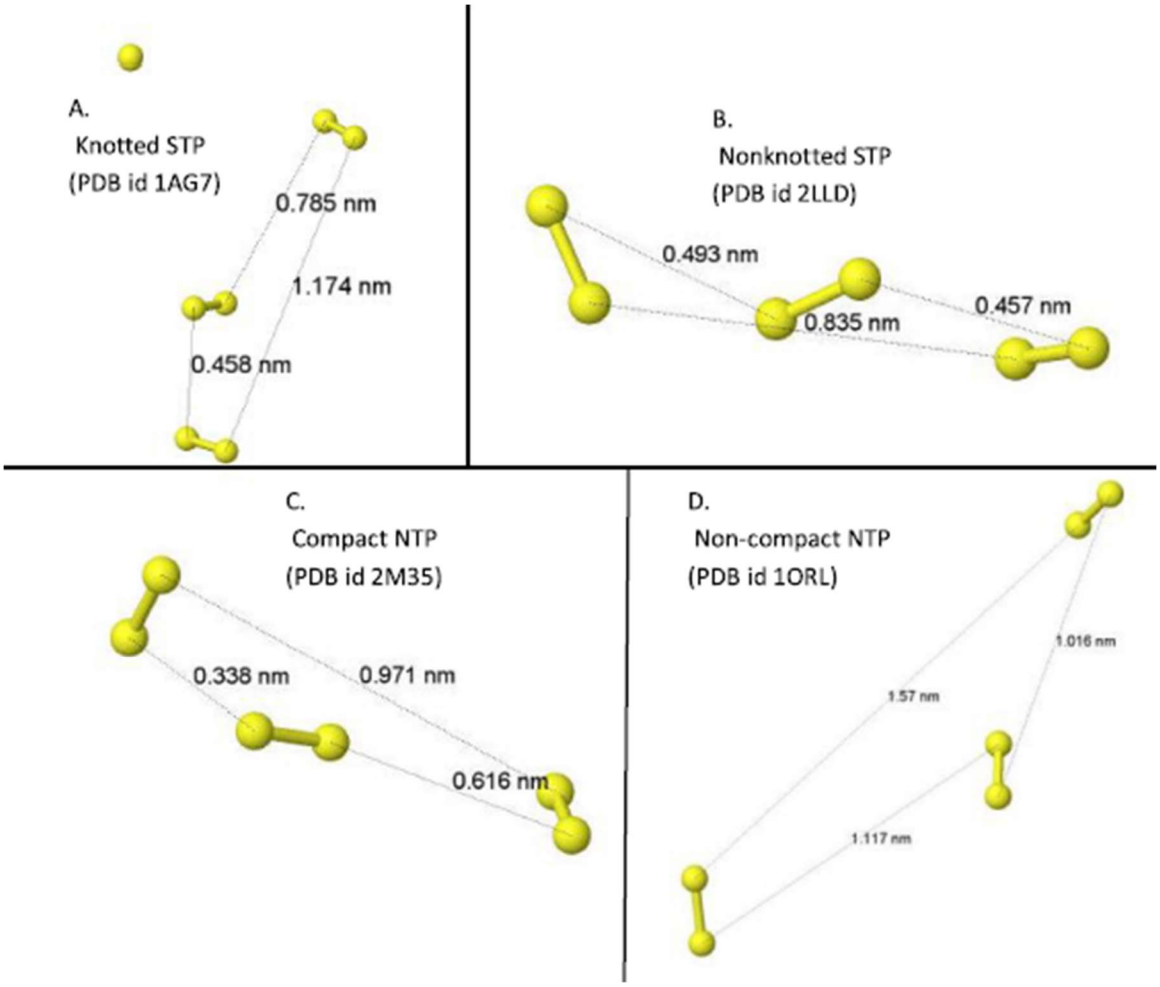

Figure 2.2: Illustration of distances among the non-pairing sulfur molecules participating in the tri-disulfide array. Distances between different sulfur molecule pairs (yellow balls) were measured using jmol software. The mean of these distances indicates the average distance among the disulfide bonds demonstrating the compactness of the tri-disulfide fold in the peptide. A, B, C and D show distances of a sample representative of knotted STPs, nonknotted STPs, compact NTPs and non-compact NTPs, respectively, together with their PDB ids. The average of distance in STP toxins (A and B) is typically less than 0.85nm, while it is more than 1.2nm in other tri-disulfide peptides (Non-compact NTPs, data not shown) (D). Some NTPs demonstrate a similar compactness (average distance) to STPs and can be designated as compact NTPs (C).

Methods and Materials

Known STP Sequence Collection

Sequence of ICKs and cyclotides (knotted STPs) were collected from the Knottin database (http://knottin. cbs.cnrs.fr/) and 167 sequences with solved 3D structures were obtained from this source. An additional 36 sequences of nonknotted STPs with known 3D structures were collected from PDB with 90% sequence identity (http://www.rcsb.org/ , June, 2013). Our total set of 204 candidate sequences (167 from the knottin database and 37 from PDB) were further reduced to remove redundant sequences, defined as sequences sharing ≥ 90% sequence identity using CD-HIT (Huang et al. 2010; Li and Godzik 2006). A total of 108 sequences were retained from the knottin database set and 36 sequences were from the PDB set, leaving 144 canonical STPs (Table S2.1). The mean, standard deviation and range of the number of residues in the positive training set are 42.20, 15.70 and 23-143, respectively, with an average number of 6 cysteines per chain.

Control Negative Sequence Collection

Sequences classified as negative control were collected from PDB using a criterion that was species agnostic and stipulated the exclusion of STPs through positive matches to PDB small proteins (Table S2.2). 393 sequences were classified as non-STP sequences for the purposes of this study. The mean, standard deviation and range of the number of residues in the chains of the negative training set are 63.16, 25.92 and 9-160, respectively, with an average number of 6 cysteines per chain.

Independent Test Sequence Collection

Seven independent sets of sequences were collected to verify the robustness of the model (Table 2.1). Among these were sets classified according to Protein Data Bank (PDB, July 2013) criteria as Eukaryote, Bacteria, Archaea, Virus and Unassigned. In addition, a set of proteins whose sequences were recently solved by NMR and deposited in PDB (July 04, 2012 to March 25, 2014) (NewNMR751) and also the Structural Classification of Protein (SCOP) PDB subset were used (Smallprotein163). Small protein sequences were retrieved with the following parameters: (a) resolution < 1.5 Å, (b) protein chain but not DNA/RNA/Hybrid, and (c) limited to small disulfide rich proteins and have similarity in size, number of disulfide bonds, cystine number and cystine arrangements in their primary structure. The result included STPs, rubredoxins, BPTI-like, snake toxin-like, crambin-like, insulin-like, and high potential iron proteins among others.

Defining the Putative STP Cystine Motif

STP motifs consist of six cysteine residues (C1-C6) flanked by varying number of non-cysteine residues (Figure 2.1). This set of consecutive cysteines is identified here by elucidating the distance between each consecutive pair of cysteines, *i* and *i+1 as* $\Delta C_{i,i+1}$ (cysteine loops). Based on our global analysis of STP motifs, if the $min(\Delta C_{i,i+1})$ is greater than three, then the motif is not considered to contain a STP and is discarded (Figure S2.1). Likewise, if the $min(\Delta C_{i,i+1})$ is less than or equal to three and located between C1 and C2 or C2 and C3 the motifs are disregarded as these motifs are often found within electron transport-like proteins such as ferredoxin, rubredoxin, and iron-sulfur proteins (Emeleus 1959; van Beilen et al. 2002). Otherwise, the $min(\Delta C_{i,i+1})$ was defined to exist

between cysteines C3 and C4. This default pair of cysteines is shifted to a higher pair of cysteines if there exist less than 2 additional c-terminus cysteines. For example, if after the default C3 and C4 cysteines are identified, there is only one c-terminus cysteine, then the *min* ($\Delta C_{i,i+1}$) is defined as cysteines C4 and C5.

Table 2.1: Description of independent test sets analyzed by the new model (PredSTP)

Independent test sample	Query Parameters (PDB*)	Number of Proteins	Number of Chains
Small protein 92	SCOP: Small Proteins Experimental Method: X-RAY Resolution: 1.499 or less	92	163
Only Eukaryote	TAXONOMY: Eukaryota	45751	102748
Only Bacteria	TAXONOMY: Bacteria (eubacteria)	31664	80664
Only Archaea	TAXONOMY: Archaea	3127	8366
Only Virus	TAXONOMY: Viruses	4629	18642
Unassigned	TAXONOMY: Unassigned	479	980
Recently deposited proteins solved by NMR in PDB (July 2012 to March 25 2014)	Experimental Method: solution NMR	657	751

* PDB date August, 2013 unless otherwise noted. Protein chain types only.

Proximity Length (P) and Normalized Proximity Length (NP)

After putative STP motifs are identified, a set of three proximity lengths are calculated: $P_1 = \Delta C_{1,4}$; $P_2 = \Delta C_{2,5}$; $P_3 = \Delta C_{3,6}$. Motifs of less than six cysteines, or motifs defined as invalid by our criteria, were assigned $P_1 = P_2 = P_3 = 0$. A Normalized Proximity Length (NP) was then assigned for each proximity length, P, resulting in three

new values: NP_1, NP_2, and NP_3. The NP identifies the distance from the observed mean proximity lengths of known STPs to the corresponding bonded cysteines involved in STP cysteine loops in the training set. For example, the average *P* for all STP sequences in the training set is subtracted from the calculated *P* value associated with its corresponding proximity length and normalized as described in *Eq. 1*, where $\bar{x}P_j$ is the average of the proximity lengths of known STPs derived from the training set.

$$\mathrm{NP}_{j\in\{1,2,3\}} = \frac{100}{(|P_j - \bar{x}\mathrm{P_j}| + 10)} \quad \text{(eq. 1)}$$

Detecting Least Loop Length Ratio

The least loop length is defined as the *min* ($\Delta C_{i,i+1}$) divided by the total length of the peptide. This feature is used as part of feature sets 5 and 6, see Table S2.3.

Detecting Presence of Amino Acid Between C4 -C5 and C5-C6

Data published describing loop lengths of ICKs and cyclotides, which comprise a large subset of STPs (Gracy et al. 2008a), motivated a Boolean feature for the presence of inter-loop amino acids. A result of 'true' is returned if there is a presence of a minimum of one amino acid in both of the last two loops (C4-C5 and C5-C6) in a putative STP motif.

Algorithm

We used a Support Vector Machine (SVM) classifier/predictor implementation to elucidate STP toxins. The SVM was implemented using the e1071 library in R (2.15.1). Feature sets were assigned as described in the Table 2.3, and sensitivity, specificity, precision and accuracy were determined after ten-fold cross validation. Initial gamma and

cost were set to 0.1 and 0.1, respectively, with the best output at 0.0587. Given 144 STP and 393 non-STP chains, 100 and 300 random samples were chosen, respectively, for a training set over 200 iterations. Feature sets were prioritized based on accuracy. STP sequences were predicted from the test sets described previously (Table 2.1) using feature set 6. Due to the limited throughput of the Knoter1D interface, only the "NewNMR751" and "Smallprotein163" (predicted STP chains from the SCOPs derived subset) predictions where compared against Knoter 1D predictions (http://knottin.cbs.cnrs.fr/Tools_1D.php) and validated with Jmol by analyzing the disulfide connectivity using the corresponding PDB files. Results from only the eukaryotic test sets were filtered to remove sequences with ≥ 30% chain identity and compared against Jmol analysis. Chains exhibiting canonical STP connectivity (C1-C4, C2-C5, C3-C6) were initially considered as true positives. True positives were further cross matched with their PDB annotations to make the final confirmation.

Confusion Matrix Creation

A confusion matrix was created to perform the cross-validation test. True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) were determined from the confusion matrix. Sensitivity [TP/(TP+FN)], specificity [TN/(TN+FP)], precision [TP/(TP+FP)], accuracy [(TP+TN)/(TP+FN+TN+FP)] and Mathews Correlation Coefficient (MCC) [(TPXTNFPXFN)/sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN) were calculated to evaluate the performance of the algorithm.

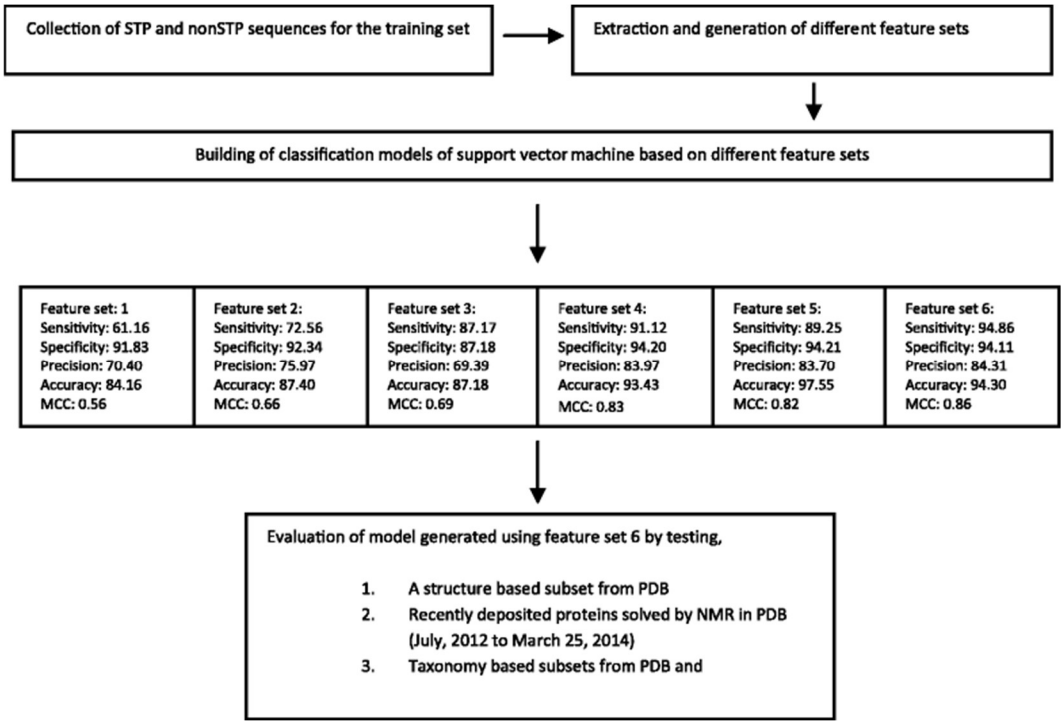

Figure 2.3: Schematic of the process followed to develop and evaluate the SVM based STP toxin classifier.

PSI BLAST

The BLAST suite (blast-2.2.29+) was installed on a local machine along with the appropriate dataset. The dataset was the chains of proteins deposited in PDB, solved by the NMR method, from July 04, 2012 to March 25, 2014. The selected threshold e-values PSI BLAST (Altschul et al. 1997) were 0.01, 0.1 and 0.5. The number of iterations for PSI BLAST was 5. All other parameters were set as default.

Results

Evaluation of Feature Sets for Machine Learning Outcomes

The training data set of 144 STP and 393 non-STP chains was evaluated using

randomized sampling over 200 iterations to determine the optimal feature sets. All of the 6 feature sets were examined (Table S 2.3), and the sensitivity, specificity, precision, accuracy and MCC scores were calculated (Figure 2.3). Feature set 6 demonstrated the best accuracy and MCC with values of 94.30 %, and 0.86, respectively, and was used for the basis of the remainder of the study. The Receptor Operating Curve (ROC) for feature set 6 is provided in the Figure 2.4. In the rest of the article, the model is referred to as PredSTP.

Classifying STPs from the Smallprotein163 Subset from PDB

The SmallProtein163 data subset from PDB was analyzed to determine potential automated STP classification. The median residue number of the chains in the Smallprotein163 subset is 54, which is similar to the number of residues in STP chains. In addition, 94 out of the 163 chains contain at least 6 cysteines in their primary sequences. From this subset, PredSTP was able to identify 21 of the 163 potential chains as STP-containing. These putative STP structures were verified by examining their disulfide bonding patterns in Jmol. Of the 21 identified chains by PredSTP, 14 of them were confirmed as true positives (Table 2.2). An analysis of the 142 negative STP chains predicted by PredSTP demonstrated only one false negative. The sensitivity, specificity, precision and accuracy for this particular dataset were 94.86%, 95.27 %, 66.66% and 95.09%, respectively (Table 2.3). PDB ids and functions for the positive predicted chains are provided in Table S2.5.

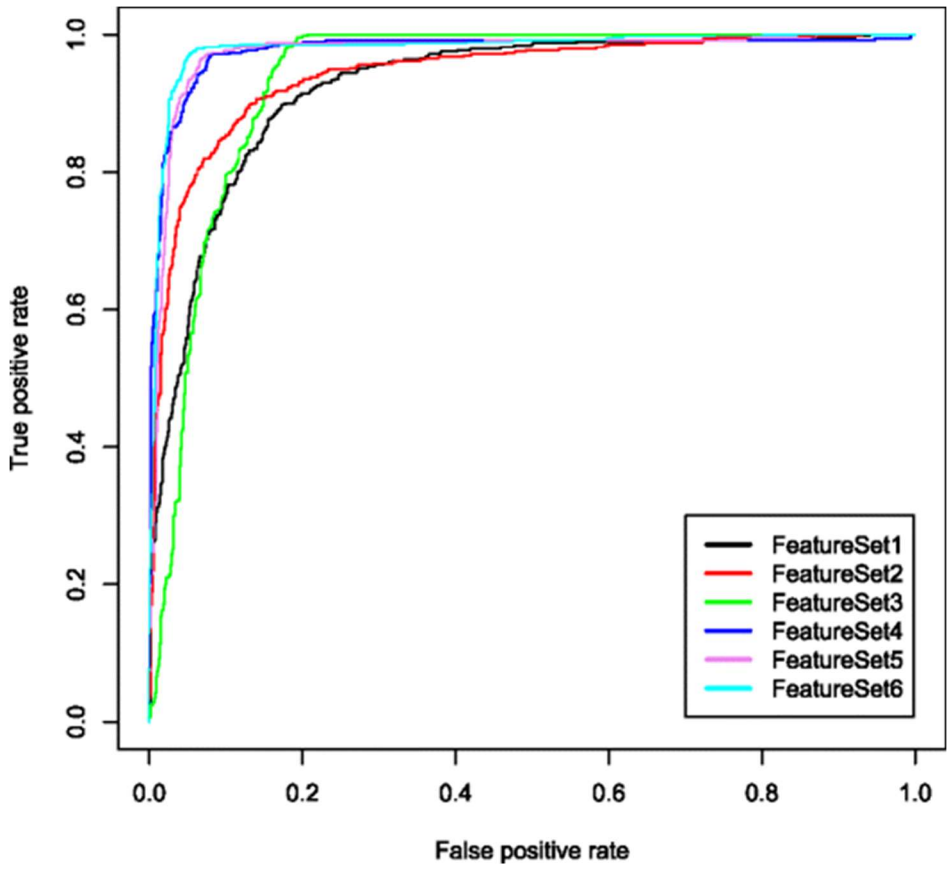

Figure 2.4: Receiver operating characteristic curves (ROC) for the models generated using 6 different feature sets. The area under curve (AUC) generated by feature set 1, 2, 3, 4, 5 and 6 are 0.84, 0.87, 0.87. 0.93, 0.92 and 0.94, respectively.

Table 2.2: Analysis of PredSTP positive hits from smallprotein92 subset

Total PredSTP positive chains	**TRUE positive**	**Knoter1D positive**
21	14/21	1/21

Testing Primary Sequences of Recently Deposited Proteins Solved by NMR (newNMR 751)

PredSTP was tested against protein sequences with less than 90% sequence identity and recently solved (July 04, 2012 to March 25, 2014) by NMR. This set of 751 amino acid chains is denoted as newNMR751 and has a median number of 82 residues

with 118 chains containing more than 6 cysteines. The model detected 23 chains from 23 different proteins. Analyzing the disulfide connectivity of the positive hits by Jmol, 21 chains were confirmed as true positive. Based on the number of the predicted outcomes, the sensitivity, specificity, precision and accuracy for this particular dataset were 91.30%, 99.72%, 91.30% and 99.46%, respectively (Table 2.3). The true positive chains were further classified into 9 ICKs, 5 cyclotides and 7 nonknotted STPs. PDB ids and functions for positive predictions are provided in Table S2.6. This set was also analyzed by PSI BLAST (Altschul et al. 1997) and Knoter1D (Gracy et al. 2008b). Knoter1D detected 5 cyclotides, 3 of the 9 ICKs and none of the nonknotted STPs. PSI BLAST (e-value 0.01) detected 12 chains comprising 1 ICK, 5 cyclotides, 5 nonknotted STPs and 1 false positive; PSI BLAST (e-value 0.1) detected 21 chains comprising 5 ICK, 5 cyclotides, 7 nonknotted STPs and 4 false positives; PSI BLAST (e-value 0.5) detected 52 chains comprising 5 ICK, 5 cyclotides, 7 nonknotted STPs and 35 false positives (Figure 2.5, Table 2.4, Table S2.7).

Table 2.3: Comparison of evaluation matrices generated by PredSTP using the training set, Smallprotein163 and NewNMR751 subsets from PDB. The confusion matrix generated by PredSTP using the corresponding datasets are provided in Table S2.4.

Source of data	Sensitivity	Specificity	Precision	Accuracy
Training set over 200 iterations	94.86	94.11	84.31	94.30
Smallprotein163	93.00	95.27	66.66	95.09
NewNMR751	91.30	99.58	91.30	99.46

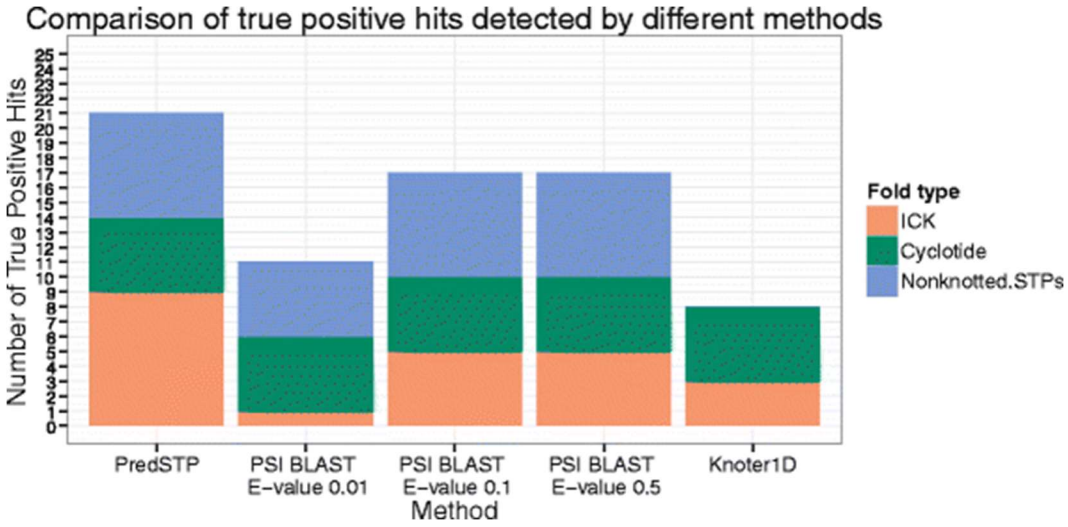

Figure 2.5: Bar diagram of a comparison the number of true positive hits detected by testing recently deposited proteins chains solved by NMR in PDB (July, 4 2012 to March, 25 2014) using different methods. Each stack color represents a different type of fold. PredSTP detected 9 ICKs, 5 Cyclotides and 6nonknotted STPs; PSI BLAST with E-value 0.01 detected 1 ICK, 5 Cyclotides and 5 nonknotted STPs; PSI BLAST with E-value 0.1 and 0.5 detected 5 ICKs, 5 Cyclotides and 7 nonknotted STPs; Knoter1D detected 3 ICKs and 5 Cyclotides.

Evaluation of the PredSTP through Scanning and Analyzing the Taxonomy Subsets from PDB

Finally, after testing the performance of PredSTP against chains from the "SmallProtein163" and "NewNMR751" subsets, which consist of sequences of similar size to the training set, we tested against a set based on diverse taxonomy. We analyzed "Eukaryota", "Bacteria", "Viruses", "Archaea" and "Unassigned" subsets of proteins from the PDB (Table 2.5).The percentage of positive chains in "Eukaryote" (0.61) is more than the percentage of predicted positive chains for the other three major super kingdoms. In "Eukaryotes", 636 chains were predicted as STP positive. This number was reduced to 139 chains when chains sharing > 30% sequence similarity were removed and the first 100 chains (based on PDB id) were manually cross-matched with Jmol analysis to determine true positives. This resulted in a 82% precision rate (Table S2.8). In

“bacteria”, “virus” and “unassigned” subsets, the precisions were 50%, 33.33% and 90%, respectively (Table 2.6). In the "Archaea" subset, PredSTP did not predict any potential STP toxins, resulting in no precision. In total, 115 positive hits were analyzed from the "Taxonomy" subset and 93 chains were found as true positive with an overall 80.86% precision. Individual precision rates for bacteria and viruses were low; this is potentially an artifact of their small sizes. In addition, some bacteria may contain iron-sulfur like transport proteins that mimic STPs by primary structure but are functionally distinct. The number of protein chains containing a minimum of six cysteines and consisting of a maximum 75 residues were also calculated for the same taxonomy subsets from PDB, and the percentages of predicted STPs were 30.08, 6.66, 0, 14.81 and 47.61 for Eukaryotes, bacteria, archaea, virus and unassigned , respectively (Table 2.7).

Discussion

A wide array of toxic peptides, with varying bonding patterns, can be stabilized by disulfide bonds. A large number of these peptides include a sequentially paired disulfide bonding pattern (C1-C4, C2-C5, C3-C6), confirming a compact array of this cystine trio which we refer to here as Sequential Tri-disulfide Peptides (STP). This array includes the well-defined knottin and cyclotide groups that have knotted tertiary structures. They also include a large number of stable toxins that contain the STP bonding pattern but lack the knotted motif typically created by C3-C6 in knottins and cyclotides. Going beyond these groupings, there are other stable toxins that exhibit compact tri-disulfide bonding patterns, but not in the sequentially paired model, including the ladder-type toxins and what we have distinguished as NTPs (Figure 2.1).

Table 2.4: Comparison of number of hits detected by different methods in recently deposited proteins solved by NMR in PDB (July 2012 to March 25, 2014).

Method	Positive hits	**True positive hits**	**False positive hits**	**Calculated sensitivity (%) for STPs***	**Calculated precision (%) for STPs**
PredSTP	23	21	2	91.30	91.30
PSI BLAST with e-value 0.01	13	12	1	52.17	92.30
PSI BLAST with e-value 0.1	21	17	4	73.90	80.95
PSI BLAST with e-value 0.5	52	17	35	73.90	32.69
Knoter1D	8	8	0	57.14	100

* Sensitivity for PredSTP and PSI BLAST was calculated based on total experimentally positive STPs (22 chains) in the NewNMR subset from PDB, while sensitivity for Knoter1D was calculated only for Knottins (knotted STPs)

Table 2.5: Discovery of STPs across major domains using PDB protein sequence data and PredSTP.

PDB subset	**Total # of proteins analyzed**	**Total # of chains**	**Positive chains predicted by PredSTP**	**# of proteins containing positive chains**	**Percentage of positive chains**
Eukaryotes	45751	102748	636	139*	0.61
Eubacteria	31664	80664	3	2	0.003
Archaea	3127	8366	0	0	0
Viruses	4629	18642	4	3	0.02
Unassigned	479	980	10	10	1.02

*For eukaryotes, 139chains were obtained after screening 636 chains and removing those with ≥ 30% sequence identity.

Table 2.6: Comparison of positive hits detected by PredSTP in different taxonomy based subsets from PDB.

PDB subset	PredSTP positive hits	True (structurally) positives	% of true positives (Precision)
Eukaryotes	139	82(100)*	82
Bacteria	2	1	50
Archaea	0	0	NA
Viruses	3	1	33
Unassigned	10	9	90
Total	115*	93	80.86

*For eukaryotes, 100 of the 139 proteins were analyzed in Jmol to find true positives.

It is imperative that successful machine learning algorithms select proper training sets and features. We constructed our negative training set with a collection of small proteins verified from the NMR subset deposited in PDB between 2000 and 2010. They contain a similar number of total residues as STPs, and a number have tri-disulfide bonds (NTPs) in their 3D structure. After evaluating several feature sets, a combination of motif-based features and features based on individual amino acids (C, S, H, K, L) generated the best predictions, indicating that differentiation between STPs and nonSTPs lies in both inclusive motifs and primary sequences.

In order to evaluate the performance of PredSTP on out of sample data we developed several independent test sets. The Smallprotein163 and NewNMR751 sets from PDB consist of a substantial number of cysteine rich small proteins. PredSTP showed a better accuracy (95.09%) for Smallprotein163 than it did for the training set (94.30%), while the precision was comparatively low (66.66%). The only STP not detected (PDB id 2C4B) was a heterogonous fusion protein of an STP and a catalytically inactive variant of RNase barnase (Niemann et al. 2006). On the other hand, a test of

performance of PredSTP on the NewNMP751 subset showed an excellent accuracy (99.46%) with a better precision (90.30%) than it showed on the training set (Table 2.3). These results indicate that PredSTP retained its performance when distinguishing STPs from out of sample cysteine rich small proteins.

Knoter1D (Gracy et al. 2008a) and Cypred (Kedarisetti et al. 2014) are examples of related software to discover cystine stabilized peptide toxins. Cypred is dedicated for detecting cyclic peptides. Knoter 1D is optimized to identify only knotted STPs using an algorithm that implements BLAST and is dependent on sequence identity with known knotted STPs. This approach does not allow Knoter1D to expand the inclusion of knotted STPs beyond a threshold of sequence identity. However, both knotted and non-knotted STPs vary in their sequences depending on the source organism. To compare our sequence independent algorithm to these approaches, we used the recently deposited protein structure in PDB (NewNMR751). Knoter1D detected only 8 out of 14 knotted STPs (ICKs and cyclotides) and did not detect six new ICKs as they differ significantly from the sequences of the known ICKs (knotted STPs)(Figure 2.5). While we compared PredSTP with PSI-BLAST, we used three different E-values to obtain the optimum result from PSI BLAST. Among the three versions, PSI BLAST with E-value 0.1 can detect 21 chains that exhibit the highest sensitivity with a minimum number of 4 false positives. On the other hand, PredSTP detected 21 STPs including the six new ICKs missed by the detection method of Knoter 1D and PSI BLAST. Therefore, in terms of detecting all type of STPs (cyclotides, ICKs and nonknotted STPs), PredSTP demonstrates better sensitivity and precision than PSI BLAST (Table 2.4).

In order to illustrate the capability of predicting tri-disulfide bonded peptides using PredSTP, we utilized the known paucity of disulfide bonding in bacteria and archaea as compared to eukaryotes (Bosnjak et al. 2014). We anticipated a higher proportion of STPs in eukaryotes with respect to the total number of cysteine chains with a maximum of 75 residues and a minimum of six cysteines. The threshold of 75 is chosen because it is well below the length of the longest chain (86 residues long) detected as STP by PredSTP among taxonomy subsets. After testing protein chains from different organismal taxonomy subsets in PDB, we confirmed this by observing that only 6.66% and 0% of chains possessing a minimum of six cysteines and maximum 75 residues were predicted as STPs in bacteria and archaea, respectively (Table 2.7). In contrast, 30% of the small cysteine-containing chains were predicted as STPs in eukaryotes.

Conclusion

PredSTP is capable of predicting STP toxins containing a compact tri-disulfide domain and exhibiting identical functional properties in a sequence identity independent manner. Our algorithm implements an automated method to find cystine stabilized toxins containing a compact arrangement of tri-disulfide domain with minimal sequence identity. Therefore, this approach provides useful directions for enhancement of theoretical and experimental research to find new antimicrobial peptides, insecticides and other stable peptide drug candidates by shortening the discovery time of potential bioactive peptides. Further research may benefit from a model that classifies all cystine stabilized peptide toxins (inhibitor or antimicrobial) into the different subgroups based on source, mode of action, and target organisms.

Authors Contributions

SI conducted the primary investigation, including data aggregation and computation and drafted the manuscript, TS developed the machine learning approach and on-line analysis tools, CK designed the study and participated in the manuscript, EB aided in the design of the study and developed the manuscript. All authors have read and approved the manuscript.

References

Aloush, Valerie, Shiri Navon-Venezia, Yardena Seigman-Igra, Shaltiel Cabili, and Yehuda Carmeli. 2006. "Multidrug-Resistant Pseudomonas Aeruginosa: Risk Factors and Clinical Impact." *Antimicrobial Agents and Chemotherapy* 50 (1): 43–48. https://doi.org/10.1128/AAC.50.1.43-48.2006.

Altschul, S F, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25 (17): 3389–3402.

Beilen, J. B. van, M. Neuenschwander, T. H. M. Smits, C. Roth, S. B. Balada, and B. Witholt. 2002. "Rubredoxins Involved in Alkane Oxidation." *Journal of Bacteriology* 184 (6): 1722–32. https://doi.org/10.1128/JB.184.6.1722-1732.2002.

Bende, Niraj S., Sławomir Dziemborowicz, Mehdi Mobli, Volker Herzig, John Gilchrist, Jordan Wagner, Graham M. Nicholson, Glenn F. King, and Frank Bosmans. 2014. "A Distinct Sodium Channel Voltage-Sensor Locus Determines Insect Selectivity of the Spider Toxin Dc1a." *Nature Communications* 5: 4350. https://doi.org/10.1038/ncomms5350.

Bock, J R, and D A Gough. 2001. "Predicting Protein--Protein Interactions from Primary Structure." *Bioinformatics (Oxford, England)* 17 (5): 455–60.

Bosnjak, I., V. Bojovic, T. Segvic-Bubic, and A. Bielen. 2014. "Occurrence of Protein Disulfide Bonds in Different Domains of Life: A Comparison of Proteins from the Protein Data Bank." *Protein Engineering Design and Selection* 27 (3): 65–72. https://doi.org/10.1093/protein/gzt063.

Braunstein, A., N. Papo, and Y. Shai. 2004. "In Vitro Activity and Potency of an Intravenously Injected Antimicrobial Peptide and Its DL Amino Acid Analog in Mice Infected with Bacteria." *Antimicrobial Agents and Chemotherapy* 48 (8): 3127–29. https://doi.org/10.1128/AAC.48.8.3127-3129.2004.

Brooke, B D, R H Hunt, F Chandre, P Carnevale, and M Coetzee. 2002. “Stable Chromosomal Inversion Polymorphisms and Insecticide Resistance in the Malaria Vector Mosquito Anopheles Gambiae (Diptera: Culicidae).” *Journal of Medical Entomology* 39 (4): 568–73.

Bulet, P, C Hetru, J L Dimarcq, and D Hoffmann. 1999. “Antimicrobial Peptides in Insects; Structure and Function.” *Developmental and Comparative Immunology* 23 (4–5): 329–44.

Cai, C Z, L Y Han, Z L Ji, X Chen, and Y Z Chen. 2003. “SVM-Prot: Web-Based Support Vector Machine Software for Functional Classification of a Protein from Its Primary Sequence.” *Nucleic Acids Research* 31 (13): 3692–97.

Cai, Y. D., X. J. Liu, X. Xu, and G. P. Zhou. 2001. “Support Vector Machines for Predicting Protein Structural Class.” *BMC Bioinformatics* 2: 3.

Carlini, Célia R, and Maria Fátima Grossi-de-Sá. 2002. “Plant Toxic Proteins with Insecticidal Properties. A Review on Their Potentialities as Bioinsecticides.” *Toxicon: Official Journal of the International Society on Toxinology* 40 (11): 1515–39.

Circo, Raffaella, Barbara Skerlavaj, Renato Gennaro, Antonio Amoroso, and Margherita Zanetti. 2002. “Structural and Functional Characterization of HBD-1(Ser35), a Peptide Deduced from a DEFB1 Polymorphism.” *Biochemical and Biophysical Research Communications* 293 (1): 586–92. https://doi.org/10.1016/S0006-291X(02)00267-X.

Conibear, Anne C., Alexander Bochen, K. Johan Rosengren, Petar Stupar, Conan Wang, Horst Kessler, and David J. Craik. 2014. “The Cyclic Cystine Ladder of Theta-Defensins as a Stable, Bifunctional Scaffold: A Proof-of-Concept Study Using the Integrin-Binding RGD Motif.” *ChemBioChem* 15 (3): 451–59. https://doi.org/10.1002/cbic.201300568.

Conibear, Anne C., K. Johan Rosengren, Norelle L. Daly, Sónia Troeira Henriques, and David J. Craik. 2013. “The Cyclic Cystine Ladder in θ-Defensins Is Important for Structure and Stability, but Not Antibacterial Activity.” *The Journal of Biological Chemistry* 288 (15): 10830–40. https://doi.org/10.1074/jbc.M113.451047.

Edgar, R. C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97. https://doi.org/10.1093/nar/gkh340.

Emeleus, H. J. 1959. *Advances in Inorganic Chemistry*. New York: Academic Press.

Gelly, Jean-Christophe, Jérôme Gracy, Quentin Kaas, Dung Le-Nguyen, Annie Heitz, and Laurent Chiche. 2004. “The KNOTTIN Website and Database: A New Information System Dedicated to the Knottin Scaffold.” *Nucleic Acids Research* 32 (Database issue): D156-159. https://doi.org/10.1093/nar/gkh015.

Góngora-Benítez, Miriam, Judit Tulla-Puche, and Fernando Albericio. 2014. "Multifaceted Roles of Disulfide Bonds. Peptides as Therapeutics." *Chemical Reviews* 114 (2): 901–26. https://doi.org/10.1021/cr400031z.

Gordon, Y. Jerold, Eric G. Romanowski, and Alison M. McDermott. 2005. "A Review of Antimicrobial Peptides and Their Therapeutic Potential as Anti-Infective Drugs." *Current Eye Research* 30 (7): 505–15. https://doi.org/10.1080/02713680590968637.

Gould, Andrew, Yanbin Ji, Teshome L. Aboye, and Julio A. Camarero. 2011. "Cyclotides, a Novel Ultrastable Polypeptide Scaffold for Drug Discovery." *Current Pharmaceutical Design* 17 (38): 4294–4307.

Gracy, Jérôme, Dung Le-Nguyen, Jean-Christophe Gelly, Quentin Kaas, Annie Heitz, and Laurent Chiche. 2008a. "KNOTTIN: The Knottin or Inhibitor Cystine Knot Scaffold in 2007." *Nucleic Acids Research* 36 (Database issue): D314-319. https://doi.org/10.1093/nar/gkm939.

Hemingway, J, and H Ranson. 2000. "Insecticide Resistance in Insect Vectors of Human Disease." *Annual Review of Entomology* 45: 371–91. https://doi.org/10.1146/annurev.ento.45.1.371.

Henriques, Sónia Troeira, and David J. Craik. 2010. "Cyclotides as Templates in Drug Design." *Drug Discovery Today* 15 (1–2): 57–64. https://doi.org/10.1016/j.drudis.2009.10.007.

Herzig, Volker, David L A Wood, Felicity Newell, Pierre-Alain Chaumeil, Quentin Kaas, Greta J Binford, Graham M Nicholson, Dominique Gorse, and Glenn F King. 2011. "ArachnoServer 2.0, an Updated Online Resource for Spider Toxin Sequences and Structures." *Nucleic Acids Research* 39 (Database issue): D653-657. https://doi.org/10.1093/nar/gkq1058.

Hiramatsu, K, N Aritaka, H Hanaki, S Kawasaki, Y Hosoda, S Hori, Y Fukuchi, and I Kobayashi. 1997. "Dissemination in Japanese Hospitals of Strains of Staphylococcus Aureus Heterogeneously Resistant to Vancomycin." *Lancet* 350 (9092): 1670–73. https://doi.org/10.1016/S0140-6736(97)07324-8.

Hua, S., and Z. Sun. 2001. "Support Vector Machine Approach for Protein Subcellular Localization Prediction." *Bioinformatics* 17 (8): 721–28. https://doi.org/10.1093/bioinformatics/17.8.721.

Hua, Sujun, and Zhirong Sun. 2001. "A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach." *Journal of Molecular Biology* 308 (2): 397–407. https://doi.org/10.1006/jmbi.2001.4580.

Huang, Y., B. Niu, Y. Gao, L. Fu, and W. Li. 2010. “CD-HIT Suite: A Web Server for Clustering and Comparing Biological Sequences.” *Bioinformatics* 26 (5): 680–82. https://doi.org/10.1093/bioinformatics/btq003.

Jennings, Cameron V., K. Johan Rosengren, Norelle L. Daly, Manuel Plan, Jackie Stevens, Martin J. Scanlon, Clement Waine, David G. Norman, Marilyn A. Anderson, and David J. Craik. 2005. “Isolation, Solution Structure, and Insecticidal Activity of Kalata B2, a Circular Protein with a Twist: Do Möbius Strips Exist in Nature?” *Biochemistry* 44 (3): 851–60. https://doi.org/10.1021/bi047837h.

Kaas, Quentin, Rilei Yu, Ai-Hua Jin, Sébastien Dutertre, and David J Craik. 2012. “ConoServer: Updated Content, Knowledge, and Discovery Tools in the Conopeptide Database.” *Nucleic Acids Research* 40 (Database issue): D325-330. https://doi.org/10.1093/nar/gkr886.

Kedarisetti, Pradyumna, Marcin J. Mizianty, Quentin Kaas, David J. Craik, and Lukasz Kurgan. 2014. “Prediction and Characterization of Cyclic Proteins from Sequences in Three Domains of Life.” *Biochimica Et Biophysica Acta* 1844 (1 Pt B): 181–90. https://doi.org/10.1016/j.bbapap.2013.05.002.

Lehrer, R I, A K Lichtenstein, and T Ganz. 1993. “Defensins: Antimicrobial and Cytotoxic Peptides of Mammalian Cells.” *Annual Review of Immunology* 11: 105–28. https://doi.org/10.1146/annurev.iy.11.040193.000541.

Lewis, Richard J., and Maria L. Garcia. 2003. “Therapeutic Potential of Venom Peptides.” *Nature Reviews Drug Discovery* 2 (10): 790–802. https://doi.org/10.1038/nrd1197.

Li, W., and A. Godzik. 2006. “Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.” *Bioinformatics* 22 (13): 1658–59. https://doi.org/10.1093/bioinformatics/btl158.

Marr, Alexandra K., William J. Gooderham, and Robert Ew Hancock. 2006. “Antibacterial Peptides for Therapeutic Use: Obstacles and Realistic Outlook.” *Current Opinion in Pharmacology* 6 (5): 468–72. https://doi.org/10.1016/j.coph.2006.04.006.

Matsumura, M., G. Signor, and B. W. Matthews. 1989. “Substantial Increase of Protein Stability by Multiple Disulphide Bonds.” *Nature* 342 (6247): 291–93. https://doi.org/10.1038/342291a0.

Monroc, Sylvie, Esther Badosa, Lidia Feliu, Marta Planas, Emili Montesinos, and Eduard Bardají. 2006. “De Novo Designed Cyclic Cationic Peptides as Inhibitors of Plant Pathogenic Bacteria.” *Peptides* 27 (11): 2567–74. https://doi.org/10.1016/j.peptides.2006.04.019.

Muggleton, S, R D King, and M J Sternberg. 1992. “Protein Secondary Structure Prediction Using Logic-Based Machine Learning.” *Protein Engineering* 5 (7): 647–57.

Mulvenna, Jason P, Conan Wang, and David J Craik. 2006. “CyBase: A Database of Cyclic Protein Sequence and Structure.” *Nucleic Acids Research* 34 (Database issue): D192-194. https://doi.org/10.1093/nar/gkj005.

Niemann, Hartmut H., Hans-Ulrich Schmoldt, Alexander Wentzel, Harald Kolmar, and Dirk W. Heinz. 2006. “Barnase Fusion as a Tool to Determine the Crystal Structure of the Small Disulfide-Rich Protein McoEeTI.” *Journal of Molecular Biology* 356 (1): 1–8. https://doi.org/10.1016/j.jmb.2005.11.005.

Ovchinnikova, Tatiana V., Sergey V. Balandin, Galina M. Aleshina, Andrey A. Tagaev, Yulia F. Leonova, Eugeny D. Krasnodembsky, Alexander V. Men’shenin, and Vladimir N. Kokryakov. 2006. “Aurelin, a Novel Antimicrobial Peptide from Jellyfish Aurelia Aurita with Structural Features of Defensins and Channel-Blocking Toxins.” *Biochemical and Biophysical Research Communications* 348 (2): 514–23. https://doi.org/10.1016/j.bbrc.2006.07.078.

Reddy, K. V. R., R. D. Yedery, and C. Aranha. 2004. “Antimicrobial Peptides: Premises and Promises.” *International Journal of Antimicrobial Agents* 24 (6): 536–47. https://doi.org/10.1016/j.ijantimicag.2004.09.005.

Schroeder, Bjoern O., Zhihong Wu, Sabine Nuding, Sandra Groscurth, Moritz Marcinowski, Julia Beisner, Johannes Buchner, Martin Schaller, Eduard F. Stange, and Jan Wehkamp. 2011. “Reduction of Disulphide Bonds Unmasks Potent Antimicrobial Activity of Human β-Defensin 1.” *Nature* 469 (7330): 419–23. https://doi.org/10.1038/nature09674.

Tugyi, Regina, Gábor Mezö, Erzsébet Fellinger, David Andreu, and Ferenc Hudecz. 2005. “The Effect of Cyclization on the Enzymatic Degradation of Herpes Simplex Virus Glycoprotein D Derived Epitope Peptide.” *Journal of Peptide Science: An Official Publication of the European Peptide Society* 11 (10): 642–49. https://doi.org/10.1002/psc.669.

Wang, Conan K L, Quentin Kaas, Laurent Chiche, and David J Craik. 2008. “CyBase: A Database of Cyclic Protein Sequences and Structures, with Applications in Protein Discovery and Engineering.” *Nucleic Acids Research* 36 (Database issue): D206-210. https://doi.org/10.1093/nar/gkm953.

Ye, Mingyu, Keith K. Khoo, Shaoqiong Xu, Mi Zhou, Nonlawat Boonyalai, Matthew A. Perugini, Xiaoxia Shao, et al. 2012. “A Helical Conotoxin from Conus Imperialis Has a Novel Cysteine Framework and Defines a New Superfamily.” *The Journal of Biological Chemistry* 287 (18): 14973–83. https://doi.org/10.1074/jbc.M111.334615.

Zhu, Shunyi, Steve Peigneur, Bin Gao, Lan Luo, Di Jin, Yong Zhao, and Jan Tytgat. 2011a. “Molecular Diversity and Functional Evolution of Scorpion Potassium Channel Toxins.” *Molecular & Cellular Proteomics: MCP* 10 (2): M110.002832. https://doi.org/10.1074/mcp.M110.002832.

Table S2.1: PDB ID of control STP chains

>1ACW:A|PDBID
>1AG7:A|PDBID
>1AGG:A|PDBID
>1AGT:A|PDBID
>1AHO:A|PDBID
>1AXH:A|PDBID
>1AYJ:A|PDBID
>1BCG:A|PDBID
>1BH4:A|PDBID
>1BIG:A|PDBID
>1BK8:A|PDBID
>1BKT:A|PDBID
>1BX7:A|PDBID
>1C49:A|PDBID
>1C56:A|PDBID
>1C6W:A|PDBID
>1CHL:A|PDBID
>1CIX:A|PDBID
>1CLV:I|PDBID
>1CMR:A|PDBID
>1CN2:A|PDBID
>1DJT:A|PDBID
>1DKC:A|PDBID
>1DL0:A|PDBID
>1DQ7:A|PDBID
>1EIT:A|PDBID
>1EMX:A|PDBID
>1EYO:A|PDBID
>1F3K:A|PDBID
>1FH3:A|PDBID
>1FJN:A|PDBID
>1FSB:A|PDBID
>1FU3:A|PDBID
>1FYG:A|PDBID
>1G1Z:A|PDBID
>1G9P:A|PDBID
>1GPS:A|PDBID
>1GPT:A|PDBID
>1H20:A|PDBID
>1HA9:A|PDBID

>1HAE:A|PDBID
>1HLY:A|PDBID
>1HP2:A|PDBID
>1HVW:A|PDBID
>1HY9:A|PDBID
>1I26:A|PDBID
>1IE6:A|PDBID
>1IP0:A|PDBID
>1IXT:A|PDBID
>1J5J:A|PDBID
>1JKZ:A|PDBID
>1JLZ:A|PDBID
>1JU8:A|PDBID
>1JXC:A|PDBID
>1K36:A|PDBID
>1K48:A|PDBID
>1KAL:A|PDBID
>1KCP:A|PDBID
>1KOZ:A|PDBID
>1KQI:A|PDBID
>1KV0:A|PDBID
>1L3Y:A|PDBID
>1L4V:A|PDBID
>1LA4:A|PDBID
>1LIR:A|PDBID
>1LMM:A|PDBID
>1LMR:A|PDBID
>1LU0:A|PDBID
>1LUP:A|PDBID
>1M2S:A|PDBID
>1MB6:A|PDBID
>1MCT:I|PDBID
>1MCV:I|PDBID
>1MM0:A|PDBID
>1MMC:A|PDBID
>1MR4:A|PDBID
>1MTX:A|PDBID
>1MVJ:A|PDBID
>1MYN:A|PDBID
>1N8M:A|PDBID

>1NBJ:A|PDBID
>1NH5:A|PDBID
>1NIX:A|PDBID
>1NPI:A|PDBID
>1NRA:A|PDBID
>1OAV:A|PDBID
>1OMY:A|PDBID
>1OZZ:A|PDBID
>1P8B:A|PDBID
>1PE4:A|PDBID
>1PJV:A|PDBID
>1PNH:A|PDBID
>1PT4:A|PDBID
>1PVZ:A|PDBID
>1PX9:A|PDBID
>1Q3J:A|PDBID
>1Q9B:A|PDBID
>1QDP:A|PDBID
>1QK6:A|PDBID
>1QK7:A|PDBID
>1QKY:A|PDBID
>1R1F:A|PDBID
>1R1G:A|PDBID
>1RMK:A|PDBID
>1RYG:A|PDBID
>1SCO:A|PDBID
>1SCY:A|PDBID
>1SEG:A|PDBID
>1SIS:A|PDBID
>1SN4:A|PDBID
>1SNB:A|PDBID
>1SXM:A|PDBID
>1T0Z:A|PDBID
>1TSK:A|PDBID
>1TTK:A|PDBID
>1TTL:A|PDBID
>1TYK:A|PDBID
>1UDK:A|PDBID
>1UGL:A|PDBID
>1UOY:A|PDBID

>1V7F:A|PDBID	>1WPD:A|PDBID	>2B3C:A|PDBID
>1VB8:A|PDBID	>1WQJ:B|PDBID	>2BMT:A|PDBID
>1VNA:A|PDBID	>1WT7:A|PDBID	>2BRZ:A|PDBID
>1VTX:A|PDBID	>1XDT:R|PDBID	>2C4B:A|PDBID
>1W7Z:A|PDBID	>1Y29:A|PDBID	>2PO8:A|PDBID
>1WM7:A|PDBID	>1ZNU:A|PDBID	>2SN3:A|PDBID
>1WM8:A|PDBID	>2A9H:E|PDBID	>2UVS:A|PDBID
>1WMT:A|PDBID	>2ASC:A|PDBID	>2Z3S:A|PDBID

Table S2.2: PDB ID of control nonSTP chains

>1ADZ:A|PDBID	>1E88:A|PDBID	>1IYM:A|PDBID
>1AFP:A|PDBID	>1E8P:A|PDBID	>1J7M:A|PDBID
>1ANS:A|PDBID	>1E9T:A|PDBID	>1JC6:A|PDBID
>1APJ:A|PDBID	>1ED0:A|PDBID	>1JFN:A|PDBID
>1ATA:A|PDBID	>1EFE:A|PDBID	>1JMN:A|PDBID
>1AW6:A|PDBID	>1F5Y:A|PDBID	>1JMP:A|PDBID
>1B2I:A|PDBID	>1F81:A|PDBID	>1JRF:A|PDBID
>1B9G:A|PDBID	>1F8Z:A|PDBID	>1K18:A|PDBID
>1BBG:A|PDBID	>1FAQ:A|PDBID	>1K7B:A|PDBID
>1BEI:A|PDBID	>1FBR:A|PDBID	>1KBE:A|PDBID
>1BF0:A|PDBID	>1FRE:A|PDBID	>1KDU:A|PDBID
>1BGK:A|PDBID	>1FVL:A|PDBID	>1KG1:A|PDBID
>1BOR:A|PDBID	>1FYB:A|PDBID	>1KGM:A|PDBID
>1BUS:A|PDBID	>1G25:A|PDBID	>1KIO:A|PDBID
>1C2U:A|PDBID	>1G4F:A|PDBID	>1KJ0:A|PDBID
>1C9Q:A|PDBID	>1GKG:A|PDBID	>1KMA:A|PDBID
>1CCV:A|PDBID	>1GKN:A|PDBID	>1KMX:A|PDBID
>1CE3:A|PDBID	>1H0Z:A|PDBID	>1KS0:A|PDBID
>1CLD:A|PDBID	>1H7V:A|PDBID	>1KSQ:A|PDBID
>1CO4:A|PDBID	>1HA8:A|PDBID	>1KUN:A|PDBID
>1COU:A|PDBID	>1HCC:A|PDBID	>1L3H:A|PDBID
>1CR8:A|PDBID	>1HD4:A|PDBID	>1L3X:A|PDBID
>1CXW:A|PDBID	>1HFH:A|PDBID	>1LD6:A|PDBID
>1D2L:A|PDBID	>1HFI:A|PDBID	>1LDL:A|PDBID
>1D4U:A|PDBID	>1HKY:A|PDBID	>1LDR:A|PDBID
>1D6G:A|PDBID	>1HN6:A|PDBID	>1LPV:A|PDBID
>1D6G:B|PDBID	>1HPJ:A|PDBID	>1M8B:A|PDBID
>1DEM:A|PDBID	>1HX2:A|PDBID	>1M9O:A|PDBID
>1DQC:A|PDBID	>1IGL:A|PDBID	>1MGX:A|PDBID
>1DTK:A|PDBID	>1IRH:A|PDBID	>1MKC:A|PDBID
>1DX8:A|PDBID	>1IW4:A|PDBID	>1MKN:A|PDBID
>1E4U:A|PDBID	>1IYC:A|PDBID	>1MPZ:A|PDBID

>1N0Z:A|PDBID
>1N5G:A|PDBID
>1N87:A|PDBID
>1NBL:A|PDBID
>1NJ3:A|PDBID
>1NWV:A|PDBID
>1ORL:A|PDBID
>1OSX:A|PDBID
>1PB5:A|PDBID
>1PCE:A|PDBID
>1PCP:A|PDBID
>1PDC:A|PDBID
>1PK2:A|PDBID
>1PMC:A|PDBID
>1PMX:A|PDBID
>1PMX:B|PDBID
>1PPQ:A|PDBID
>1PS2:A|PDBID
>1PXE:A|PDBID
>1PYC:A|PDBID
>1Q3M:A|PDBID
>1QBH:A|PDBID
>1QGB:A|PDBID
>1QO6:A|PDBID
>1R79:A|PDBID
>1RMJ:A|PDBID
>1RO3:A|PDBID
>1RO4:A|PDBID
>1SHP:A|PDBID
>1SJU:A|PDBID
>1SP7:A|PDBID
>1SRZ:A|PDBID
>1SS3:A|PDBID
>1SSL:A|PDBID
>1SSU:A|PDBID
>1T1H:A|PDBID
>1T50:A|PDBID
>1TBN:A|PDBID
>1TCP:A|PDBID
>1TFI:A|PDBID
>1TFQ:A|PDBID
>1TIH:A|PDBID
>1TOT:A|PDBID
>1TPG:A|PDBID
>1TPM:A|PDBID

>1U34:A|PDBID
>1U5M:A|PDBID
>1UL4:A|PDBID
>1UL5:A|PDBID
>1URK:A|PDBID
>1UUA:A|PDBID
>1UUC:A|PDBID
>1V5N:A|PDBID
>1V87:A|PDBID
>1VD4:A|PDBID
>1VFI:A|PDBID
>1VIB:A|PDBID
>1WD2:A|PDBID
>1WEO:A|PDBID
>1WFE:A|PDBID
>1WFF:A|PDBID
>1WFH:A|PDBID
>1WFL:A|PDBID
>1WFP:A|PDBID
>1WG2:A|PDBID
>1WGE:A|PDBID
>1WGM:A|PDBID
>1WHE:A|PDBID
>1WII:A|PDBID
>1WIM:A|PDBID
>1WJ0:A|PDBID
>1WJ2:A|PDBID
>1WO9:A|PDBID
>1WVK:A|PDBID
>1X4S:A|PDBID
>1XFE:A|PDBID
>1XU6:A|PDBID
>1XUT:A|PDBID
>1YWS:A|PDBID
>1Z60:A|PDBID
>1ZFI:A|PDBID
>2AQA:A|PDBID
>2C6A:A|PDBID
>2CKU:A|PDBID
>2CON:A|PDBID
>2CQE:A|PDBID
>2CR8:A|PDBID
>2CS3:A|PDBID
>2CSV:A|PDBID
>2CT7:A|PDBID

>2D8Q:A|PDBID
>2D8U:A|PDBID
>2D8V:A|PDBID
>2DAN:A|PDBID
>2DID:A|PDBID
>2DIP:A|PDBID
>2DJ8:A|PDBID
>2DJA:A|PDBID
>2DKT:A|PDBID
>2DQ5:A|PDBID
>2ERS:A|PDBID
>2EYA:A|PDBID
>2FC6:A|PDBID
>2FC7:A|PDBID
>2FFT:A|PDBID
>2FN2:A|PDBID
>2GQE:A|PDBID
>2HGF:A|PDBID
>2JQ8:A|PDBID
>2JW6:A|PDBID
>2PJF:A|PDBID
>2UZG:A|PDBID
>3ALC:A|PDBID
>3LRI:A|PDBID
>1ARD:A|PDBID
>1BBO:A|PDBID
>1BHI:A|PDBID
>1FV5:A|PDBID
>1JN7:A|PDBID
>1M36:A|PDBID
>1NCS:A|PDBID
>1NJQ:A|PDBID
>1P7A:A|PDBID
>1PAA:A|PDBID
>1SRK:A|PDBID
>1U85:A|PDBID
>1U86:A|PDBID
>1VA2:A|PDBID
>1VA3:A|PDBID
>1WIR:A|PDBID
>1WJP:A|PDBID
>1WJV:A|PDBID
>1X3C:A|PDBID
>1X5W:A|PDBID
>1X6E:A|PDBID

>1X6F:A|PDBID
>1X6H:A|PDBID
>1XF7:A|PDBID
>1XRZ:A|PDBID
>1ZFD:A|PDBID
>1ZNF:A|PDBID
>1ZR9:A|PDBID
>1ZU1:A|PDBID
>2ADR:A|PDBID
>2COT:A|PDBID
>2CSH:A|PDBID
>2CT1:A|PDBID
>2CT5:A|PDBID
>2CTD:A|PDBID
>2DLK:A|PDBID
>2DLQ:A|PDBID
>2DMD:A|PDBID
>2EPP:A|PDBID
>2EPQ:A|PDBID
>2EPR:A|PDBID
>2EPS:A|PDBID
>2GHF:A|PDBID
>2VRD:A|PDBID
>2VY5:A|PDBID
>2YRK:A|PDBID
>2YT9:A|PDBID
>3ZNF:A|PDBID
>7ZNF:A|PDBID
>1ARD:A|PDBID
>1BBO:A|PDBID
>1BHI:A|PDBID
>1FV5:A|PDBID
>1JN7:A|PDBID
>1M36:A|PDBID
>1NCS:A|PDBID
>1NJQ:A|PDBID
>1P7A:A|PDBID
>1PAA:A|PDBID
>1SRK:A|PDBID
>1U85:A|PDBID
>1U86:A|PDBID
>1VA2:A|PDBID
>1VA3:A|PDBID
>1WIR:A|PDBID
>1WJP:A|PDBID

>1WJV:A|PDBID
>1X3C:A|PDBID
>1X5W:A|PDBID
>1X6E:A|PDBID
>1X6F:A|PDBID
>1X6H:A|PDBID
>1XF7:A|PDBID
>1XRZ:A|PDBID
>1ZFD:A|PDBID
>1ZNF:A|PDBID
>1ZR9:A|PDBID
>1ZU1:A|PDBID
>2ADR:A|PDBID
>2COT:A|PDBID
>2CSH:A|PDBID
>2CT1:A|PDBID
>2CT5:A|PDBID
>2CTD:A|PDBID
>2DLK:A|PDBID
>2DLQ:A|PDBID
>2DMD:A|PDBID
>2EPP:A|PDBID
>2EPQ:A|PDBID
>2EPR:A|PDBID
>2EPS:A|PDBID
>2GHF:A|PDBID
>2VRD:A|PDBID
>2VY5:A|PDBID
>2YRK:A|PDBID
>2YT9:A|PDBID
>3ZNF:A|PDBID
>7ZNF:A|PDBID
>1AQS:A|PDBID
>1DFS:A|PDBID
>1DFT:A|PDBID
>1DMC:A|PDBID
>1DME:A|PDBID
>1FMY:A|PDBID
>1J5L:A|PDBID
>1J5M:A|PDBID
>1JI9:A|PDBID
>1M0G:A|PDBID
>1M0J:A|PDBID
>1MHU:A|PDBID
>1MRT:A|PDBID

>1QJK:A|PDBID
>1QJL:A|PDBID
>1T2Y:A|PDBID
>2MHU:A|PDBID
>2MRB:A|PDBID
>2MRT:A|PDBID
>1F62:A|PDBID
>1FP0:A|PDBID
>1HYI:A|PDBID
>1MM2:A|PDBID
>1MM3:A|PDBID
>1WE9:A|PDBID
>1WEE:A|PDBID
>1WEM:A|PDBID
>1WEN:A|PDBID
>1WEP:A|PDBID
>1WEQ:A|PDBID
>1WES:A|PDBID
>1WEU:A|PDBID
>1WEV:A|PDBID
>1WEW:A|PDBID
>1WFK:A|PDBID
>1WIL:A|PDBID
>2JWO:A|PDBID
>2K1J:A|PDBID
>1CDQ:A|PDBID
>1CHV:S|PDBID
>1COD:A|PDBID
>1CVO:A|PDBID
>1CXO:A|PDBID
>1DRS:A|PDBID
>1ERA:A|PDBID
>1FFJ:A|PDBID
>1G6M:A|PDBID
>1I02:A|PDBID
>1IJC:A|PDBID
>1JE9:A|PDBID
>1JGK:A|PDBID
>1KBS:A|PDBID
>1KS6:A|PDBID
>1LSI:A|PDBID
>1LXG:A|PDBID
>1LXG:B|PDBID
>1MR6:A|PDBID
>1NEA:A|PDBID

>1NTX:A|PDBID	>1APF:A|PDBID	>1KJ6:A|PDBID
>1PLO:A|PDBID	>1ATX:A|PDBID	>1SHI:A|PDBID
>1RGJ:A|PDBID	>1B8W:A|PDBID	>1UT3:A|PDBID
>1RGJ:B|PDBID	>1BDS:A|PDBID	>1Z99:A|PDBID
>1TFS:A|PDBID	>1BNB:A|PDBID	>1ZUF:A|PDBID
>1TXA:A|PDBID	>1E4R:A|PDBID	>2GW9:A|PDBID
>1VYC:A|PDBID	>1E4T:A|PDBID	>2JTO:A|PDBID
>1W6B:A|PDBID	>1EWS:A|PDBID	
>2CDX:A|PDBID	>1FQQ:A|PDBID	
>1AHL:A|PDBID	>1KJ5:A|PDBID	

Defined as > 95% matches to the following PDB Query: “Experimental method is SOLUTION NMR; SCOP is small proteins; chain type: there is a protein chain but not any DNA or RNA or hybrid; stoichiometry in biological assembly: stoichiometry is MONOMER and TAXONOMY is Eukaryota (eucaryotes) ;released between 2000 and 2010”

Table S2.3: Feature Sets Tested for SVM STP prediction. We extracted six unique sets of features for use in our machine learning protocol (Supplement Table 6). The first feature set was derived from a multiple sequence alignment (MSA) using MUSCLE(Edgar 2004) in MEGA 5.10. Here, each column was considered an independent feature, providing 318 unique features. Feature sets 2 – 6 were derived from a variety of sequence metadata, including composition and frequency of different amino acids, hydrophobicity, hydrophilicity, neutrality, bonding proximity score (defined below), total length of a chain and least loop to total length ratio (defined below), creating sets of 3, 23, 23, 28 and 28 features, respectively.

Feature Set 1 Comprises 23 distinct features, derived from calculating the frequency of occurrence of each amino acid plus the *frequency of occurrence* of aggregate hydrophobic (F,Y,L,I,A,M,C,W,V), hydrophilic (R,K,N,D,A,P) and neutral (G,H,S,T,Q) amino acids.

Feature Set 2: Comprises 23 distinct features, derived from calculating the number of occurrences of each amino acid plus the *aggregate number of occurrences* of hydrophobic (F,Y,L,I,A,M,C,W,V), hydrophilic (R,K,N,D,A,P), and neutral (G,H,S,T,Q) amino acids.**Feature Set 3:** Comprises three features derived from the Normalized Bonding Distance (NBD) between C1-C4, C2-C5 and C3-C6.

Feature Set 4: Comprises 7 distinct features, derived from the Normalized Bonding Distance (NBD) between C1-C4, C2-C5 and C3-C6, Presence of amino acid between C4 -C5 and C5-C6 , presence of double consecutive cysteines in the sequence, total peptide length and the least loop length ratio. The latter was calculated by dividing the length of the shortest ΔCij by the total length of the peptide.

Feature Set 5: Comprises 11 distinct features, derived from Feature set 4 , plus calculating the frequency of occurrences of cysteine, serine, arginine, histindine, lysine (C,S,R,H,K)(plus the *aggregate number of occurrences* of hydrophobic (F,Y,L,I,A,M,C,W,V), hydrophilic (R,K,N,D,A,P), and neutral (G,H,S,T,Q) amino acids.

Feature Set 6: Comprises 11 distinct features, derived from Feature set 4 , plus calculating the frequency of occurrences of cysteine, serine, arginine, histindine, lysine (C,S,R,H,K)(plus the *aggregate number of occurrences* of hydrophobic (F,Y,L,I,A,M,C,W,V), hydrophilic (R,K,N,D,A,P), and neutral (G,H,S,T,Q) amino acids.

Table S2.4: Confusion matrices generated by PredSTP using the training set, Smallprotein163 and NewNMR751 subsets from PDB.

Source of data	True positive	True negative	False positive	False negative
Training set over 200 iterations	18959	56537	3463	1041
Smallprotein163	14	141	7	1
NewNMR751	21	726	2	2

Table S2.5: List and description of 21 positively predicted proteins in "Smallprotein163" subset from PDB.

PDB ID[1]	Domain stabilized by tri-disulfide bonds	Disulfide connectivity[2]	Knoter1D	Function/Class
*1AHO	Yes	(C1-C4, C2-C5, C3-C6)	No	Scorpion Neurotoxin
1BX7	Yes	Array is not compact or absent	No	Serine Protease Inhibitor
1BX8	Yes	Array is not compact or absent	No	Serine Protease Inhibitor
*1DJT:A	Yes	(C1-C4, C2-C5, C3-C6)	No	Alpha-like Neurotoxin
*1DJT:B	Yes	(C1-C4, C2-C5, C3-C6)	No	Alpha-like Neurotoxin
*1KV0:A	Yes	(C1-C4, C2-C5, C3-C6)	No	Alpha-like Toxin
*1KV0:A	Yes	(C1-C4, C2-C5, C3-C6)	No	Alpha-like Toxin
*1LU0:A	Yes	(C1-C4, C2-C5, C3-C6)	Yes	Hydrolase Inhibitor
*1LU0:B	Yes	(C1-C4, C2-C5, C3-C6)	Yes	Hydrolase Inhibitor
*1NPI	Yes	(C1-C4, C2-C5, C3-C6)	No	Neurotoxin
1P9G	Yes	C1-C4, C2-C6,C3-C5	No	Antifungal Protein

*1PTX	Yes	(C1-C4, C2-C5, C3-C6)	No	Scorpion Toxin
1R0R	Yes	Array is not compact or absent	No	Serine Protease
*1SEG	Yes	(C1-C4, C2-C5, C3-C6)	No	Scorpion Alpha Toxin
1SGP	No	Array is not compact or absent	No	Serine Protease/ Inhibitor
*1SN4	Yes	(C1-C4, C2-C5, C3-C6)	No	Scorpion Neurotoxin
*1T7E	Yes	(C1-C4, C2-C5, C3-C6)	No	Alpha-Like Neurotoxin
*2ASC	Yes	(C1-C4, C2-C5, C3-C6)	No	Scorpion Toxin
2GKR	Yes	Array is not compact or absent	No	Hydrolase Inhibitor
*2SN3	Yes	(C1-C4, C2-C5, C3-C6)	No	Scorpion Neurotoxin
2UUY	Yes	C1-C3, C2-C6, C4-C5	No	Tryptase Inhibitor

1. * = true positives

Table S2.6: List and description of 23 positively predicted proteins in NewNMR751 set,deposited in PDB from July 04, 2012 to March 25, 2014.

PDB id	Functional classification	Disulfide connectivity in the putative motif	True positive	Predicted by Knotter1D
*2LIX	Potassium channel toxin	C1-C4, C2-C5, C3-C6	Yes	No
*2LJ7	Antimicrobial Peptide	C1-C4, C2-C5, C3-C6	Yes	No
*2LJS	Cyclotide	C1-C4, C2-C5, C3-C6	Yes	Yes
*2LL1	Spider toxin	C1-C4, C2-C5, C3-C6	Yes	Yes
*2LN4	Antimicrobial Peptide	C1-C4, C2-C5, C3-C6	Yes	No
*2LT8	Antimicrobial Peptide	C1-C4, C2-C5, C3-C6	Yes	No

*2LU9	Potassium channel toxin	C1-C4, C2-C5, C3-C6	Yes	No
*2LUR	Cyclotide	C1-C4, C2-C5, C3-C6	Yes	Yes
*2LY5	Defensin-like	C1-C4, C2-C5, C3-C6	Yes	No
*2LZX	New ICK toxin from sponge	C1-C4, C2-C5, C3-C6	Yes	No
*2M2Q	New ICK toxin from bitter melon	C1-C4, C2-C5, C3-C6	Yes	No
*2M2R	New ICK toxin from bitter melon	C1-C4, C2-C5, C3-C6	Yes	No
*2M36	Spider toxin	C1-C4, C2-C5, C3-C6	Yes	Yes
2M3H	Apoptotic protein	**Array is not compact or absent	No	No
*2M3J	New ICK toxin from sponge	C1-C4, C2-C5, C3-C6	Yes	No
*2M4Z	Spider toxin	C1-C4, C2-C5, C3-C6	Yes	Yes
*2M86	Cyclotide	C1-C4, C2-C5, C3-C6	Yes	Yes
*2M9O	Cyclotide	C1-C4, C2-C5, C3-C6	Yes	Yes
2MD7	Transcription	**Array is not compact or absent	No	No
*2MH1	Cyclotide	C1-C4, C2-C5, C3-C6	No	Yes
*4B2U	New ICK toxin *sicarius spiders*	C1-C4, C2-C5, C3-C6	Yes	No
*4B2V	New ICK toxin *sicarius spiders*	C1-C4, C2-C5, C3-C6	Yes	No
*4BMF	Hydrolase	C1-C4, C2-C5, C3-C6	Yes	No

[1]. * = true positives

Table S2.7: PDB ID of proteins detected by PSI BLAST with different E-values

PSI BLAST hits with E-value 0.01

>2LJS:A|PDBID	>2LU9:A|PDBID	>2M4Z:A|PDBID
>2LN4:A|PDBID	>2LUR:A|PDBID	>2M86:A|PDBID
>2LR3:A|PDBID	>2LY5:A|PDBID	>2M9O:A|PDBID
>2LT8:A|PDBID	>2M2D:A|PDBID	>2MH1:A|PDBID

PSIBLAST hits with E-value 0.1

>2LIX:A|PDBID	>2LT8:A|PDBID	>2M2Q:A|PDBID
>2LJ7:A|PDBID	>2LU9:A|PDBID	>2M3J:A|PDBID
>2LJS:A|PDBID	>2LUR:A|PDBID	>2M4Z:A|PDBID
>2LL1:A|PDBID	>2LY1:A|PDBID	>2M86:A|PDBID
>2LN4:A|PDBID	>2LY5:A|PDBID	>2M9O:A|PDBID
>2LR1:A|PDBID	>2LZX:A|PDBID	>2MBC:A|PDBID
>2LR3:A|PDBID	>2M2D:A|PDBID	
>2MH1:A|PDBID		

PSIBLAST hits with E-value 0.5

>2LE8:A|PDBID	>2LWW:A|PDBID	>2M6P:A|PDBID
>2LGC:A|PDBID	>2LXF:A|PDBID	>2M74:A|PDBID
>2LIX:A|PDBID	>2LY1:A|PDBID	>2M7P:A|PDBID
>2LIY:A|PDBID	>2LY5:A|PDBID	>2M86:A|PDBID
>2LJ7:A|PDBID	>2LZO:A|PDBID	>2M9O:A|PDBID
>2LJS:A|PDBID	>2LZX:A|PDBID	>2MAB:A|PDBID
>2LL1:A|PDBID	>2M16:A|PDBID	>2MAB:B|PDBID
>2LN4:A|PDBID	>2M1U:A|PDBID	>2MBC:A|PDBID
>2LNC:A|PDBID	>2M1X:A|PDBID	>2ME0:A|PDBID
>2LR1:A|PDBID	>2M2D:A|PDBID	>2MGX:A|PDBID
>2LR3:A|PDBID	>2M2Q:A|PDBID	>2MH1:A|PDBID
>2LSQ:A|PDBID	>2M3J:A|PDBID	>2MJV:B|PDBID
>2LT8:A|PDBID	>2M3V:A|PDBID	>2ML5:A|PDBID
>2LU9:A|PDBID	>2M48:A|PDBID	>3ZFJ:A|PDBID
>2LUL:A|PDBID	>2M4E:A|PDBID	>4B2R:A|PDBID
>2LUR:A|PDBID	>2M4I:A|PDBID	>4BF8:A|PDBID
>2LW3:A|PDBID	>2M4V:A|PDBID	
>2LWR:A|PDBID	>2M4Z:A|PDBID	

Table S2.8: PDB ids of 100 proteins from the "Eukaryote" subset analyzed manually. From 636 chains detected as STP by PredSTP in the Eukaryote dataset, 139 chains were obtained having maximum 30% sequence identity using CD-hit. Out of the 139 chains, the first 100 chains (based on PDB id) were manually analyzed for sequential tri-disulfide bonds using Jmol.

>1ACW:A|PDBID
>1ADX:A|PDBID
>1AG7:A|PDBID
>1AGG:A|PDBID
>1APQ:A|PDBID
>1ATA:A|PDBID
>1AXH:A|PDBID
>1AYJ:A|PDBID
>1B9G:A|PDBID
>1BBG:A|PDBID
>1BCG:A|PDBID
>1BGK:A|PDBID
>1BMR:A|PDBID
>1BRZ:A|PDBID
>1C4E:A|PDBID
>1C6W:A|PDBID
>1C9P:B|PDBID
>1CCV:A|PDBID
>1CE3:A|PDBID
>1CGI:I|PDBID
>1CHL:A|PDBID
>1CIX:A|PDBID
>1CLV:I|PDBID
>1CMR:A|PDBID
>1CN2:A|PDBID
>1CNN:A|PDBID
>1D1H:A|PDBID
>1DF6:A|PDBID
>1DKC:A|PDBID
>1DS3:I|PDBID
>1ERD:A|PDBID
>1FLE:I|PDBID
>1FU3:A|PDBID
>1G1P:A|PDBID
>1G9P:A|PDBID
>1GL0:I|PDBID
>1GPS:A|PDBID
>1H20:A|PDBID
>1H9H:I|PDBID
>1HD6:A|PDBID
>1HEV:A|PDBID
>1HI7:B|PDBID
>1HLY:A|PDBID
>1HY9:A|PDBID
>1HYK:A|PDBID
>1I26:A|PDBID
>1I2U:A|PDBID
>1IOX:A|PDBID
>1IXT:A|PDBID
>1JJZ:A|PDBID
>1JLZ:A|PDBID
>1JU8:A|PDBID
>1K36:A|PDBID
>1KCP:A|PDBID
>1KLI:L|PDBID
>1KOZ:A|PDBID
>1KQH:A|PDBID
>1L3Y:A|PDBID
>1LMR:A|PDBID
>1LU8:A|PDBID
>1LUP:A|PDBID
>1M2S:A|PDBID
>1MCT:I|PDBID
>1MM0:A|PDBID
>1MM2:A|PDBID
>1MR4:A|PDBID
>1MYN:A|PDBID
>1N89:A|PDBID
>1NE5:A|PDBID
>1NIY:A|PDBID
>1OMC:A|PDBID
>1P9G:A|PDBID
>1PJV:A|PDBID
>1PVZ:A|PDBID
>1Q2K:A|PDBID
>1Q3J:A|PDBID
>1QK7:A|PDBID
>1R1F:A|PDBID
>1RMK:A|PDBID
>1S8K:A|PDBID
>1UDK:A|PDBID
>1UGL:A|PDBID
>1UR6:B|PDBID
>1V91:A|PDBID
>1VIB:A|PDBID
>1WHE:A|PDBID
>1WMT:A|PDBID
>1WQB:A|PDBID
>1X5V:A|PDBID
>1Y29:A|PDBID
>1YP8:A|PDBID
>1YZ2:A|PDBID
>1ZA8:A|PDBID
>1ZAQ:A|PDBID
>1ZFU:A|PDBID
>1ZNT:A|PDBID
>2B68:A|PDBID
>2D56:A|PDBID
>2E2F:A|PDBID
>2E2S:A|PDBID

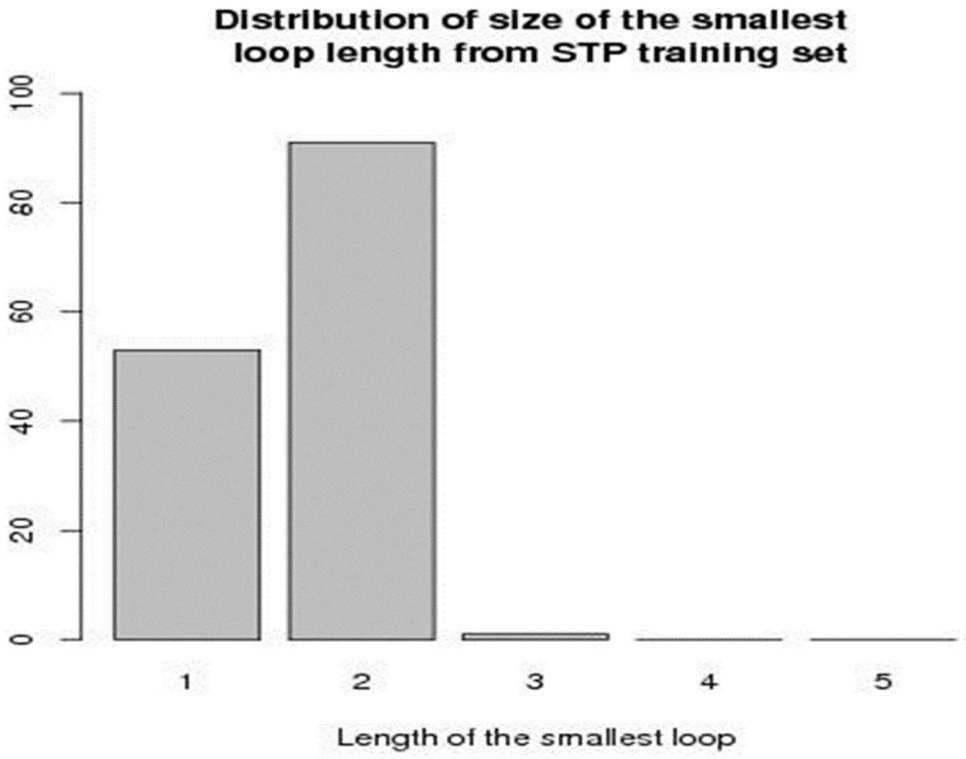

Figure S2.1: Distribution of size of the smallest loop lengths of control STP chains from the training set.

CHAPTER THREE

Protein Classification Using Modified *N-grams* and *Skip-grams*

This chapter is published as: Islam, SM Ashiqul, Benjamin J. Heil, Christopher Michel Kearney, and Erich J. Baker. "Protein classification using modified n-grams and skip-grams." Bioinformatics (2017): btx823.

Abstract

Classification by supervised machine learning greatly facilitates the annotation of protein characteristics from their primary sequence. However, the feature generation step in this process requires detailed knowledge of attributes used to classify the proteins. Lack of this knowledge risks the selection of irrelevant features, resulting in a faulty model. In this study, we introduce a supervised protein classification method with a novel means of automating the work-intensive feature generation step via a Natural Language Processing (NLP)-dependent model, using a modified combination of n-grams and skip-grams (m-NGSG). A meta-comparison of cross-validation accuracy with twelve training datasets from nine different published studies demonstrates a consistent increase in accuracy of m-NGSG when compared to contemporary classification and feature generation models. We expect this model to accelerate the classification of proteins from primary sequence data and increase the accessibility of protein characteristic prediction to a broader range of scientists. M-NGSG is freely available at Bitbucket: https://bitbucket.org/sm islam/mngsg/src A web server is available at watson.ecs.baylor.edu/ngsg

Introduction

It is well appreciated that primary polypeptide sequence informs higher order protein structure. The primary sequence provides the blueprint which encodes the purpose of the protein, ultimately determining the protein's characteristics, functions, subcellular localization and interactions (Pour-El 1978). However, classical approaches using primary sequence alignment for the prediction of remote homology detection are problematic due to low signal to noise ratios in polypeptide strings (Teichert et al. 2010). To circumvent this problem, non-alignment based methodologies are being investigated to demonstrate remote homology (Bonham-Carter, Steele, and Bastola 2014; Vinga and Almeida 2003; Du, Gu, and Jiao 2014; Liu et al. 2014). Here we illustrate a novel approach that relies on Natural Language Processing (NLP) to produce generalized feature sets for machine learning classification of protein characteristics.

A polypeptide string can be treated as a text string where hidden information is deciphered by implementing NLP techniques. Generating n-grams(Cavnar and John 1994) and skip- grams(Guthrie et al. 2006) from text documents is a feature extraction method which can produce meaningful information for machine learning (ML) classification algorithms(Cavnar and John 1994; Guthrie et al. 2006), and has been used for the categorization and sorting of documents based on their subject matter(Hu and Liu 2004; Pang, Lee, and Vaithyanathan 2002; Tan, Wang, and Lee 2002). Treating a primary protein sequence as a textual string is a natural extension of this approach. Indeed, text mining has been used previously for protein clustering and classification, protein-protein interaction (PPI), protein folding, and cnRNA identification(Zeng et al. 2015). Linguistic methodologies

based on primary sequence features have also been applied in areas of secondary structure prediction(Ding, Lin, et al. 2014).

Sequence classification using supervised and unsupervised machine learning methods is becoming popular due to algorithm accessibility in conjunction with increasing amounts of available biological data. Recent work in this area includes the classification of protein structure(Islam et al. 2015), localization(Yu et al. 2006), function(Cai et al. 2003), family(Chou 2005) and protein-protein interaction (PPI)(Zhao, Ma, and Yin 2012; Yu et al. 2006) (Zhao et al., 2012; Yu and Hwang, 2008) based on primary sequence. These studies consistently report that ML approaches are superior to alignment-based predictions when deriving protein characteristics from primary sequence and perform effectively in protein groups with low sequence similarity. However, the success of ML models depends heavily on training data, feature extraction, classifier algorithm selection and optimization.

Among these steps, robust results are disproportionately influenced by feature selection. Thus, substantial effort is required to obtain meaningful features from protein data. While universal methods for feature extraction are problematic due to the wide range of classification strategies, several generalized feature generation methods have been proposed. Many of these methods aim to address specific classification problems(Bock and Gough 2001; Islam et al. 2015; Du, Gu, and Jiao 2014), while others may be implemented as semi-automated feature generators. For example, amino acid composition(Verma and Melcher 2012) and pseudo-amino acid composition(Du, Gu, and Jiao 2014) based feature extraction schemes have been successfully used to solve a range of classification problems(Garg, Bhasin, and Raghava 2005; Qiu et al. 2016; Xu et al. 2013; Tiwari 2016). There are also hybrid feature generation strategies which include both generalized and data specific feature

selection methods(Ettayapuram Ramaprasad et al. 2015; Sharma et al. 2013; Chaudhary et al. 2016). In each case, however, manual intervention is required to produce the optimal set of features.

Using n-grams and skip-grams in biological applications driven by ML is not without precedent. For example, the n-gram model has been used to classify protein sequences into super families using extreme machine learning(Cao and Xiong 2014). Homology between proteins with low sequence similarity has also been successfully revealed using distances between Top-n-gram and amino acid residue pairs(Liu et al. 2014). “Spaced words” is a derivative of n-gram feature selection in biological sequence analysis where the letters of one or more indices in each word are replaced by blanks except the first and last letters. This method of feature extraction is used along with another method called kmacs to perform alignment-free comparison in both DNA and protein sequences(Horwege et al. 2014) .

Through the application of a modified NLP n-gram and skip-gram (m-NGSG) approach, we developed a novel supervised protein classification procedure with an automated feature generation model without the requirement of expert intervention for optimal feature selection. Here, we used modified (optimized for protein sequence) n-grams and skip-grams to extract features in a protein-family agnostic fashion which is integrated with a logistic regression classifier. Further, we have performed a meta-comparison between our generalized classification model with several other published specialized protein classification models using the corresponding benchmark datasets and cross-validation methods to validate our new model.

Materials and Methods

Feature Generation, Vectorization and Model Construction

The n-gram and k-skip-bi-gram profiles are initially extracted from each candidate protein sequence. They are given a position identity with respect to the C-terminus of the protein sequence. Thereafter, modifications of the length of k-skip-bi-grams and positional identity are performed to obtain potential motifs (or words). The whole procedure is described in the following subsections.

Binary Profile of N-Grams in a Protein Sequence

N-grams, strings of contiguous sequences consisting of n items, are valuable features extracted from text or speech, and are useful in NLP and sentiment analysis(Cui, Mittal, and Datar, n.d.; Socher et al. 2013; Ghiassi, Skinner, and Zimbra 2013). Given that a primary protein sequence can be treated as a string of amino acids, n-gram-based feature extraction methods can be applied to predict functionality from a sequence. Interestingly, n-grams from a protein sequence also offer biologically meaningful information, as each n-gram represents a protein sequence motif. N-gram motifs provide information helpful in inferring protein functionality, and can be represented as:

$$GM_p^s \qquad equation\ (1)$$

where GM stands for Gram Motif, and s is a positive integer not longer than the length, L, of the corresponding protein sequence (s|s∈N,s<=L). s = 0 represents a null motif, s = 1 represents all single residue motifs (uni-grams), s = 2 represents all dipeptide motifs (bi-grams excluding their uni-gram components) and s = n represents all n-peptide motifs. p is the permutation index of the participating residue(s) parameterized by s. Since there are 20

different amino acids, there can be 20^s different values of p for an s-gram. For example, if we consider the amino acid sequence MISHW, then M is one of the 20 possible elements of uni-gram (s = 1) as $p=20^s=20^1=20$. Similarly, MI is one of the 400 possible elements of dipeptides (s = 2) as $p=20^s=20^2=400$.

Binary Profile of K-Skip-Bi-Grams

Skip-grams are a technique largely used in the field of speech processing that allow items, or in our case substrings, to be ignored during processing(Guthrie et al. 2006). In m-NGSG we adopted the k-skip-bi-gram approach where the skip distance, k, allows a total of k or fewer skips to construct the bi-gram.

For example, for protein sequence MISHW, the 2-skip-bi-grams will be MI, IS, SH, HW, MXS, IXH, SXW, MXXH and IXXW where skips are represented by X. The k = 0 skips are MI, IS, SH and HW, the k = 1 skips are MXS, IXH, SXW and the k = 2 skips are MXXH, IXXW. This approach can be useful in comparing k-length mutational events across protein sequences. In order to avoid duplicating features extracted with the n-gram method, we exclude the motifs produced where k = 0.

$$SM_p^b \qquad\qquad equation\ (2)$$

SM stands for Skip Motif and b is the number of skips between two amino acids. b is a positive integer that is at most two less than the length of the protein sequence ($b|b\in N, b\leq L-2$). b = 0 represents no skips between a specific permutation of two residues, b = 1 represents one skip, and b = 2 represents two skips. p is the permutation index of the participating residue(s) parameterized by s. Since there are 20 different amino acids, there can be 20^2 different values of p for a given value of b.

Modification of Skips in K-Skip-Bi-Gram Motifs

The m-NGSG employs a modification of the k-skip-bi-gram model that allows buffering on the number of skips. That is, after obtaining the exact number of skips from a k-skip-bi-gram, an estimated number of skips is determined as:

$$SM_p^c \qquad equation\ (3)$$

where c represents the estimated number of skips based on the given parameter a, and b is the number of skips in a motif as determined from the k-skip-bi-gram.

$$c = b + ((a - b)\%a) \qquad \text{equation (4)}$$

An expanded example is described in the Supplementary text.

Modification of Estimated C-Terminus Position in N-Grams and K-Skip-Bi-Grams

During feature extraction from a protein sequence m-NGSG determines the relative position of the motifs with respect to the C-terminus. N-gram or k-skip-bi-gram motifs are tagged with a maximum position identity, noted as s^{th} gram (for an n-gram) and for a b^{th}-skip-bi-gram(k-skip-bi-gram), respectively. This position is measured after obtaining the exact distance from the C-terminus and applying a buffering distance to capture shared positional identity for n-gram motifs,

$$GM_p^s, (x; y) \qquad equation\ (5)$$

and k-skip-bi-gram motifs,

$$SM_p^c, (x; y) \qquad equation\ (6)$$

$$x=z+((y-z)\%y) \qquad equation\ (7)$$

where x represents the distance identity of motif GM_p^s or SM_p^c based on the given parameter y, and z is the distance of the onset of the corresponding motif from the C-terminus of the sequence buffered by y. m-NGSG initializes y based on y0, defined by ModifiedGridSearch, and increases with the length, l, of the motif, as:

$$y = y0 + l - 1 \qquad \text{equation (8)}$$

An expanded example is described in the Supplementary text.

Finally, the motifs are vectorized to construct feature vectors with a simultaneous noise filtration. The length of initial n-gram and k-skip-bi-gram motifs, and amplitude of their modification are determined by six parameters (described in Supplementary Table S1). The parameters are optimized using a modified grid search algorithm (see Algorithm 3.1 and 3.2 in Supplementary text) depending on the training set of a five-fold cross-validation using a logistic regression classifier. As the modified grid search is seeded using different initial n-grams, they are defined as seeds in this study (see the Methods section in Supplementary Text for details). To observe the scalability of the optimization algorithm, a rum-time study was also performed on different size of datasets (see Supplementary Method).

Meta-Comparison

The performance of m-NGSG was compared with other methodologies that use generalized or data-specific feature extraction methods for model construction. Comparison models were chosen based on the availability of benchmark data reported by those models, the diversity of protein characteristics classified, and the ability of the model to report functional or structural classification of proteins with regard to their sequence. The performance was compared with the published models using logistic regression

(Supplementary Table S3.6). The number and size of different classes in each dataset are described in Supplementary Table S3.4.

In addition, we have conducted a performance comparison of n-gram and skip-gram-based feature generation models with and without modifiers. We also demonstrate a comparison among different feature generation methods and classifier combinations followed by a comparison among the models with induced noise on the Subchlo60 dataset to test the robustness of the m-NGSG model (see Supplementary Method and Supplementary Figures S3.5, S3.6 and S3.7).

Results

Parameter Optimization Analysis

This study illustrates that n-gram and skip-gram text mining approaches can be exploited to develop a generalized feature extraction method for protein classification. N-gram and skip-gram models are not used directly; rather, the models are modified according to six parameters based on sequence (Supplementary Table S3.1). The parameters themselves are optimized by using the modified grid search-based algorithm (see Algorithm 3.2 in Supplementary text) and compared to 12 benchmark datasets. In each case, the automated generalized feature extraction algorithm obtained features that outperformed the originally published feature sets for linear regression.

For the benchmark datasets iAMP-2L, Cypred, TumorHPD 1, TumorHPD 2, IGPred and PVPred, the optimization strategy for m-NGSG reported the same parameters (see Supplementary Figure. S3.1) with identical accuracy (Supplementary Figure. S3.2) regardless of the initial seed, indicating convergence in these datasets. For the subchlo raw

training set, parameters n, k and y showed variation with some seeds (see Supplementary Figure. S3.1). Overall, the subchlo raw training set accuracy for different seeds ranged from 89 to 89.70% (Supplementary Figure. S3.2). For the subchloro60 training set, parameters n, k, y and c demonstrated variability over the first four seeds and then became stable while the accuracy ranged from 65.76 to 68.07%. In the PredSTP training set, there was slight variation in parameters n, k and y which was also reflected in the variation of accuracies for the corresponding seeds. Parameters for the HemoPI 1 training set varied for seed three, and training set HemoPI 2, which classifies between hemolytic and semihemolytic peptides, presented variation in parameters n, k, kp, y and c for seed 3, 4 and 5 (see Supplementary Figures S3.1 and S3.2).

The goal of parameter optimization is to identify parameters that contribute to the best accuracy after five-fold cross-validation. Although the principle approach is a modified grid-search, it demonstrates an ability to converge on accuracy regardless of initiating seeds. Supplementary Figure S3.3 illustrates the convergence characteristic of the optimization algorithm which calculates the mode value of accuracies generated from different seeds against the percent change of the accuracies from each seed for a specific training set when compared to the mode accuracy. Flat areas in Supplementary Figure S3.3 indicate low percentage change compared to the mode which suggests convergence.

To observe the influence of modifying parameters of n-grams and skip-grams, we compared the performance of m-NGSG with and without modifying n-grams and skip-grams and observed the change of accuracy without and with modifications. Supplementary Figure S3.5 illustrates an increase of accuracy with modifications for most of the datasets with an average 2.2%. This result indicates that modifications are not necessary for all datasets;

however, the classification performance on some dataset noticeably improved by the modifications (see Supplementary Figure. S3.5 for an elaborated description). This study explains why the modification of n-grams and skip-grams is necessary to generalize the usability of m-NGSG.

Meta-Comparison of Prediction Performance on Benchmark Datasets

Once the parameters were optimized for each benchmark training set, the reported accuracy was compared to the m-NGSG model built with the optimized feature set. A logistic regression classifier was used for all models. To compare the cross-validation accuracy, we mimicked the approach published as part of the original dataset, either five-fold, ten-fold or jackknife validation.

Subchlo

Subchlo is a multi-class classifier designed to predict the localization of chloroplast proteins. Subchlo raw is a dataset of protein sequences based on their location in chloroplast and the Subchloro60 dataset consists of proteins with approximately 60% sequence identity(Du, Cao, and Li 2009) (Du et al., 2009). Subchlo raw and Subchlo60 were both cross-validated by a jackknife method in the original publication, resulting in a combined accuracy of 89.69 and 67.18%, respectively. The accuracy of the m-NGSG model is 91.59 and 73.92% for the same datasets (see Supplementary Table S3.2). This indicates a 2.12 and 10.03% increase of accuracy by our model compared to the reported model for the two given datasets (Figure. 3.1A).

To check the suitability of other classifiers for use with the automated feature selection method in m-NGSG, we compared the accuracy of linear regression to SVM with

linear and non-linear (rbf) kernels and K-nearest neighbors (KNN). We used features generated by m-NGSG with logistic regression and SVM linear kernel models, which yielded accuracies of 73.92 and 75.09%, respectively (see Supplementary Figure. S3.6). Also, Supplementary Figure S3.6 illustrates that the m-NGSG-SVM and m-NGSG-LR (logistic regression) combinations generate higher accuracy among other QSAR based features (Simeon et al., 2016) and classifier combinations. However, we integrated logistic regression with m-NGSG for other studies as we used that for the m-NGSG parameter optimization step for and does not need a further optimization of the parameters of the classifier. We also performed a performance comparison of feature-classifier combination with different k-fold cross-validations on the Subchlo60 with 8% noised data-points. There also m-NGSG-LR showed the highest accuracy among all other feature-classifier combinations for all forms of cross-validation (see Supplementary Figure. S3.7).

OsFP

The osFP model classifies fluorescent proteins into monomer or oligomeric states(Simeon et al. 2016). In the original study, different QSAR-based feature selection models were investigated. The best model yielded an average of 72.13 and 72.89% accuracy for the training and test sets after 100 iterations (see Supplementary Table S3.5). In contrast, m-NGSG generated an average of 78.02 and 79.21% accuracy for the same sets, yielding an 8.16 and 8.6% increase of accuracy respectively (Figure. 3.1A and B). To confirm the superiority of m-NGSG model over the QSAR based feature selection method, we also performed a comparison on Subchlo60 dataset. The comparison demonstrated that m-NGSG's performance is better than that of other feature generation methods (see Supplementary Figures S3.6 and S3.7).

iAMP-2L

The iAMP-2L(Xiao et al. 2013) model classifies antimicrobial peptides from nonantimicrobial peptides. Supplementary Tables S3.2 and S3.3 illustrate the increased performance of m-NGSG over the iAMP-2L when using the jackknife cross-validation method. The accuracy of m-NGSG on the training set was 91.25%, yielding a 5.71% rise over the previously reported accuracy. When we used m-NGSG to evaluate the performance on the benchmark independent test set, we achieved a 4.6% rise from the accuracy reported by the original model (Figure. 3.1A and B).

Cypred and PredSTP

Both Cypred (Kedarisetti et al. 2014) and PredSTP (Islam et al. 2015) classify proteins based on their structural characteristics. While Cypred performed comparably to m-NGSG (99.20% accuracy after 10-fold cross-validation in the original publication versus 99.53% for m-NGSG), m-NGSG did provide a modest 0.35% increase. On a bechmark out of sample test dataset, the m-NGSG model narrowly outperformed Cypred by 0.28%. On the other hand, a comparison on training set cross-validation accuracy between PredSTP and m-NGSG produces a 2.50% gain of accuracy from the original model (see Supplementary Tables S3.2 and S3.3 and Figure. 3.1*).*

TumorHPD 1 and 2

TumorHPD classifies tumor homing peptides to identify analogs of tumor homing ability(Sharma et al. 2013) (Sharma et al., 2013). Two training sets were used to generate the models: raw tumor homing peptides, TumorHPD 1 and tumor homing peptides less than or equal to ten residues long TumorHPD 2. Among three different generation methods they

used(Sharma et al. 2013), amino acid composition yielded the best accuracy 82.52 and 80.28% for the training set TumorHPD 1 and 2, respectively. The accuracy of m-NGSG the same datasets were 83.40 and 82.55%, respectively (see Supplementary Table S3.2) which using logistic regression yielding a 1.07 and 2.83% rise from the original model (Figure. 3.1A).

HemoPI 1 and 2

HemoPI 1 model classifies hemolytic and nonhemolytic proteins, while HemoPI 2 classifies hemolytic and semi hemolytic peptides(Chaudhary et al. 2016). The performance data for the training and test sets were available for the models developed from hybrid feature sets. The original model searched for the best accuracy by considering whole proteins and fractions of the proteins. Here, we compared the m-NGSG accuracy with only the whole length proteins. Our model generated 97.97% accuracy for HemoPI 1 and 79.5% accuracy (Supplementary Table S3.2) for HemoPI 2 training sets offering a 2.8 and 1.92% increase from the original models respectively. When we compared m-NGSG on the benchmark independent test sets, it achieved an increase of 3.26 and 0.7% for HemoPI 1 and HemoPI 2 respectively (Figure. 3.1).

IGPred and PVPred

IGPred(Tang, Chen, and Lin 2016) predicts immunoglobulin proteins, and PVPred(Ding, Feng, et al. 2014) predicts virion proteins from primary sequence data. The size of these proteins is very different from that of previously classified proteins. Immunoglobulin and virion proteins have very long sequences. In both models, important features were selected using ANOVA analysis before performing the jackknife cross-

validation. Therefore, we also performed jackknife cross-validation with and without an ANOVA-based feature selection method where we used the minimum number of features offering the best cross-validation accuracy (see Supplementary Figure. S3.4). The accuracy of m-NGSG model was 100% with ANOVA-based feature selection, and 92.60% with jackknife cross-validation (Supplementary Table S3.2), while the accuracy of the original IGPred model with jackknife test was 96.60%. The accuracy for the independent test set was 100% regardless the model (Supplementary Table S3.3). For PVPred, the accuracy of jackknife cross-validation with and without feature selection was 89.25 and 77.19% respectively, with corresponding accuracies of 90 and 93.33% on the benchmark independent test sets. The original feature selection assisted model showed 85.02% accuracy for jackknife cross-validation and 86.66% accuracy for the independent test set (Supplementary Table S3.3).

Discussion

The crucial steps of machine learning-based classifications are the selection of datasets that unambiguously represent informative classes, creation of meaningful features from the dataset that can optimally correlate to different classes, and an appropriate choice of machine learning algorithms which effectively classify the data based on the data points and descriptors. Predicting protein characteristics from primary sequence is becoming popular as appropriate data sources experience rapid growth and computer libraries for machine learning algorithms become accessible to bench biologists. However, generating effective features from protein sequences continues to require enormous manual intervention, and automated approaches have narrowly scoped structure prediction. Chemical property-based feature generation algorithms and dipeptide or tripeptide motif-specific

approaches(Chaudhary et al. 2016; Kedarisetti et al. 2014) account for the majority of these feature generation methods. In particular, Pseudo Amino Acid Composition (PseAAC) has been the most frequently used approach to classify proteins per their functional properties(Mohabatkar et al. 2013; Xiao et al. 2013), subfamilies(Chou 2005), interactions with other proteins(Jia et al. 2015) and subcellular localizations(Lin et al. 2008). Methods that classify based on physicochemical or biochemical properties rely heavily on the AAindex database(Kawashima 2000).

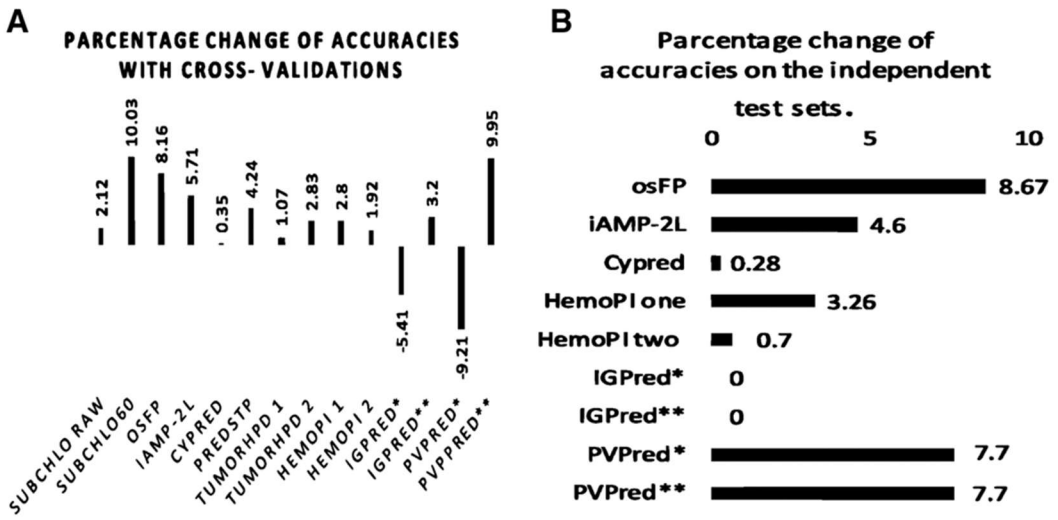

Figure 3.1: The percentage changes of accuracies m-NGSG in cross-validation compared to the original models for each dataset. IGPred* and PVPred* shows the comparative accuracy changes without feature selection while IGPred** and PVPred** shows accuracy changes after mimicking the feature selection method of the original model (A). The percentage changes of accuracies m-NGSG on the independent test sets (depending on availability) compared to the original models. IGPred* and PVPred* shows the comparative accuracy changes without feature selection while IGPred** and PVPred** shows accuracy changes after mimicking the feature selection method of the original model (B)

However, as protein sequences are strings of amino acid residues, they can be treated as normal text that can be interpreted through NLP-based techniques. The m-NGSG algorithm presented herein generates features in a text mining manner where words are artificially generated from protein sequences using modified n-gram and skip-gram models.

The models themselves are optimized based on the combination of six parameters (Supplementary Table S3.1). NLP processing of protein strings creates a corpus of words that is subsequently used for vectorization to generate features for each individual data point. To fully automate the classification process, a modified grid search algorithm is employed to obtain the optimal values of the six parameters. The parameter optimization itself is performed after 5-fold cross-validation to confirm the whole training set is not exposed to the classifier during the optimization step, limiting the risk of bias during the meta-comparison. Moreover, all the optimization was done with a logistic regression classifier with the same regularization parameter value to avoid disparity in this step.

Interestingly, although the optimization algorithm primarily depends on a modified grid search, in most cases parameters converge to a single value regardless of the initial seed (Supplementary Figure. S3.1). Also, in many cases, the different starting seeds yield the same accuracy (Supplementary Figure. S3.2). These outcomes indicate that the optimization algorithm searches for the maximum value while retaining the ability to converge.

A collection of contemporary models was chosen for meta-comparison based on their diversity of classification topic (such as functional, structural and subcellular localization), database size, sequence length and feature selection methods (Supplementary Table S3.6). Benchmark training datasets from comparison model publications were used (Supplementary datasets). With the exclusion of the osFP dataset, the meta-analysis comprised six of the eleven independent test sets (five were unavailable). In the case of osFP, the original dataset was divided into training and test sets and ten-fold cross-validation was performed only on the training set. For the models without an independent test set, evaluation with cross-

validations on the benchmark datasets was performed as an adequate replacement to reveal the comparative performance between the models.

The m-NGSG model outperformed the cross-validation accuracy of each model it was compared against, with the increase in accuracy ranging from 0.35 to 9.95% over the original models (Figure. 3.1A). Moreover, we observed up to an 8.67% increase in accuracy over the original model when compared to independent test sets (Figure. 3.1B). As shown in Figure 3.1A, the cross-validation accuracy of IGPred and PVPred without feature selection was considerably less than the original model where ANOVA based feature selection was performed before the execution of jackknife cross-validation accuracy, while the same ANOVA based feature selection method in m-NGSG model displayed higher jackknife cross-validation accuracy on the same training set. The accuracy on the independent test set demonstrated a 0 and 7.7% increment from the original IGPred and PVPred, respectively, regardless of which feature selection was used (Figure. 3.1B). This result illustrates that using a feature selection method followed by cross-validation test biases the cross-validation process without improving the performance of a model.

The Subchlo60 and osFP datasets were used to compare the performance of the m-NGSG model with motif composition, represented by AAC/DPC/TPC, and chemical property-based feature generation methods, represented by AC, CTD, Ctriad, SOCN, QSO and PseAAC methods (Supplementary Table S3.5, Figures S3.6 and S3.7). The m-NGSG model demonstrates a 2.12% increase over the PseAAC-based model on the Subchlo raw dataset. However, with the low sequence identity Subchlo60 dataset we observed a 10.03% increase in accuracy (Figure. 3.1A). This result indicates that m-NGSG performs comparatively better than chemical property-based method when the sequence identity in the

training dataset is lower. In addition, the accuracy of m-NGSG outperformed all of the competitors in the osFP model (Supplementary Table S3.5), illustrating the robustness of the m-NGSG model for feature generation when compared to presently available approaches.

Although there are a number of automated models related to protein sequence classification such as spectrum kernels(C. Leslie, Eskin, and Noble 2001), their mismatch variant(C. S. Leslie et al. 2004) (Leslie et al., 2004) and vector quantization techniques(Clark and Radivojac 2013), these kernel-based approaches are associated most often with feature transformation than novel feature generation, such as that proposed by the m-NGSG framework. Moreover, as the kernel-based studies do not offer any distinct benchmark dataset or any module/library for their algorithm, we were not able compare our automated module with those. Another study(Asgari and Mofrad 2015) demonstrated the use of the n-grams with a skip-gram model where the skip-gram model relies on the probability of word associations and not skips (https://arxiv.org/abs/1402.3722) as it is applied here.

Conclusion

The meta-comparison results outlined in this study illustrate that the m-NGSG is an effective fully automated feature generation method. This framework will benefit the machine learning-based protein classification community, particularity those interested in classification based on primary protein sequence. It is expected that m-NGSG will significantly reduce the work load for the feature generation step regardless of protein characteristics and sequence size. Moreover, by analyzing the feature importance, the distinguishing part of the sequence (motif) in a protein class can be revealed, which is often difficult to discover using multiple sequence alignment.

References

Asgari, Ehsaneddin, and Mohammad R. K. Mofrad. 2015. "Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics." *PloS One* 10 (11): e0141287. https://doi.org/10.1371/journal.pone.0141287.

Bock, J. R., and D. A. Gough. 2001. "Predicting Protein--Protein Interactions from Primary Structure." *Bioinformatics (Oxford, England)* 17 (5): 455–60.

Bonham-Carter, Oliver, Joe Steele, and Dhundy Bastola. 2014. "Alignment-Free Genetic Sequence Comparisons: A Review of Recent Approaches by Word Analysis." *Briefings in Bioinformatics* 15 (6): 890–905. https://doi.org/10.1093/bib/bbt052.

Cai, C. Z., L. Y. Han, Z. L. Ji, X. Chen, and Y. Z. Chen. 2003. "SVM-Prot: Web-Based Support Vector Machine Software for Functional Classification of a Protein from Its Primary Sequence." *Nucleic Acids Research* 31 (13): 3692–97.

Cao, Jiuwen, and Lianglin Xiong. 2014. "Protein Sequence Classification with Improved Extreme Learning Machine Algorithms." *BioMed Research International* 2014: 103054. https://doi.org/10.1155/2014/103054.

Cavnar, William B., and M. Trenkle John. 1994. "N-Gram-Based Text Categorization." *Ann Arbor Mi* 48113 (2): 161–75.

Chaudhary, Kumardeep, Ritesh Kumar, Sandeep Singh, Abhishek Tuknait, Ankur Gautam, Deepika Mathur, Priya Anand, Grish C. Varshney, and Gajendra P. S. Raghava. 2016. "A Web Server and Mobile App for Computing Hemolytic Potency of Peptides." *Scientific Reports* 6 (March): 22843. https://doi.org/10.1038/srep22843.

Chou, Kuo-Chen. 2005. "Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes." *Bioinformatics (Oxford, England)* 21 (1): 10–19. https://doi.org/10.1093/bioinformatics/bth466.

Clark, Wyatt T., and Predrag Radivojac. 2013. "VECTOR QUANTIZATION KERNELS FOR THE CLASSIFICATION OF PROTEIN SEQUENCES AND STRUCTURES." In , 316–27. WORLD SCIENTIFIC. https://doi.org/10.1142/9789814583220_0031.

Cui, Hang, Vibhu Mittal, and Mayur Datar. n.d. "Comparative Experiments on Sentiment Classification for Online Product Reviews." In .

Ding, Hui, Peng-Mian Feng, Wei Chen, and Hao Lin. 2014. "Identification of Bacteriophage Virion Proteins by the ANOVA Feature Selection and Analysis." *Mol. BioSyst.* 10 (8): 2229–35. https://doi.org/10.1039/C4MB00316K.

Ding, Hui, Hao Lin, Wei Chen, Zi-Qiang Li, Feng-Biao Guo, Jian Huang, and Nini Rao. 2014. “Prediction of Protein Structural Classes Based on Feature Selection Technique.” *Interdisciplinary Sciences: Computational Life Sciences* 6 (3): 235–40. https://doi.org/10.1007/s12539-013-0205-6.

Du, Pufeng, Shengjiao Cao, and Yanda Li. 2009. “SubChlo: Predicting Protein Subchloroplast Locations with Pseudo-Amino Acid Composition and the Evidence-Theoretic K-Nearest Neighbor (ET-KNN) Algorithm.” *Journal of Theoretical Biology* 261 (2): 330–35. https://doi.org/10.1016/j.jtbi.2009.08.004.

Du, Pufeng, Shuwang Gu, and Yasen Jiao. 2014. “PseAAC-General: Fast Building Various Modes of General Form of Chou’s Pseudo-Amino Acid Composition for Large-Scale Protein Datasets.” *International Journal of Molecular Sciences* 15 (3): 3495–3506. https://doi.org/10.3390/ijms15033495.

Ettayapuram Ramaprasad, Azhagiya Singam, Sandeep Singh, Raghava Gajendra P S, and Subramanian Venkatesan. 2015. “AntiAngioPred: A Server for Prediction of Anti-Angiogenic Peptides.” *PloS One* 10 (9): e0136990. https://doi.org/10.1371/journal.pone.0136990.

Garg, Aarti, Manoj Bhasin, and Gajendra P. S. Raghava. 2005. “Support Vector Machine-Based Method for Subcellular Localization of Human Proteins Using Amino Acid Compositions, Their Order, and Similarity Search.” *The Journal of Biological Chemistry* 280 (15): 14427–32. https://doi.org/10.1074/jbc.M411789200.

Ghiassi, M., J. Skinner, and D. Zimbra. 2013. “Twitter Brand Sentiment Analysis: A Hybrid System Using n-Gram Analysis and Dynamic Artificial Neural Network.” *Expert Systems with Applications* 40 (16): 6266–82. https://doi.org/10.1016/j.eswa.2013.05.057.

Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. 2006. “A Closer Look at Skip-Gram Modelling.” In , 1–4. sn.

Horwege, Sebastian, Sebastian Lindner, Marcus Boden, Klas Hatje, Martin Kollmar, Chris-André Leimeister, and Burkhard Morgenstern. 2014. “Spaced Words and Kmacs: Fast Alignment-Free Sequence Comparison Based on Inexact Word Matches.” *Nucleic Acids Research* 42 (Web Server issue): W7-11. https://doi.org/10.1093/nar/gku398.

Hu, Minqing, and Bing Liu. 2004. “Mining and Summarizing Customer Reviews.” In , 168. ACM Press. https://doi.org/10.1145/1014052.1014073.

Islam, S. M. Ashiqul, Tanvir Sajed, Christopher Michel Kearney, and Erich J. Baker. 2015. “PredSTP: A Highly Accurate SVM Based Model to Predict Sequential Cystine Stabilized Peptides.” *BMC Bioinformatics* 16 (July): 210. https://doi.org/10.1186/s12859-015-0633-x.

Jia, Jianhua, Zi Liu, Xuan Xiao, Bingxiang Liu, and Kuo-Chen Chou. 2015. “IPPI-Esml: An Ensemble Classifier for Identifying the Interactions of Proteins by Incorporating Their Physicochemical Properties and Wavelet Transforms into PseAAC.” *Journal of Theoretical Biology* 377 (July): 47–56. https://doi.org/10.1016/j.jtbi.2015.04.011.

Kawashima, S. 2000. “AAindex: Amino Acid Index Database.” *Nucleic Acids Research* 28 (1): 374–374. https://doi.org/10.1093/nar/28.1.374.

Kedarisetti, Pradyumna, Marcin J. Mizianty, Quentin Kaas, David J. Craik, and Lukasz Kurgan. 2014. “Prediction and Characterization of Cyclic Proteins from Sequences in Three Domains of Life.” *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1844 (1): 181–90. https://doi.org/10.1016/j.bbapap.2013.05.002.

Leslie, Christina, Eleazar Eskin, and William Stafford Noble. 2001. “THE SPECTRUM KERNEL: A STRING KERNEL FOR SVM PROTEIN CLASSIFICATION.” In , 564–75. WORLD SCIENTIFIC. https://doi.org/10.1142/9789812799623_0053.

Leslie, Christina S., Eleazar Eskin, Adiel Cohen, Jason Weston, and William Stafford Noble. 2004. “Mismatch String Kernels for Discriminative Protein Classification.” *Bioinformatics (Oxford, England)* 20 (4): 467–76. https://doi.org/10.1093/bioinformatics/btg431.

Lin, Hao, Hui Ding, Feng-Biao Guo, An-Ying Zhang, and Jian Huang. 2008. “Predicting Subcellular Localization of Mycobacterial Proteins by Using Chous Pseudo Amino Acid Composition.” *Protein & Peptide Letters* 15 (7): 739–44. https://doi.org/10.2174/092986608785133681.

Liu, Bin, Jinghao Xu, Quan Zou, Ruifeng Xu, Xiaolong Wang, and Qingcai Chen. 2014. “Using Distances between Top-n-Gram and Residue Pairs for Protein Remote Homology Detection.” *BMC Bioinformatics* 15 (Suppl 2): S3. https://doi.org/10.1186/1471-2105-15-S2-S3.

Mohabatkar, Hassan, Majid Mohammad Beigi, Kolsoum Abdolahi, and Sasan Mohsenzadeh. 2013. “Prediction of Allergenic Proteins by Means of the Concept of Chou’s Pseudo Amino Acid Composition and a Machine Learning Approach.” *Medicinal Chemistry (Shariqah (United Arab Emirates))* 9 (1): 133–37.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. “Thumbs up?: Sentiment Classification Using Machine Learning Techniques.” In , 10:79–86. Association for Computational Linguistics. https://doi.org/10.3115/1118693.1118704.

Pour-El, A. 1978. “Functionality and Protein Structure; Based on a Symposium.” In .

Qiu, Wang-Ren, Bi-Qian Sun, Xuan Xiao, Zhao-Chun Xu, and Kuo-Chen Chou. 2016. “IHyd-PseCp: Identify Hydroxyproline and Hydroxylysine in Proteins by Incorporating Sequence-Coupled Effects into General PseAAC.” *Oncotarget* 7 (28): 44310–21. https://doi.org/10.18632/oncotarget.10027.

Sharma, Arun, Pallavi Kapoor, Ankur Gautam, Kumardeep Chaudhary, Rahul Kumar, Jagat Singh Chauhan, Atul Tyagi, and Gajendra P. S. Raghava. 2013. "Computational Approach for Designing Tumor Homing Peptides." *Scientific Reports* 3: 1607. https://doi.org/10.1038/srep01607.

Simeon, Saw, Watshara Shoombuatong, Nuttapat Anuwongcharoen, Likit Preeyanon, Virapong Prachayasittikul, Jarl E. S. Wikberg, and Chanin Nantasenamat. 2016. "OsFP: A Web Server for Predicting the Oligomeric States of Fluorescent Proteins." *Journal of Cheminformatics* 8 (1). https://doi.org/10.1186/s13321-016-0185-8.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. "Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank." In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–42.

Tan, Chade-Meng, Yuan-Fang Wang, and Chan-Do Lee. 2002. "The Use of Bigrams to Enhance Text Categorization." *Information Processing & Management* 38 (4): 529–46. https://doi.org/10.1016/S0306-4573(01)00045-0.

Tang, Hua, Wei Chen, and Hao Lin. 2016. "Identification of Immunoglobulins Using Chou's Pseudo Amino Acid Composition with Feature Selection Technique." *Molecular BioSystems* 12 (4): 1269–75. https://doi.org/10.1039/C5MB00883B.

Teichert, Florian, Jonas Minning, Ugo Bastolla, and Markus Porto. 2010. "High Quality Protein Sequence Alignment by Combining Structural Profile Prediction and Profile Alignment Using SABER-TOOTH." *BMC Bioinformatics* 11 (May): 251. https://doi.org/10.1186/1471-2105-11-251.

Tiwari, Arvind Kumar. 2016. "Prediction of G-Protein Coupled Receptors and Their Subfamilies by Incorporating Various Sequence Features into Chou's General PseAAC." *Computer Methods and Programs in Biomedicine* 134 (October): 197–213. https://doi.org/10.1016/j.cmpb.2016.07.004.

Verma, Ruchi, and Ulrich Melcher. 2012. "A Support Vector Machine Based Method to Distinguish Proteobacterial Proteins from Eukaryotic Plant Proteins." *BMC Bioinformatics* 13 Suppl 15: S9. https://doi.org/10.1186/1471-2105-13-S15-S9.

Vinga, Susana, and Jonas Almeida. 2003. "Alignment-Free Sequence Comparison-a Review." *Bioinformatics (Oxford, England)* 19 (4): 513–23.

Xiao, Xuan, Pu Wang, Wei-Zhong Lin, Jian-Hua Jia, and Kuo-Chen Chou. 2013. "IAMP-2L: A Two-Level Multi-Label Classifier for Identifying Antimicrobial Peptides and Their Functional Types." *Analytical Biochemistry* 436 (2): 168–77. https://doi.org/10.1016/j.ab.2013.01.019.

Xu, Yan, Jun Ding, Ling-Yun Wu, and Kuo-Chen Chou. 2013. "ISNO-PseAAC: Predict Cysteine S-Nitrosylation Sites in Proteins by Incorporating Position Specific Amino Acid Propensity into Pseudo Amino Acid Composition." *PloS One* 8 (2): e55844. https://doi.org/10.1371/journal.pone.0055844.

Yu, Chin-Sheng, Yu-Ching Chen, Chih-Hao Lu, and Jenn-Kang Hwang. 2006. "Prediction of Protein Subcellular Localization." *Proteins: Structure, Function, and Bioinformatics* 64 (3): 643–51. https://doi.org/10.1002/prot.21018.

Zeng, Zhiqiang, Hua Shi, Yun Wu, and Zhiling Hong. 2015. "Survey of Natural Language Processing Techniques in Bioinformatics." *Computational and Mathematical Methods in Medicine* 2015: 1–10. https://doi.org/10.1155/2015/674296.

Zhao, Xiao-Wei, Zhi-Qiang Ma, and Ming-Hao Yin. 2012. "Predicting Protein-Protein Interactions by Combing Various Sequence- Derived Features into the General Form of Chou's Pseudo Amino Acid Composition." *Protein & Peptide Letters* 19 (5): 492–500. https://doi.org/10.2174/092986612800191080.

Supplementary Data

Feature Extraction

Example of modification of skips in k-skip-bi-gram motifs: For example, if X represents a single skip, the motifs MXXH and MXH are considered unique without buffering. However, if the skips are buffered by 2 (a = 2), the buffered skip value of motif MXXH will be c = 2 + ((2 − 2)%2) = 2, yielding the original motif MXXH. On the other hand, skip buffering MXH gives the value c = 1 + ((2 − 1)%2) = 2, and yields a new motif MXXH. This motif is different from the original MXH, but is identical to the previous example MXXH. In this way, the buffered skip model can account for insertion/deletion events.

Example of modification of estimated C-terminus position in n-grams and k-skip-bi-grams: As an example, if we consider NTerm-AYHGFTVCKY-CTerm as a protein sequence, then two tyrosines will be members of the set of uni-gram motifs, and should be considered as identical. However, if we choose to account for position, each will be assigned position identity information as defined by equation (5). If the initial buffer value y0 equals 5 then the positional identity of the first Y and the last Y will be x = 9 + ((5 − 9)%5) = 10 and x = 1 + ((5 − 1)%5) = 5, respectively. Here the distance of first Y is 9 and the second Y is 1 from the C-terminus. In this way, rather being identical, the tyrosines will be recorded as Y10 and Y5 in the feature set. This approach can be generalized to n-grams. The bi-gram AY has a positional identity of 12, because its onset is 10 residues away from the C-terminus, and the buffer value will be 6 because y0 is 5 and the length of the motif is 2.

Feature Selection and Model Construction

Ultimately, six parameters determine the final set of features to be generated from a given sequence (Table S3.1). The feature extraction algorithm generates descriptors (motifs) from a list of protein sequences, which function as words in a document. To reduce noise, words that make up more than 30% of the corpus and words that appear less than 3 times are removed as an alternative to tf-idf (Joachims T 1996). Next, the model creates a sparse matrix using a vectorization method where each of the retained words or motifs composes a vector. The value of the vectors for data-points in the sparse matrix describes the presence or absence of the feature in a corresponding data-point. In other words, each row of a vector reports the presence of a selected motif in a protein sequence. Finally, a logistic regression model (Ruczinski I et al.,2003) is trained with the training data set, and its accuracy is calculated with five-fold cross-validation. The model construction scheme is done in python 2.7 using the numpy, pandas and scikit-learn packages (Pedregosa F et al., 2011). When running logistic regression, a regularization constant of 1 and default parameters are used.

Parameter Optimization Algorithm

The feature generation function depends on the six parameters described in Table S3.1. Here, a modified grid-search optimization algorithm, Algorithm 3.2, chooses parameters for generalized classification problems based on the accuracy of five-fold cross-validation using a logistic regression model. Briefly, it iterates over pairs of parameters to maximize accuracy, using maximal previous knowledge to inform future iterations. Each grid-search is initiated from a value of the parameter for the n-gram motifs (n) which is referred to here as the seed.

Algorithm 3.1 Logistic Regression Accuracy

1: **procedure** logRegAcc(k, n, kp, np, y, c)
2: Generate features using the given parameters
3: Run logistic regression using the given features, and determine accuracy by five-fold cross validation
4: **return** the accuracy from logistic regression

Algorithm 3.2 Modified Grid Search

1: **procedure** ModifiedGridSearch(The parameters that yielded the best accuracy in logistic regression on the test set)
2: **for** superSeed = 1,3,...25 **do**
3: Initialize all parameters to superSeed
4: **while** True **do**
5: $k = \text{argmax}_k$ logRegAcc(k,n,kp,np,y,c)
6: $n = \text{argmax}_n$ logRegAcc(k,n,kp,np,y,c)
7: **if** k and n are unchanged **then**
8: break
9: **while** True **do**
10: $kp = \text{argmax}_{kp}$ logRegAcc(k,n,kp,np,y,c)
11: $np = \text{argmax}_{np}$ logRegAcc(k,n,kp,np,y,c)
12: **if** kp and np are unchanged **then**
13: break
14: **while** True **do**
15: $y = \text{argmax}_y$ logRegAcc(k,n,kp,np,y,c)
16: $c = \text{argmax}_c$ logRegAcc(k,n,kp,np,y,c)
17: **if** y and c are unchanged **then**
18: break
19: **return** parameter values with best 5-fold cross validation accuracy from all trials.

Run-Time Study of Optimization Algorithm

A run-time study was executed on the optimization step using i2AMP-2L, Cypred, PreSTP, TumorHPD 1, TumorHPD 2, HemoPI 1, HemoPI 2, IGPred and PVPred dataset. Each dataset was fractioned into 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%

and 100% sequences of the different classes. Then the time required to optimize the parameters of each of fraction for each dataset were measured and plotted.

Comparison of the Performance of N-Gram and Skip-Gram Based Feature Generation Models with and without Modifiers

To observe the impact of the modifiers of n-grams and skip-grams (see equation 3, 5 and 6 in the main text) on the prediction models, we have compared the features extracted from non-modified and modified n-gram and skip-gram models. To do that, we selected the parameter combination (see Supplement Table S3.1.) from the seed generating the best accuracy during optimization (see Algorithm 3.1 and 3.2 in the Supplement text) for each benchmark dataset and compared the corresponding accuracy without modifying the n and k parameters (by changing the parameters np and kp into 0 and c into 1). Thereafter, we compared the change of accuracy due to changing the parameters with a five-fold cross-validation for each dataset.

Comparison among Different Feature Generation Method and Classifier Combinations

To check the suitability of other classifiers with the automated feature selection method in m-NGSG, we compared SVM with linear and non-linear kernels and K-nearest neighbors(KNN). We also compared the Quantitative Structure-Property Relationship (QSAR) based feature generation methods (1) combining with the same classifiers to observe the optimal combinations. For KNN and SVM with non-linear kernel (here we used rbf kernel) classifiers, we used the default parameters from the scikit-learn packages. We chose the Subchlo60 dataset from the Subchlo model (see Supplement Table S3.6.) because that data is a complex one to classify among the twelve benchmark datasets (see Supplement

Table S3.6.) as it contains the protein sequences less 60% identity which come from four different classes (see Supplementary Table S3.4.).

Comparison among the Models with Noise in Subchlo60 Dataset

To check the influence the data of random noise on different models, we added 8% random sequences for the four different classes of Subchlo60 dataset (see Supplement Table S3.6.). Thereafter, we compared the best feature-classifier combinations obtained from the classifier comparison in the previous section. This comparison was done with jackknife, five-fold and ten-fold cross validations.

Table S3.1. Description of parameters employed in m-NGSG (modified n-gram skip-gram) based feature generation from an individual sequence.

n	determines the maximum length of an n-gram motif
k	determines the maximum number of skips in a k-skip-bi-gram motif
np	determines the maximum length of an n-gram motif that gets a positional value
kp	determines the maximum skips in a k-skip-bi-gram motif that gets a positional value
y	determines the positional buffering parameter in both n-gram and k-skip-bi-gram motifs
c	determines the skip buffering parameter in k-skip-bi-gram motifs

Table S3.2. Comparison between cross-validation accuracies reported on different benchmark training datasets and the corresponding accuracies achieved employing the m-NGSG model. Accuracies are displayed in percentage values.

Classification dataset	Reported accuracy on the training set	m-NGSG accuracy on the training set
Subchlo raw	89.69	91.59
Subchlo60	67.18	73.92
iAMP-2L	86.32	91.25
Cypred	99.2	99.55
PredSTP	94.3	96.66
TumorHPD 1	82.52	83.4
TumorHPD 2	80.28	82.55
HemoPI 1	95.3	97.97
HemoPI 2	78	79.5
IGPred	96.92	91.66
IGPred with feature selection	96.9	100
PVPred	85.02	77.19
PVPred with feature selection	85.02	93.48

Table S3.3. Comparison between accuracies reported on independent test sets and the corresponding accuracies achieved employing the m-NGSG model. Accuracies are displayed in percentage values.

Classification dataset	Reported accuracy on the test set	m-NGSG accuracy on the test set
iAMP-2L	92.23	96.47
CypredL	98.7	98.98
HemoPI 1	96.4	99.54
HemoPI 2	75.7	76.23
IGPred	100	100
IGPred with feature selection	100	100
PVPred	86.66	93.33
PVPred with feature selection	86.66	93.33

Table S3.4. Description of number of classes and class size the datasets used meta-comparison.

Data-set	Training set				Test	
	Class 1	Class 2	Class 3	Class 4	Class 1	Class 2
Subchlo	60	71	516	90	NA	NA
Subchlo60	42	50	126	39	NA	NA
osFP	40	64	0	0	10	23
iAMP-2L	843	2405	0	0	920	920
Cypred	55	342	0	0	54	341
PredSTP	145	442	0	0	NA	NA
TumorHPD 1	651	651	0	0	NA	NA
TumorHPD 2	469	469	0	0	NA	NA
HemoPI 1	442	442	0	0	110	110
HemoPI 2	442	370	0	0	110	92
IGPred	109	119	0	0	20	20
PVPred	99	208	0	0	11	19

Table S3.5. Comparison of evaluation matrices of the m-NGSG model with the feature generation methods used in osFP dataset. Accuracies are displayed in percentage values.

Methods	CV Accuracy%	CV MCC	Test set Accuracy%	Test set MCC
AAC/DPC/TPC	72.13 ± 4.18	0.42 ± 0.08	72.89 ± 7.08	0.43 ± 0.15
AC	70.71 ± 4.45	0.38 ± 0.09	70.30 ± 8.55	0.38 ± 0.18
CTD	69.40 ± 4.95	0.39 ± 0.10	70.18 ± 7.79	0.38 ± 0.17
Ctriad	68.64 ± 5.99	0.34 ± 0.12	71.26 ± 8.36	0.40 ± 0.17
QSO	68.98 ± 4.21	0.34 ± 0.09	69.93 ± 6.90	0.37 ± 0.14
PseAAC	69.39 ± 4.97	0.35 ± 0.10	69.67 ± 8.03	0.36 ± 0.17
m-NGSG	78.02 ± 0.93	0.50 ± 0.02	79.21 ± 1.47	0.54 ± 0.03

Table S3.6. Description of the models those are used in meta-comparison.

Model Name	Dataset Description	Original Feature Extraction Method	Classifier
Subchlo	Two pair of training datasets of protein sequences from 4 classes based on their localization in choloplast. One dataset includes the raw sequences of proteins here annotated as "Subchlo raw". Another dataset consists of sequences less than 60% identity annotated as "Subchlo60"	Pseudo Amino acid composition (PseAAC)	Evidence theoretic K-nearest neighbor(ET-KNN) with a jackknife cross validation.
osFP	One pair of training and testing dataset of protein sequences from 2 classes (monomer vs oligomer) based on oligomeric states of proteins. We selected the benchmark protein sequences less than 95% identity.	AAC/DPC/TPC, AC, CTD, Ctriad, QSO, PseAAC	Features having a threshold 0.7 for the Pearson correlation coefficient were removed.Decision tree with ten-fold cross validation following splitting the whole training set into 80% and 20% training and test set respectively. This process was repeated for 100 time to get an unbiased confidence interval of the accuracy.
iAMP-2L	One pair of training and test dataset of two classes of protein sequences based on antimicrobial activity	PseAAC	fuzzy k nearest neighborhood (FKNN) with a Jackknife cross validation.
Cypred	One pair of training and test dataset consist of two classes of sequences based on cyclic and noncyclic structure.	AAC, cyclicpeptide specific motifs	SVM with (RBF) kernel coupled with 10-fold cross validation.

PredSTP	One training dataset divided into two classes based the cysteine bonding pattern in the 3D structure of the proteins.	Normalized distance between cystine pairs explained along with hydrophobic, hydrophilic, neutral and count of some other amino acids.	SVM with RBF kernel coupled with 200-fold cross validation.
TumorHPD	Two training datasets annotated as TumorHPD 1 and TumorHPD 2 in this paper. The sequences are divided in two classes based on their affinity to tumor cells. TumorHPD 1 consists the raw protein sequences while TumorHPD 2 consists only the sequences not more than 10 amino acids long.	AAC, DPC, BPP	SVM with 5-fold cross validation.
HemoPI	Three pairs of training and test set annotated as HemoPI, semiHemoPI and nonHemoPI two in the paper which contains hemolytic, semihemolytic and nonhemolytic peptides, respectively. Here we compared model that classifies the raw hemolytic and nonhemolytic peptides annotated as HemoPI 1, and the model that classifies hemolytic and semihemolytic peptides annotated as HemoPI 2	AAC, DPC, BPP	SVM with 5-fold cross validation.
IGPred	One pair of training and test set those are divided into two classes those fall into two groups: immunoglobulin and nonimmunoglobulin	PseAAC followed by ANOVA based feature Selection technique	SVM with RBF kernel coupled with a Jackknife cross validation.
PVPred	One pair of training and test set those are divided into two classes those fall into two groups: virion and nonvirion.	g-gap dipeptide composition plus PseAAC followed by ANOVA based feature Selection	SVM with RBF kernel coupled with a Jackknife cross validation.

Abbreviations: PseAAC = Pseudo amino acid composition; APC = Amino acid composition; DPC = Dipeptide composition; TPC= Tripeptide composition; AC = Auto correlation; CTD = Composition,Transition, Distribution; Ctraid= Conjoint triad; QSO=Quasi-sequence-order; BPP = binary profile pattern (presence or absence of a motif of interest)

Figure.S3.1. Illustrates the optimum values of individual parameters generated from each seed for a specific dataset. The values of the six parameters were optimized from thirteen different seeds (the initial value of n). Each column in the panels (n, k, np, kp, y, c) assign the parameters and each row represent individual dataset. The X axis shows the value of the parameters, while the y axis represents an individual seed.

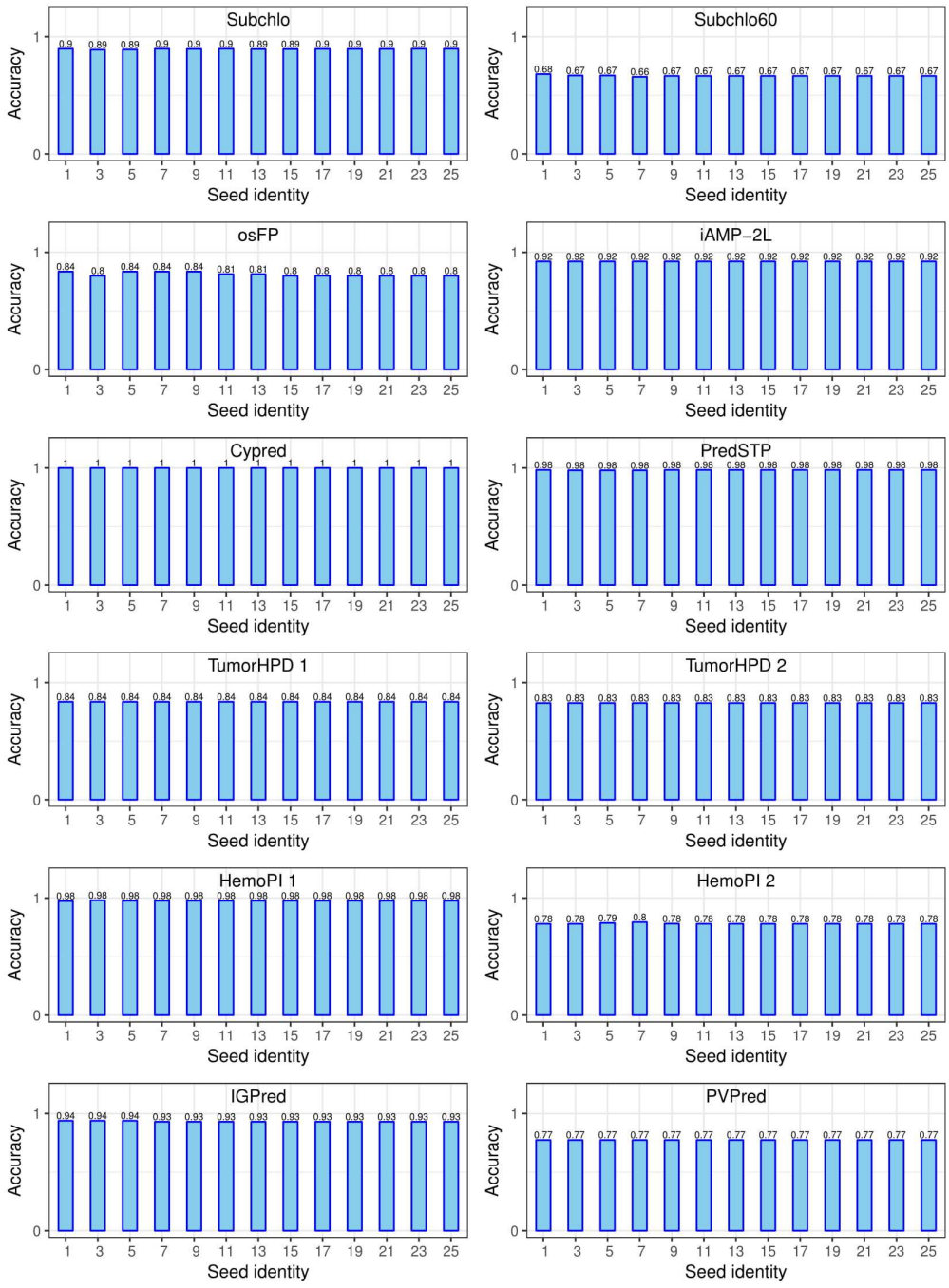

Figure S3.2. Represents the accuracies resulted from different seeds in a specific dataset. Each subplot represents an individual dataset. The x axis shows seed identities and the y axis shows the accuracy values.

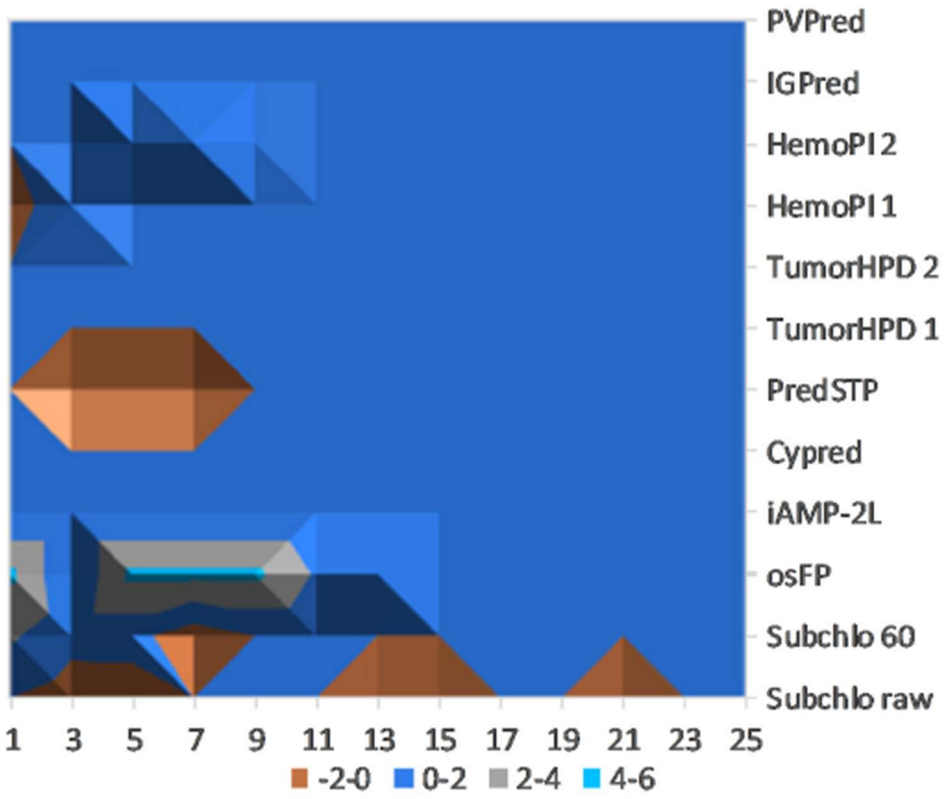

Figure S3.3. The percent change of accuracy for each seed compared to the mode accuracy of all the seeds for a specific dataset. The smaller the percentage deviation from the mode value, the better its convergence. iAMP-2L, Cypred, TumorHPD 1, TumorHPD 2, IGPred and PVPred showed perfect convergence, while the other datasets shows convergence for most of the seeds.

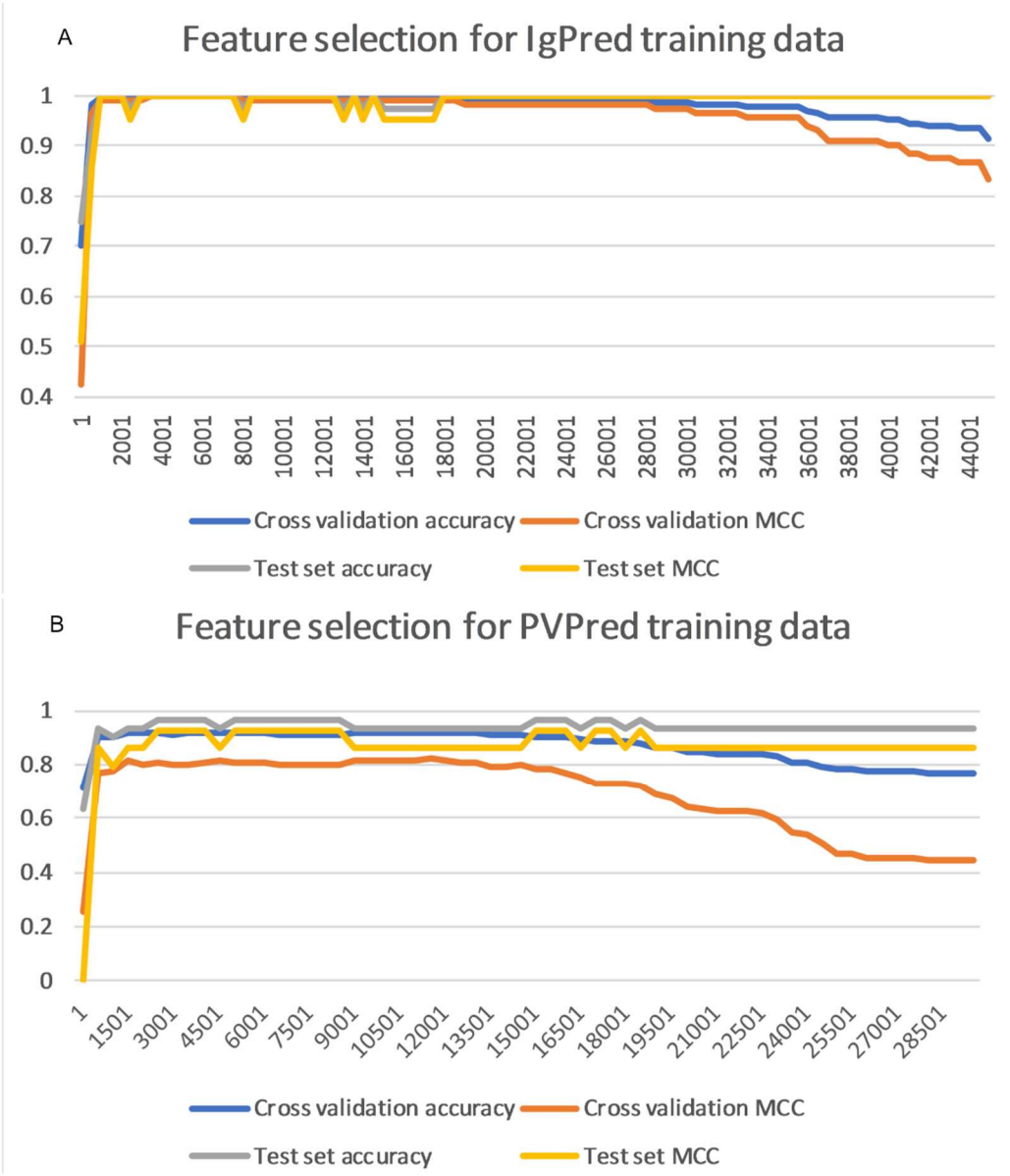

Figure S3.4. Performance of accuracies with number of features used. The features were added based on their importance according to ANOVA analysis. Using less features increase the cross-validation accuracy while a decrease of accuracy on the independent test set is evident as less features engender a bias cross validation. Figure A and B show the effect on accuracy and MCC values with increasing number of features on IGPred and PVPred datasets, respectively.

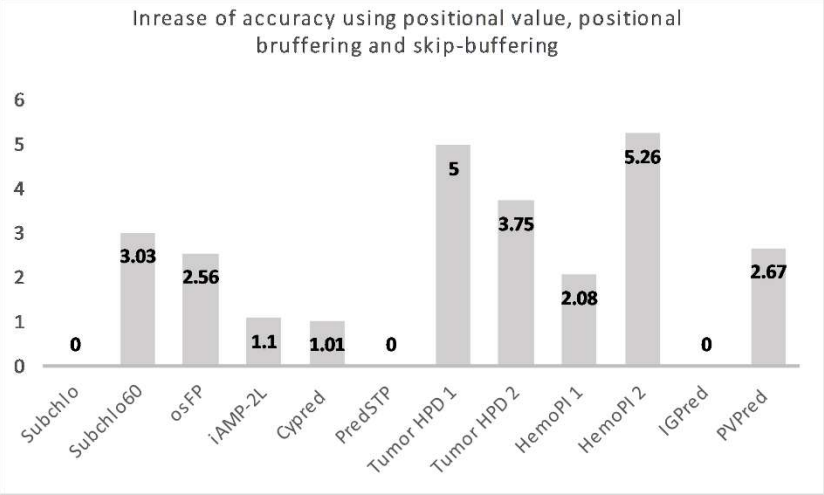

Figure S3.5. Represents the increase of five-fold cross-validation accuracies resulted from addition of positional values (np and kp), position value buffering(y) and skip-buffering(c) parameters for each dataset. Each bar represents an individual dataset and the height of the bars represent the percentage increase of accuracies compared the features generated using n-grams and skip-grams.

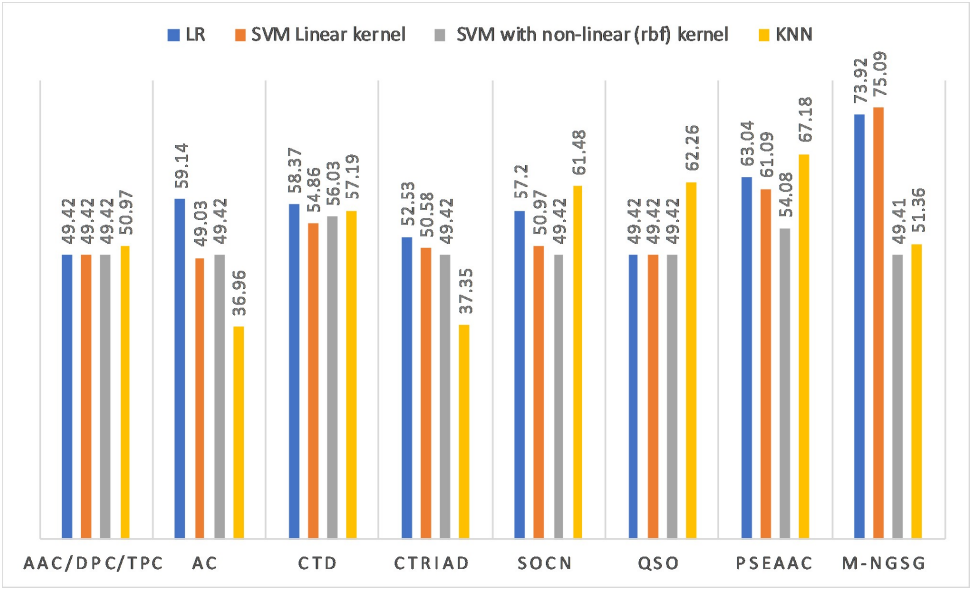

Figure S3.6. Comparison of accuracies among different combinations of feature generation models and classifiers on the Subchlo60 dataset. Combination of m-NGSG with a linear SVM or a Logistic regression shows the best accuracy. The description of the other feature generation described in the abbreviations part of Supplementary Table S6.

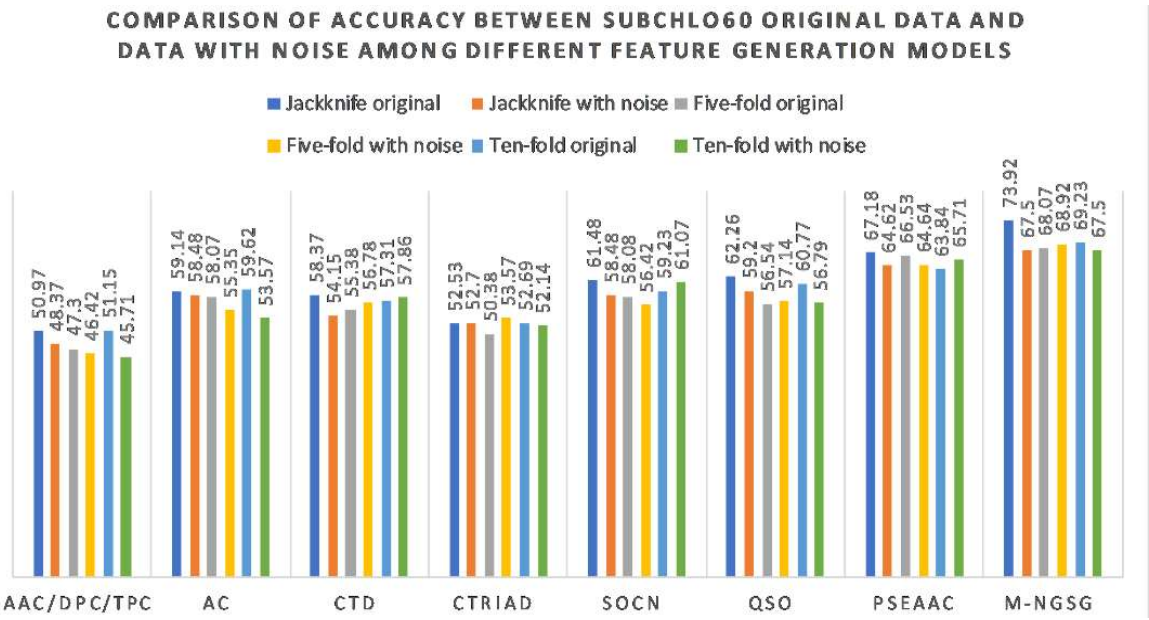

Figure S3.7. comparison of accuracies between Subchlo60 original and data and data with noise among different QSAR and Amino Acid composition related, and m-NGSG feature generation models. Each group of bars indicates the type of the feature generation model with the optimal classifier. Height of the bars represent the accuracies and color of the bars represent different k-fold cross-validations. m-NGSG coupled with a Logistic Regression shows the best accuracies for each k-fold cross-validation.

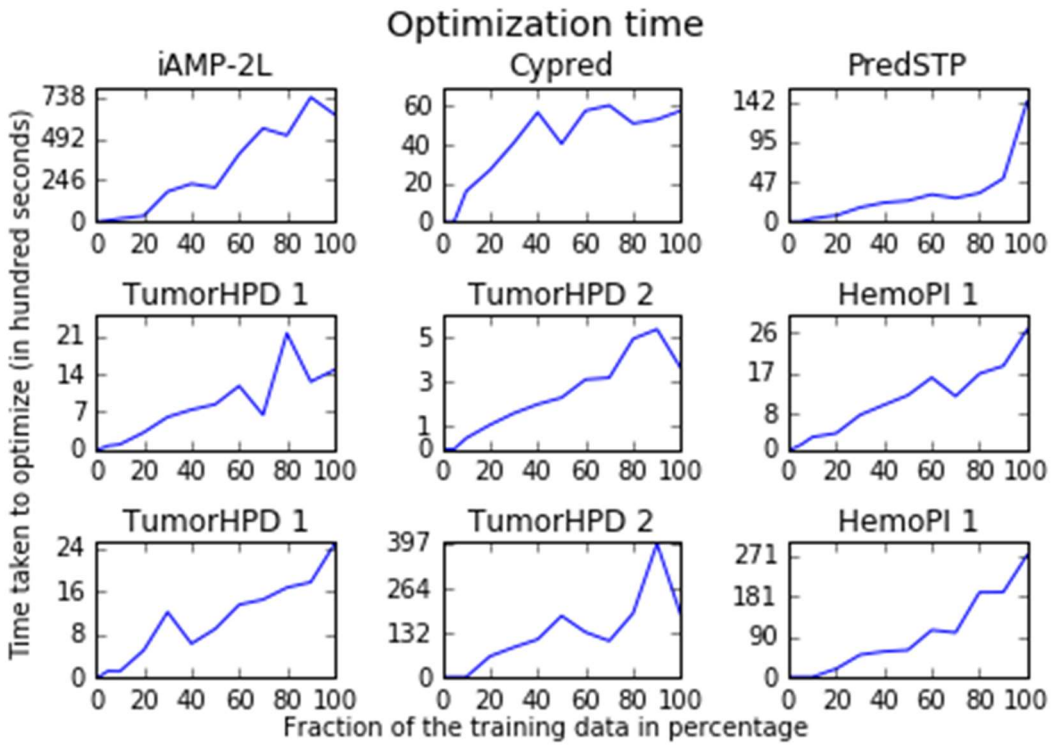

Figure S3.8. Illustrates the run-time profile during optimization for different datasets. Each subplot represent run-time for different percentage fractions of a dataset. The X-axis of a subplot indicates each percentage fraction and Y-axis indicates the run-time of optimization in seconds.The title of the subplot denotes the specific dataset. The number of sequences of 100% in iAMP-2L, Cypred, PredSTP, TumorHPD 1, TumorHPD 2, HemoPI 1, HemoPI 2, IGPred, PVPred are 3248, 397, 587, 1302, 938, 884, 812, 228, 307, respectively. Analysis was performed on a PowerEdge R630, with 32-CPUs @ 2.4Gz, 64 GB RAM, running Red Hat Enterprise Linux 7.

CHAPTER FOUR

Assigning Biological Function Using Hidden Signatures in Cystine-Stabilized Peptide Sequences

Abstract

Cystine-stabilized peptides have great utility as they naturally block ion channels, inhibit acetylcholine receptors, or inactivate microbes. However, only a tiny fraction of these peptides has been characterized. Exploration for novel peptides most efficiently starts with the identification of candidates from genome sequence data. Unfortunately, though cystine-stabilized peptides have shared structures, they have low DNA sequence similarity, restricting the utility of BLAST and even more powerful sequence alignment-based annotation algorithms, such as PSI-BLAST and HMMER. In contrast, a supervised machine learning approach may improve discovery and function assignment of these peptides. To this end, we utilized our previously described m-NGSG algorithm, which utilizes hidden signatures embedded in peptide primary sequences that define and categorize structural or functional classes of peptides. From the generalized m-NGSG framework, we derived five specific models that categorize cystine-stabilized peptide sequences into specific functional classes. When compared with PSI-BLAST, HMMER and existing function-specific models, our novel approach consistently demonstrates superior performance in discovery and function-assignment. We also report an interactive version of CSPred, available through download (https://bitbucket.org/sm_islam/cystine-stabilized-proteins/src) or web interface (watson.ecs.baylor.edu/cspred), for the discovery of cystine-stabilized peptides of specific function from genomic datasets and for genome

annotation. We fully describe, in the Availability section following in the Discussion, the quick and simple usage of the CsPred website to automatically deliver function assignments for batch submissions of peptide sequences.

Introduction

Cystine-stabilized peptides are impressively abundant and widespread across the taxa. They form the neurotoxic venom fraction of spiders (King and Hardy 2013), snakes (Chan et al. 2016), scorpions(Ortiz et al. 2015), sea anemones(Frazão, Vasconcelos, and Antunes 2012), jellyfish, corals and conch (Akondi et al. 2014)and may be specific for insects, mammals, or reptiles. Other cystine-stabilized peptides serve as antimicrobials (Nguyen, Haney, and Vogel 2011) and defensins in humans, insects, fungi, plants and most other taxa. Functionally, the venom peptides include sodium (Munasinghe and Christie 2015), calcium (Bourinet and Zamponi 2016) and potassium (Norton and Chandy 2017) ion channel blockers, acetylcholine receptor inhibitors (Dutertre, Nicke, and Tsetlin 2017), or protease inhibitors (Mourão and Schwartz 2013). Antimicrobial peptides generally act as membrane disrupters specifically against bacterial or fungal cells, but, due to their ability to penetrate cell membranes, they can also enter eukaryotic cells to act on host DNA directly and to modulate immune responses (Nguyen, Haney, and Vogel 2011). The stability of these peptides and their specific and powerful functions make them strong candidates for a variety of medical and agricultural applications, including pain relief, disruption of cancer development, and environmentally friendly insecticides, fungicides and bactericides, delivered either directly or via transgenes. However, only a tiny fraction of cystine-stabilized peptides has been characterized experimentally (Mobli, Undheim, and Rash 2017; Silverstein et al. 2007; Kuzmenkov,

Grishin, and Vassilevski 2015). To sort through the huge number of remaining cystine-stabilized peptides present in such a wide range of genomes for the purpose of classifying each of these peptides into one of the disparate functional groups, an efficient automated approach is warranted.

Sequence identity of the cystine-stabilized peptides varies broadly and can be distributed into different structural/motif and family-based (the native source of a peptide) classes (Cheek, Krishna, and Grishin 2006). For example, the scorpion toxin-like superfamily (Kuzmenkov, Grishin, and Vassilevski 2015; Santibáñez-López and Possani 2015; Possani et al. 1999), agatoxins (Adams 2004), and conotoxins (Olivera et al. 1985) are examples of family-based classes, while STPs (S. M. A. Islam et al. 2015), NTPs(S. M. A. Islam et al. 2015), cyclotides (Craik, Simonsen, and Daly 2002) and knottins (Gracy et al. 2007) are examples of structure or motif-based classes. Because of the high degree of heterogeneity in their primary sequences, several sequence alignment independent models have been reported to classify the structure the family of cystine-stabilized /disulfide-rich. For instances, Cypred (Kedarisetti et al. 2014) predicts cyclic peptides including cyclotides; Knotter 1D predict peptides with ICK motifs (Gelly 2004); iCTX-Type predicts types of Conotoxins targeting Ion Channels (Ding et al. 2014); PredCSF predicts conotoxin superfamily from the primary protein sequences(Fan et al. 2011); and, PredSTP predicts sequential tri-disulfide motifs in cysteine rich peptide (S. M. A. Islam et al. 2015). While the currently published models either classify the structural proprotein or predict the family of the peptides none of those are dedicated to predicting functional characteristics in a family or structure agnostic fashion. However, a specific functional

group of cystine-stabilized peptides often come from different family or structural classes. Thus, the family or structure/motif-based classification will not reveal the functional characteristic of a peptide. Under this context, it is necessary to develop a sequence alignment independent model to discover the functional characteristics in a family of origin or structure agnostic fashion.

Machine learning-based supervised models are widely used to predict the functional and structural class of proteins which are difficult to predict using sequence alignment-based algorithms. However, it is imperative to extract the relevant feature vectors (descriptors) and to implement an optimized classification algorithm to get expected performance from a model. Several classification algorithms have already been exploited to predict protein characteristics from the primary sequences (Sharma et al. 2013; Simeon et al. 2016; Du, Cao, and Li 2009), but, extracting proper descriptors from protein sequences remains a challenging task. A number of descriptors, such as amino acid composition (Zhang and Fang 2008) , autocorrelation (Xia, Han, and Huang 2010), CTD (composition, transition, and distribution) (Dubchak et al. 1995), conjoint triads (Chang, Syu, and Lin 2010) and pseudo amino acid compositions (Shen and Chou 2008) are routinely used to build machine learning-based models. Recently, we demonstrated a complete pipeline of a classifier constructor where the feature generation model is integrated with a logistic regression algorithm (S. M. A. Islam et al. 2017) . This training set pipeline is denoted as mNGSG (*modified n-gram* and *skip-gram*) where a modified *n-grams* (Keˇselj et al. 2003)and *skip-grams*-based (Guthrie et al. 2006) framework is used to generate descriptors from the protein sequences and utilize the hidden signatures from the descriptors for the supervised classification(S. A. Islam et al. 2017). In a this study,

we found m-NGSG to be more accurate than other commonly used descriptors which underpin this as a pipeline to construct reliable supervised prediction models (S. M. A. Islam et al. 2017).

In this study, we applied m-NGSG to build five individual models to predict ion channel blockers, antimicrobial peptides, acetylcholine receptor inhibitors, serine protease inhibitors, and hemolytic proteins from disulfide stabilized proteins. Identification of hemolytic characteristics will allow the researcher to eliminate from consideration proteins cytotoxic to humans. The results demonstrate superiority of m-NGSG-based models to PSI-BLAST (Altschul 1997) with different E-values, HMMER (Finn, Clements, and Eddy 2011) and other available models. Finally, we constructed CSPred model which combines the results of the five different models and gives a probability score for the five important functional characteristics of cystine-stabilized proteins. Since the ion channel blockers is consists of three major subclasses: sodium, potassium and calcium channel blockers, we also constructed three classifiers to see ability of m-NGSG to classify the ion channel blockers into those three subclasses.

Methods and Materials

Data Acquisition and Preparation

The positive and negative datasets for ion channel blockers (ICB), antimicrobial peptides (AMP), acetylcholine receptor inhibitors (ACRI), serine protease inhibitors (SPI), and hemolytic proteins (HLP) are generated by obtaining protein sequences from UniprotKB (knowledgebase) (Boeckmann 2003) using the search keys mentioned in Supplement Table S4.1. All the protein sequences, including positive and negative classes, contain a minimum of one disulfide bond and a chain size of less than 150 amino

acid residues. Thereafter, the protein sequences are curated manually based on the functional attribute for each entry. A part of the HLP positive dataset is collected from the HemoPI server (Chaudhary et al. 2016). Here, only the sequences containing a minimum of one pair of cysteines are selected from the dataset. The CD-HIT software (Huang et al. 2010) is used to organize sequences based on identity thresh-holds to generate final datasets for each functional group of the cysteine stabilized proteins (See Supplement Table S4.1). From the positive and negative datasets of each selected functional group, 90% of the chains are retained for training sets, while 10% of the chains are reserved for out-of-sample test sets using a random shuffle-split process. The numbers of chains in each training and test sets are mentioned in Supplement Table S4.1. Further, to construct a separate compound model to classify the ICB into three different subclasses, we made three separate models using six different training sets which are listed in Supplement Table S4.2. The ICB classifier were contracted to classify the ICBs into sodium, potassium and calcium channel blockers which are three main subclasses of ICB.

Table 4.1: Comparison of evaluation matrices between the training and the out-of-sample test sets for each functional group-based model. The precision, recall and accuracy values are shown in percentages.

Model	**Precision**		**Recall**		**F1-Score**		**Accuracy**		**MCC**	
	Training set	Test set	Training set	Test set	Training set	Test set	Training set	Test set	Training set	Test set
ICB	91.25	95.58	83.80	92.85	0.87	0.94	89.67	95.32	0.78	0.90
AMP	86.56	85.96	77.08	81.66	0.81	0.84	86.33	87.74	0.71	0.74
ACRI	100.00	100.00	0.80	63.63	0.89	0.78	95.23	92.00	0.87	0.76
SPI	97.52	96.43	79.66	81.81	0.88	0.88	91.90	92.55	0.83	0.84
HLP	86.07	92.30	86.66	80.00	0.86	0.86	89.39	89.47	0.78	0.78

Abbreviations: ICB = Ion channel blocker; AMP = Antimicrobial peptide; ACRI = Acetylcholine receptor inhibitor; SPI= Serine protease inhibitor; HLP = Hemolytic protein

Model Construction Using m-NGSG

Five different binary classifiers are constructed to predict each of the five selected functional classes using the m-NGSG algorithm(S. M. A. Islam et al. 2017) . The m-NGSG algorithm (avialable at https://bitbucket.org/sm_ islam/mngsg/src) offers an integrated and fully automated feature generation method followed by a logistic regression-based model construction, feature generation, and parameter optimization as described in(S. M. A. Islam et al. 2017). Parameter optimizations employed five-fold cross-validation using appropriate training sets. Supplement Table S4.3 illustrates the parameters selected by the m-NGSG optimizer for each functional group specific model. A combined model CSPred is further derived from the result aggregation of the five-individual function-based models. A diagram of the CSPred model construction is delineated in Figure 4.1.

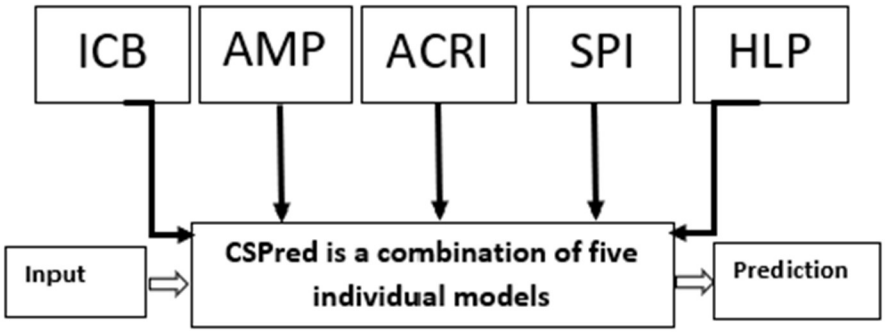

Figure 4.1: Work flow of the construction and application of CSPred.

Model Evaluation

The performances of all five models were evaluated using a five-fold cross-validation. Precision (*eq. 1*), recall (*eq. 2*), F1-score (*eq. 3*), accuracy (*eq. 4*), and Mathews Correlation Coefficient (MCC) (*eq. 5*) values are calculated for each model as the evaluation matrices. For calculation of these evaluation matrices, the confusion matrices

were constructed to calculate the True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). TP and TN are correctly predicted positive and negative data points, respectively. Similarly, FP and FN are incorrectly predicted positive and negative data points, respectively. From TP, TN, FP, and FN, the evaluation matrices were calculated using the following equations:

$$Precision = \frac{TP}{TP+FP} \quad (eq.\ 4.1)$$

$$Recall = \frac{TP}{TP+FN} \quad (eq.\ 4.2)$$

$$F1\ score = \frac{2TP}{2TP+FP+FN} \quad (eq.\ 4.3)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+F} \quad (eq.\ 4.4)$$

$$MCC = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+\quad)(TP+FN)(TN+F\quad)(TN+FN)}} \quad (eq.\ 4.5)$$

Comparison with PSI-BLAST and HMMER

The performance of each model except the subclasses of ICB is compared with PSI-BLAST (Altschul 1997) and HMMER (Finn, Clements, and Eddy 2011). The ncbi-blast-2.5.0+ standalone software was downloaded to run PSI-BLAST on a local computer. Similarly, hmmer 3.1b2 was installed in linux operating system. Phmmer function was used to run hmmer with the default parameters. The evaluation matrices are calculated for the identical training sets with PSI-BLAST using five-fold cross-validation. During cross-validation with both PSI-BLAST and phmmer, the training set was employed to make the database, and the test set is used as the query. Class of each query sequence was predicted by the highest matching score with the sequences in the database. For PSI BLAST, the same cross-validation were conducted by taking the

thresh-hold E-value 0.01, 0.05, 0.1, 0.5, 1.0 and 5.0 with five iterations. All other parameters were kept the default. Afterwards, the sequence of the out-of-sample test sets from each functional group is predicted keeping the sequence of the corresponding training sets as databases.

Comparison with Other Available Models

We searched for other available models dedicated to classifying any of the five functional groups. We found iAMP-2L (Xiao et al. 2013) and $CAMP_R3$ (Waghu et al. 2016) are available to predict antimicrobial peptides, but not limited to predict cysteine stabilized peptides. We also compare the performance of our AMP model with iAMP-2L and $CAMP_R3$. $CAMP_R3$ offers four different classifiers to predict AMPs: Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Network (ANN), Discriminant Analysis (DA). We compared our AMP model with all classifiers offered by $CAMP_R3$ using the out-of-sample test set and calculated precision, recall, F1-score, accuracy and MCC values. Similarly, there HemoPI (Chaudhary et al. 2016) model dedicated to predicting haemolytic peptides. Therefore, we also compared our HLP model with HemoPI using the out-of-sample test set of HLP and calculated evaluation matrices.

Results

Evaluation of the m-NGSG-Based Models

ICB, AMP, ACRI, SPI and HLP represent five different functional class-based models constructed using the m-NGSG algorithm(S. M. A. Islam et al. 2017) . Each model was evaluated using precision, recall, F1-score, Accuracy and MCC scores based on a five-fold cross-validation against a training set. The evaluation matrices are reported

in Table 4.1. The accuracy of the five models range from 86.33% to 95.2% where AMP and ACRI rendered the lowest and highest accuracy, respectively. The models also generated F1-Scores ranging from 0.81 to 0.89 and MCC scores ranging from 0.74 to 0.90.

To judge the robustness of our approach, it is imperative to compare the performance of the model against established reliable available methods. PSI-BLAST and HMMER are used for generalized comparison, while other comparison groups are more specific. iAMP-2L and $CAMP_R3$ are used to evaluate performance against AMPs, and MemoPI for HLP. The models those classify the ICB proteins into the sodium, potassium and calcium channel blockers were evaluated by AUC, accuracy, MCC values using a five-fold cross-validation and out of sample test set (Supplement Figure S4.5). Based on the evaluation matrices, it can be said that m-NGSG demonstrate a quite good performance to separate the subclasses of the ICB proteins.

Comparison of the Evaluation Matrices and Area Under Curve (AUC) with PSI-BLAST and HMMER

PSI-BLAST is a dependable and widely used algorithm to discover distantly related protein sequence using PSSM matrices (Altschul 1997). On the other hand, HMMER is a Hidden Markov Model-based algorithm designed to detect remote homologs with a high sensitivity (Finn, Clements, and Eddy 2011). Therefore, we compared the performance of each constructed model with PSI-BLAST and HMMER for the corresponding training sets using a five-fold cross-validation. Supplement Figure S4.1 shows an extensive comparison among the m-NGSG based models, HMMER and PSI-BLAST models made with different E-values. Precision, recall, F1-score, accuracy and

MCC values are used to evaluate the models against PSI-BLAST. Figure 4.2A and Supplement Figure S4.2 specifically shows the comparison of the MCC values of each training sets with PSI-BLAST and HMMER. Figure 4.2B and Supplement Figure S4.3 illustrates the standard deviation of the MCC values generated from different folds using different models. The area under curve (AUC) for the five-different m-NGSG-based models were also compared with PSI-BLAST and HMMER using the corresponding training sets. The E-values yielding the best MCC values for each function-based training sets were used to run a PSI-BLAST for the comparison. Figure S4.3 shows the receiver operating characteristic (ROC) curves for m-NGSG-based models with a side by side area under curve (AUC) comparison among each m-NGSG-based model and the corresponding PSI-BLAST and HMMER-based models. For the five training sets, m-NGSG based models generated better AUCs compared to the corresponding PSI BLAST and HMMER based models.

Comparison of the Evaluation Matrices with PSI-BLAST and HMMER on the Out-of-Sample Test Set

Versatility of the five m-NGSG based models were tested by comparing their performance with PSI-BLAST and HMMER on the corresponding out-of-sample test set. We imported the same E-values from the ROC curve comparison to run the PSI-BLAST on the test sets. The MCC values were measured for each model and the corresponding PSI-BLAST and HMMER to achieve an appropriate comparison. Figure 4.4 displays comparative bar plots which illustrate the MCC values on the out-of-sample test set produces five different models and PSI-BLAST. According to Figure 4.4, each of the five models shows better MCC values compared to their equivalent PSI-BLAST results while

four models except AMP show better MCC value than HMMER. In case of AMP, both m-NGSG-based model and HMMER shows the same MCC value (0.74).

Comparison of AMP and Hemolytic Peptide Prediction Models with other Currently Available Models

Along with PSI-BLAST and HMMER, we used the iAMP-2L 44 and CAMPR3 45 models to predict antimicrobial peptides (AMP), and the HemoPI 42 algorithm to predict hemolytic peptides. While it is important to note that none of these models are dedicated to the identification of only cystine- stabilized peptides, their performance parameters should generalize to their prediction. We compared performances of iAMP-L2 and CAMPR3 with our m-NGSG-based AMP model and HemoPI with the m-NGSG-based HLP model using the corresponding out-of-sample test sets. Figure 4.5 shows the comparative precision, recall, accuracy and MCC values among different models. Among the other available models, CAMP-ANN showed the highest precision score 0.43 or 43% while the precision score produced by m-NGSG-based AMP model was 0.85. CAMP-SVM showed a slightly better recall score than m-NGSG, 0.83 and 0.81, respectively. Overall, the best accuracy score was generated by iAMP-2L (0.54), but it was far less than the accuracy score produced by m-NGSG, which was 0.75. Finally, the highest MCC score was generated by CAMP-ANN (0.13) which was also well below the MCC score of m-NGSG (0.74) (see Figure 4.5). Similarly, Supplement Figure S4.4 illustrates the comparison on the out-of-sample HLP test set among Hemo PI, PSI-BLAST and m-NGSG. Here, precision, recall, accuracy and MCC scores of HemoPI are 0.55 (55%), 0.67 (66%), 0.66 (66%) and 0.31, respectively. These are lower than corresponding scores of m-NGSG-based HLP.

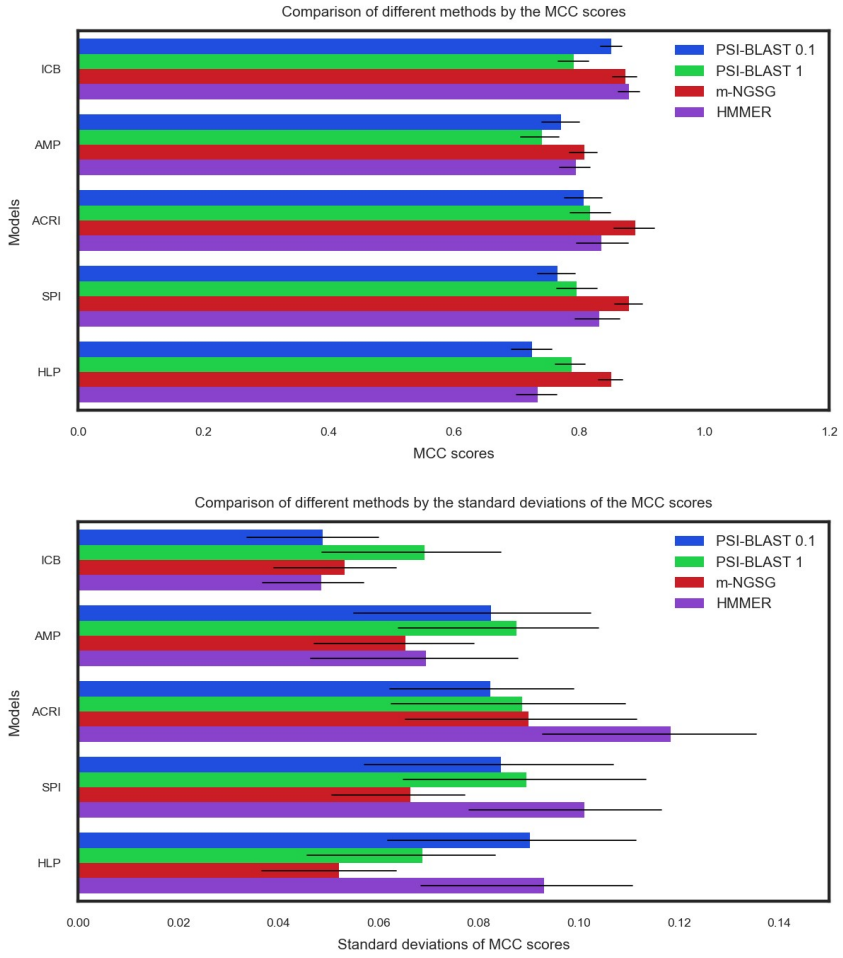

Figure 4.2: The depth of performance-consistency for each model. Figure 4.2A (upper panel) illustrates the comparison of MCC (Mathews Correlation Coefficients) among PSI-BLAST (E-value 0.1 and 1), m-NGSG and HMMER. The Y-axis indicates different function-based models; the X-axis indicates the MCC values with their standard errors. Each gray-scaled bar plot depicts the method used to build the models. Figure 4.2B (lower panel) illustrates the comparison of standard deviations of MCC (Mathews Correlation Coefficients) scores among PSI-BLAST (E-value 0.1 and 1), m-NGSG and HMMER. The Y-axis indicates different function-based models; the X-axis indicates the standard deviations of the MCC values with their standard errors. Each gray-scaled bar plot depicts the method used to build the models. Here the, higher the standard deviation, the lower the performance-consistency. The m-NGSG-based models shows standard deviations of MCC values lower than 0.05 for each model while HMMER and PSI-BLAST shows high standard deviations for a few models.Please see Supplement Figure S4.2 and S4.3 for more details.

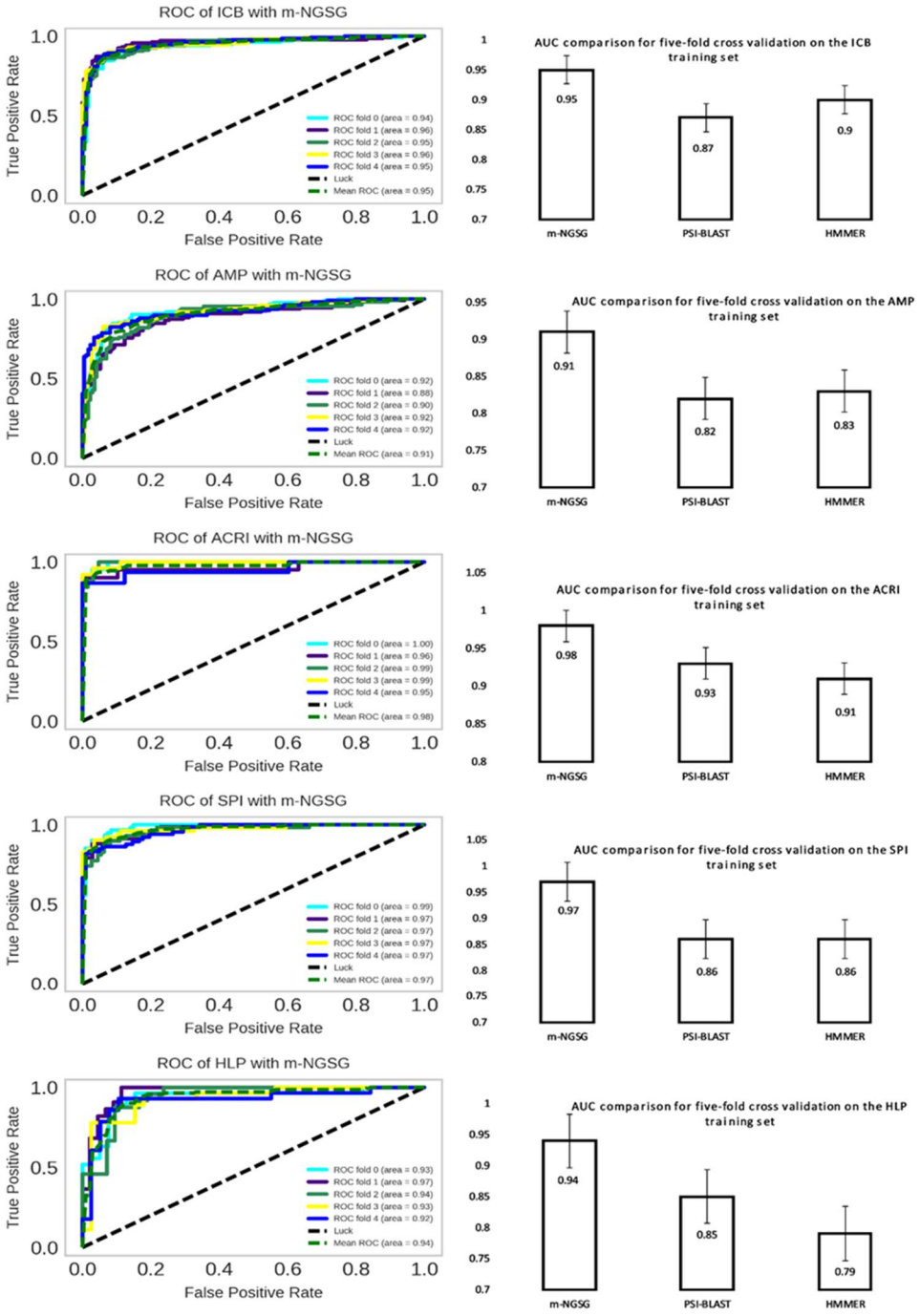

Figure 4.3**:** Description of AUC (area under the curve) among m-NGSG, HMMER and PSI-BLAST with the best MCC value. The left panel indicates the receiver operating characteristics of m-NGSG-based models. The right panel indicates the comparison of AUC among m-NGSG, PSI-BLAST and HMMER for the corresponding function-based model. The height of each bar represents the AUC for each method. m-NGSG-based models demonstrates better AUC than PSI-BLAST and HMMER.

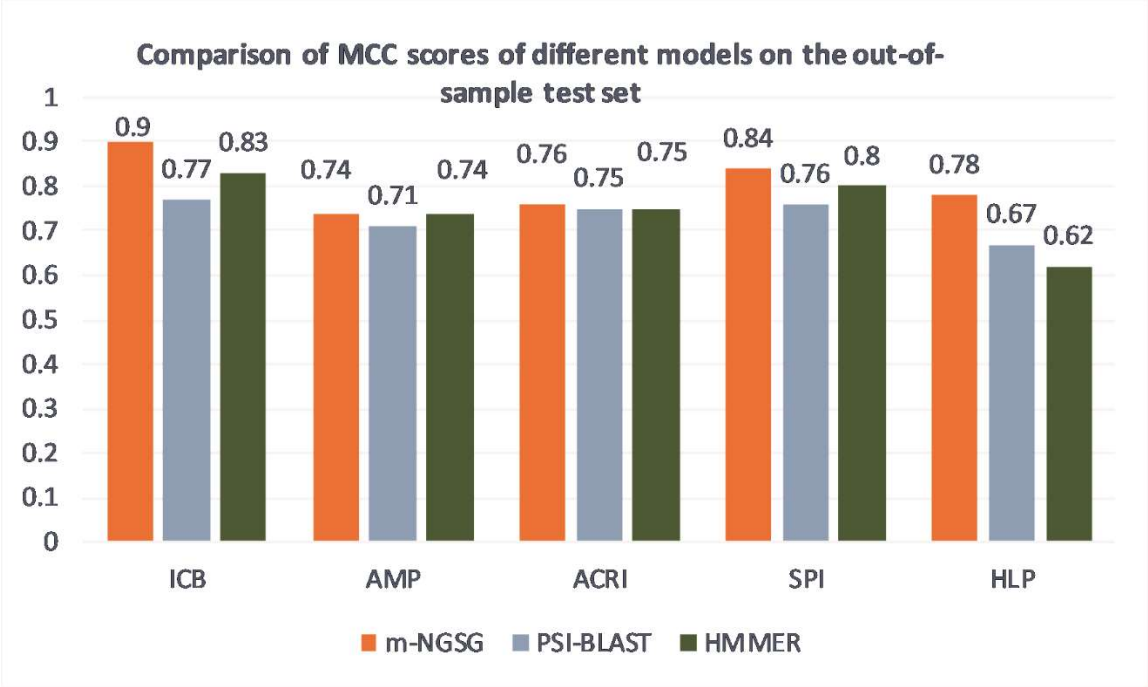

Figure 4.4: Comparison of MCC values on the out-of-sample test set with each function-based model using m-NGSG, PSI-BLAST or HMMER. While the MCC scores of HMMER are comparable for the AMP and ACRI test tests, the MCC score for ICB, SPI and HLP are noticeably lower compared to the m-NGSG-based models.

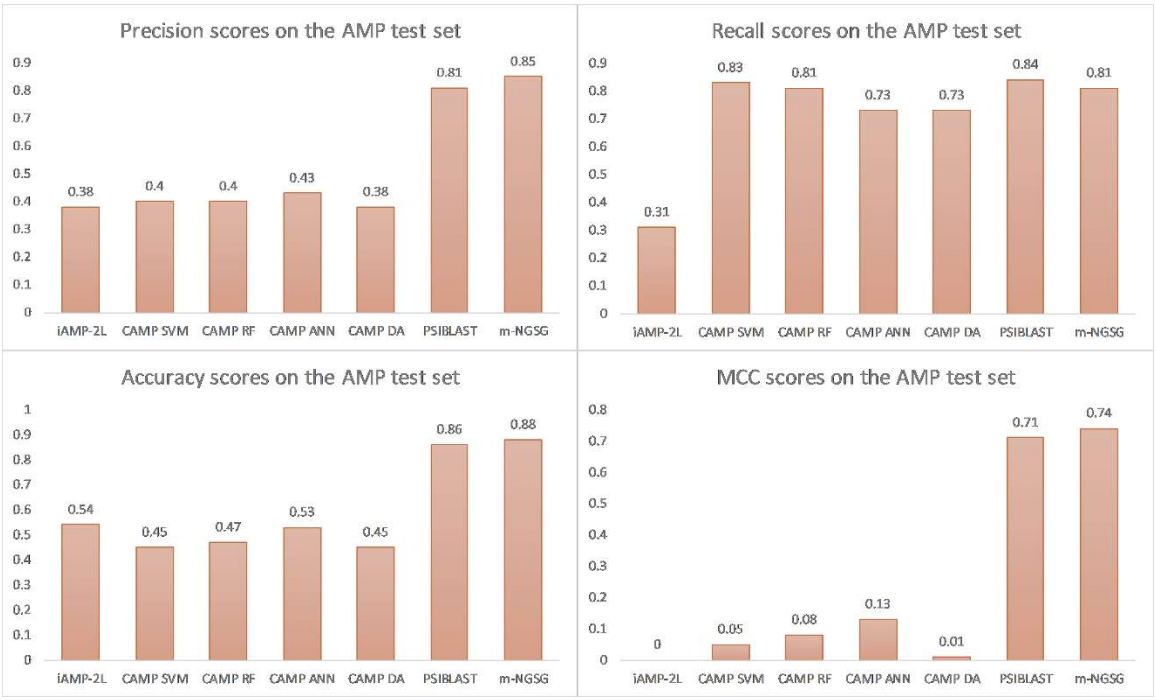

Figure 4.5: The precision, recall, accuracy and MCC values obtained applying iAMP-2L, CAMP SVM, CAMP RF, CAMP ANN, CAMP DA, PSI-BLAST E-value 0.1 and AMP (m-NGSG-based AMP model) on the out-of-sample AMP test set. Figure 4.5A illustrates that precision values of iAMP-2L and CAMP models are considerably lower than the m-NGSG based model. Figure 4.5B illustrates that the recall values of the CAMP models are comparable to the m-NGSG-based model while iAMP-2L demonstrates a noticeably lower recall value. Figures 4.5C and 5D shows considerably low MCC and accuracy values displayed by iAMP-2L and CAMP models compared to PSI-BLAST and m-NGSG.

Discussion

In this study, we constructed five different functional classifiers of cystine-stabilized peptides and combined them to build the CSPred model which predicts the probability of the selected five functional characteristics of query peptide sequences. After building a model, the most import step is to assess its performance using a k-fold cross-validation and out-of-sample test sets. We have performed this step carefully using a five-fold cross-validation and an out-of-sample test set for each of the five models. Table 4.1 shows a comprehensive comparison among evaluation matrices for each model.

The cross- validation accuracy ranged from ~86% to ~95% for the models. The accuracies on the out-of-sample test sets were concordant to the cross-validation accuracies. No big difference was detected between the accuracies for the models except the ICB model where the difference between the test and training set was ~5%. However, the increase of accuracy on the test set indicates versatility of the ICB model. The other evaluation matrices such as F1-score and MCC were also quite consistent between the training and test set (see Table 4.1). The comparative analysis between among the evaluation matrices explains the adaptability of each different function-based model.

The ultimate success and novelty of a machine learning-based model depend on its superiority than other concurrent algorithms. PSI-BLAST and HMMER are successful and widely accepted algorithms to discover distantly related protein chains. Therefore, we compared the performance of each five function-based model with PSI-BLAST and HMMER using the corresponding training and out-of-sample sets. One complexity and disadvantage to working with PSI-BLAST is choosing the optimal E-value; it is challenging to select an E-value that will give the best results. Supplement Figure S4.1 shows a clear superiority of m-NGSG-based methods over the equivalent PSI-BLAST with different E-values and HMMER except better recall values obtained by HMMER compared to m-NGSG based models. That explains only a better sensitivity of HMMER than m-NGSG models but not the overall performance. The MCC score is chosen over an accuracy score for further comparison because MCC is more robust and reflects the sensitivity, specificity, precision and false negative rate while accuracy only reflects the average of sensitivity and specificity of a model (Powers, David Martin 2011). Similar to the training set, m-NGSG-based models showed better MCC values on the out-of-sample

test sets compared to the corresponding HMMER and PSI-BLAST with the optimized E-values, Figure 4.4. This result demonstrates consistently better performance over HMMER and PSI-BLAST and unbiased behavior of the m-NGSG-based models.

In addition to PSI-BLAST and HMMER, we compared the m-NGSG-based model with other available function specific prediction models. There are two available models to predict antimicrobial peptides: iAMP-2l and $CAMP_R3$. $CAMP_R3$ also has four different classifiers to perform the antimicrobial peptide prediction which are SVM, RF, ANN, and DA. We evaluated all the models by computing the four evaluation matrices (described in Figure 4.5) on the out-of-sample test sets. The performance of iAMP-2L and $CAMP_R3$ were significantly low compared to PSI-BLAST and m-NGSG-based AMP model. The reason is possibly the training sets are not optimized to predict the cystine-stabilized AMPs. The similar results are found when we compared HemoPI (see Supplement Figure S4.4) with the m-NGSG-based HLP model. These results indicated that the m-NGSG-based models are superior to any other concurrent algorithms to classify functions of cystine-stabilized peptides.

Several cystine-stabilized peptides have already been licensed for clinical or agricultural use. This small fraction demonstrates the potential for new applications hidden among the thousands of undiscovered cystine-stabilized peptide sequences in genomes across many taxa. A voltage-gated calcium channel blocker cystine-stabilized peptide (Hv1a) from spider venom (Herzig and King 2015) is now the primary product of Vestaron, Inc., with commercial production in E. coli for broad-scale application on crops plants as an eco-friendly insecticide that degrades within two weeks after application. This same spider peptide has been fused to a targeting moiety by another group to

specifically target aphids as a transgene in plants (Bonning et al. 2013). In our own lab (CMK), antimicrobial cystine-stabilized peptides have been targeted for specific toxicity individual pathogenic bacterial species, with nontarget toxicity greatly reduced (Islam et al., unpublished data). This has implications for antibiotic treatment without the disruption of the native microbiome. Clinically, alpha-bungarotoxin has a long history of use in isolating and identifying specific acetylchloline receptors and in the diagnosis of myasthenia gravis (Dutertre, Nicke, and Tsetlin 2017). Aprotinin has been shown clinically effective against flu infection by inhibiting protease cleavage of HA0 to HA1 and HA2 (Zhirnov, Klenk, and Wright 2011). Other obvious applications for peptide-based protease inhibitors would be against HIV and HCV protease targets. The calcium channel blocker from conch, ziconotide (Prialt), is used clinically as a pain reliever (Bourinet and Zamponi 2016). The chloride channel blocker from scorpion, chlorotoxin, reached Phase III trials as a treatment for glioblastoma cancer (Cohen-Inbar and Zaaroor 2016). Linaclotide, a cystine-stabilized peptide, is licensed for clinical use orally against irritable bowel syndrome (Layer and Stanghellini 2014). Thus, a diverse array of different cystine-stabilized peptides has realized commercial application.

The present set of algorithms can be used by any researcher for data mining of genomic data in order to rapidly accrue a set of candidate cystine-stabilized peptide sequences with the desired chemical functionality. The short and simple process is detailed below in the Availability section. This analytical capability has several practical outcomes. First, such genes could be upregulated in the source organism to avoid the use of foreign transgenes, such as upregulating native antimicrobial peptides in crop plants to fight plant disease. Second, a bank of channel blockers, for example, could be

constructed from data mining a variety of genomes across several taxa in order to test an array of divergent sequences for functionality after expression of the peptides in the wet lab. Novel or more powerful activities might be found from such surveys as well as from hybrids created from this bank of sequences. Finally, some basic questions may be asked once a fuller survey of cystine-stabilized peptides is completed for a given organism. For example, Beta-amyloid peptide is an acetylcholine receptor blocker and came to prominence as the main constituent of plaque deposits symptomatic in Alzheimer's Disease (Murphy and LeVine 2010). Are there other, as yet uncharacterized, acetylchloline receptor blockers expressed from the human genome which may be associated with other diseases? It could also be asked if there is a natural pain relief functionality provided by cystine-stabilized peptides already encoded in the human genome. Existing transcriptomics data can be used to answer these questions once the peptides are properly characterized. These same sorts of questions could be asked for the many other functions associated with cystine-stabilized peptides found in other taxa as we explored the human genome for these peptide sequences.

Availability

CSPred is an open source collaborative initiative available in the bitbucket repository (https://bitbucket.org/sm_islam/cystine-stabilized-proteins/src). It is also publicly available as a free web application in watson.ecs.baylor.edu/cspred . The web server provides an accessibility to the CSPred which doesn't need any computational experience to use the model. Posting the web address (watson.ecs.baylor.edu/cspred) on a web browser will take the user to the CSPred webpage. There, the user needs to upload the fasta file of the unknown protein sequences and click the submit button. That action

will trigger the prediction process and might take a few second to display the prediction result. The result page will show six columns. The first column will be the protein ID labels of the fasta sequences. The second, third, fourth, fifth, and sixth columns will display the probability value of being an ICB, AMP, ACRI, SRI, and HLP, respectively, for each sequence submitted. Thus, the web interface provides a simple avenue to categorize submitted protein sequences according to these five functional characteristics, and does so utilizing a high-throughput, batch-style input. The sub classifiers of ICB are not included in CSPred. However, all the training and test datasets are provided as a supplement so that users can also make their own models or reproduce the same models using the m-NGSG framework that is available at watson.ecs.baylor.edu/ngsg.

Acknowledgment

We acknowledge Benjamin J. Heil for making our web-application publicly available through http://watson.ecs.baylor.edu/cspred.

References

Adams, Michael E. 2004. "Agatoxins: Ion Channel Specific Toxins from the American Funnel Web Spider, Agelenopsis Aperta." *Toxicon* 43 (5): 509–25. https://doi.org/10.1016/j.toxicon.2004.02.004.

Akondi, Kalyana B., Markus Muttenthaler, Sébastien Dutertre, Quentin Kaas, David J. Craik, Richard J. Lewis, and Paul F. Alewood. 2014. "Discovery, Synthesis, and Structure–Activity Relationships of Conotoxins." *Chemical Reviews* 114 (11): 5815–47. https://doi.org/10.1021/cr400401e.

Altschul, S. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25 (17): 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

Boeckmann, B. 2003. "The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL in 2003." *Nucleic Acids Research* 31 (1): 365–70. https://doi.org/10.1093/nar/gkg095.

Bonning, Bryony C, Narinder Pal, Sijun Liu, Zhaohui Wang, S Sivakumar, Philip M Dixon, Glenn F King, and W Allen Miller. 2013. “Toxin Delivery by the Coat Protein of an Aphid-Vectored Plant Virus Provides Plant Resistance to Aphids.” *Nature Biotechnology* 32 (1): 102–5. https://doi.org/10.1038/nbt.2753.

Bourinet, Emmanuel, and Gerald W. Zamponi. 2016. “Block of Voltage-Gated Calcium Channels by Peptide Toxins.” *Neuropharmacology*, October. https://doi.org/10.1016/j.neuropharm.2016.10.016.

Chan, Yau Sang, Randy Chi Fai Cheung, Lixin Xia, Jack Ho Wong, Tzi Bun Ng, and Wai Yee Chan. 2016. “Snake Venom Toxins: Toxicity and Medicinal Applications.” *Applied Microbiology and Biotechnology* 100 (14): 6165–81. https://doi.org/10.1007/s00253-016-7610-9.

Chang, Darby, Yu-Tang Syu, and Po-Chang Lin. 2010. “Predicting the Protein-Protein Interactions Using Primary Structures with Predicted Protein Surface.” *BMC Bioinformatics* 11 (Suppl 1): S3. https://doi.org/10.1186/1471-2105-11-S1-S3.

Chaudhary, Kumardeep, Ritesh Kumar, Sandeep Singh, Abhishek Tuknait, Ankur Gautam, Deepika Mathur, Priya Anand, Grish C. Varshney, and Gajendra P. S. Raghava. 2016. “A Web Server and Mobile App for Computing Hemolytic Potency of Peptides.” *Scientific Reports* 6 (1). https://doi.org/10.1038/srep22843.

Cheek, Sara, S. Sri Krishna, and Nick V. Grishin. 2006. “Structural Classification of Small, Disulfide-Rich Protein Domains.” *Journal of Molecular Biology* 359 (1): 215–37. https://doi.org/10.1016/j.jmb.2006.03.017.

Cohen-Inbar, Or, and Menashe Zaaroor. 2016. “Glioblastoma Multiforme Targeted Therapy: The Chlorotoxin Story.” *Journal of Clinical Neuroscience* 33 (November): 52–58. https://doi.org/10.1016/j.jocn.2016.04.012.

Craik, David J., Shane Simonsen, and Norelle L. Daly. 2002. “The Cyclotides: Novel Macrocyclic Peptides as Scaffolds in Drug Design.” *Current Opinion in Drug Discovery & Development* 5 (2): 251–60.

Ding, Hui, En-Ze Deng, Lu-Feng Yuan, Li Liu, Hao Lin, Wei Chen, and Kuo-Chen Chou. 2014. “ICTX-Type: A Sequence-Based Predictor for Identifying the Types of Conotoxins in Targeting Ion Channels.” *BioMed Research International* 2014: 1–10. https://doi.org/10.1155/2014/286419.

Du, Pufeng, Shengjiao Cao, and Yanda Li. 2009. “SubChlo: Predicting Protein Subchloroplast Locations with Pseudo-Amino Acid Composition and the Evidence-Theoretic K-Nearest Neighbor (ET-KNN) Algorithm.” *Journal of Theoretical Biology* 261 (2): 330–35. https://doi.org/10.1016/j.jtbi.2009.08.004.

Dubchak, I., I. Muchnik, S. R. Holbrook, and S. H. Kim. 1995. “Prediction of Protein Folding Class Using Global Description of Amino Acid Sequence.” *Proceedings of the National Academy of Sciences of the United States of America* 92 (19): 8700–8704.

Dutertre, Sébastien, Annette Nicke, and Victor I. Tsetlin. 2017. “Nicotinic Acetylcholine Receptor Inhibitors Derived from Snake and Snail Venoms.” *Neuropharmacology*, June. https://doi.org/10.1016/j.neuropharm.2017.06.011.

Fan, Yong-Xian, Jiangning Song, Xiangzeng Kong, and Hong-Bin Shen. 2011. “PredCSF: An Integrated Feature-Based Approach for Predicting Conotoxin Superfamily.” *Protein & Peptide Letters* 18 (3): 261–67. https://doi.org/10.2174/092986611794578341.

Finn, R. D., J. Clements, and S. R. Eddy. 2011. “HMMER Web Server: Interactive Sequence Similarity Searching.” *Nucleic Acids Research* 39 (suppl): W29–37. https://doi.org/10.1093/nar/gkr367.

Frazão, Bárbara, Vitor Vasconcelos, and Agostinho Antunes. 2012. “Sea Anemone (Cnidaria, Anthozoa, Actiniaria) Toxins: An Overview.” *Marine Drugs* 10 (12): 1812–51. https://doi.org/10.3390/md10081812.

Gelly, J.-C. 2004. “The KNOTTIN Website and Database: A New Information System Dedicated to the Knottin Scaffold.” *Nucleic Acids Research* 32 (90001): 156D–159. https://doi.org/10.1093/nar/gkh015.

Gracy, J., D. Le-Nguyen, J.-C. Gelly, Q. Kaas, A. Heitz, and L. Chiche. 2007. “KNOTTIN: The Knottin or Inhibitor Cystine Knot Scaffold in 2007.” *Nucleic Acids Research* 36 (Database): D314–19. https://doi.org/10.1093/nar/gkm939.

Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. 2006. “A Closer Look at Skip-Gram Modelling.” In *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, 1–4.

Herzig, Volker, and Glenn King. 2015. “The Cystine Knot Is Responsible for the Exceptional Stability of the Insecticidal Spider Toxin ω-Hexatoxin-Hv1a.” *Toxins* 7 (10): 4366–80. https://doi.org/10.3390/toxins7104366.

Huang, Ying, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li. 2010. “CD-HIT Suite: A Web Server for Clustering and Comparing Biological Sequences.” *Bioinformatics* 26 (5): 680–82. https://doi.org/10.1093/bioinformatics/btq003.

Islam, S M Ashiqul, Benjamin J Heil, Christopher Michel Kearney, and Erich J Baker. 2017. “Protein Classification Using Modified N-Grams and Skip-Grams.” *Bioinformatics*, December. https://doi.org/10.1093/bioinformatics/btx823.

Islam, S. M. Ashiqul, Tanvir Sajed, Christopher Michel Kearney, and Erich J. Baker. 2015. “PredSTP: A Highly Accurate SVM Based Model to Predict Sequential Cystine Stabilized Peptides.” *BMC Bioinformatics* 16 (July): 210. https://doi.org/10.1186/s12859-015-0633-x.

Islam, SM Ashiqul, Christopher Michel Kearney, Ankan Choudhury, and Erich J. Baker. 2017. “Protein Classification Using Modified *N-Gram* and *Skip-Gram* Models: Extended Abstract.” In , 586–586. ACM Press. https://doi.org/10.1145/3107411.3108193.

Keˇselj, Vlado, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. “N-GRAM-BASED AUTHOR PROFILES FOR AUTHORSHIP ATTRIBUTION.” In *Pacific Association for Computational Linguistics*.

Kedarisetti, Pradyumna, Marcin J. Mizianty, Quentin Kaas, David J. Craik, and Lukasz Kurgan. 2014. “Prediction and Characterization of Cyclic Proteins from Sequences in Three Domains of Life.” *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1844 (1): 181–90. https://doi.org/10.1016/j.bbapap.2013.05.002.

King, Glenn F., and Margaret C. Hardy. 2013. “Spider-Venom Peptides: Structure, Pharmacology, and Potential for Control of Insect Pests.” *Annual Review of Entomology* 58 (1): 475–96. https://doi.org/10.1146/annurev-ento-120811-153650.

Kuzmenkov, A. I., E. V. Grishin, and A. A. Vassilevski. 2015. “Diversity of Potassium Channel Ligands: Focus on Scorpion Toxins.” *Biochemistry (Moscow)* 80 (13): 1764–99. https://doi.org/10.1134/S0006297915130118.

Layer, P., and V. Stanghellini. 2014. “Review Article: Linaclotide for the Management of Irritable Bowel Syndrome with Constipation.” *Alimentary Pharmacology & Therapeutics* 39 (4): 371–84. https://doi.org/10.1111/apt.12604.

Mobli, Mehdi, Eivind A. B. Undheim, and Lachlan D. Rash. 2017. “Modulation of Ion Channels by Cysteine-Rich Peptides: From Sequence to Structure.” *Advances in Pharmacology (San Diego, Calif.)* 79: 199–223. https://doi.org/10.1016/bs.apha.2017.03.001.

Mourão, Caroline, and Elisabeth Schwartz. 2013. “Protease Inhibitors from Marine Venomous Animals and Their Counterparts in Terrestrial Venomous Animals.” *Marine Drugs* 11 (6): 2069–2112. https://doi.org/10.3390/md11062069.

Munasinghe, Nehan, and MacDonald Christie. 2015. “Conotoxins That Could Provide Analgesia through Voltage Gated Sodium Channel Inhibition.” *Toxins* 7 (12): 5386–5407. https://doi.org/10.3390/toxins7124890.

Murphy, M. Paul, and Harry LeVine. 2010. “Alzheimer’s Disease and the Amyloid-β Peptide.” Edited by Mark A. Lovell. *Journal of Alzheimer’s Disease* 19 (1): 311–23. https://doi.org/10.3233/JAD-2010-1221.

Nguyen, Leonard T., Evan F. Haney, and Hans J. Vogel. 2011. “The Expanding Scope of Antimicrobial Peptide Structures and Their Modes of Action.” *Trends in Biotechnology* 29 (9): 464–72. https://doi.org/10.1016/j.tibtech.2011.05.001.

Norton, Raymond S., and K. George Chandy. 2017. “Venom-Derived Peptide Inhibitors of Voltage-Gated Potassium Channels.” *Neuropharmacology*, July. https://doi.org/10.1016/j.neuropharm.2017.07.002.

Olivera, B. M., W. R. Gray, R. Zeikus, J. M. McIntosh, J. Varga, J. Rivier, V. de Santos, and L. J. Cruz. 1985. “Peptide Neurotoxins from Fish-Hunting Cone Snails.” *Science (New York, N.Y.)* 230 (4732): 1338–43.

Ortiz, Ernesto, Georgina B. Gurrola, Elisabeth Ferroni Schwartz, and Lourival D. Possani. 2015. “Scorpion Venom Components as Potential Candidates for Drug Development.” *Toxicon* 93 (January): 125–35. https://doi.org/10.1016/j.toxicon.2014.11.233.

Possani, Lourival D., Baltazar Becerril, Muriel Delepierre, and Jan Tytgat. 1999. “Scorpion Toxins Specific for Na+-Channels.” *European Journal of Biochemistry* 264 (2): 287–300. https://doi.org/10.1046/j.1432-1327.1999.00625.x.

Powers, David Martin. 2011. “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation.” *Journal of Machine Learning Technologies* 2 (1): 37–63.

Santibáñez-López, Carlos E., and Lourival D. Possani. 2015. “Overview of the Knottin Scorpion Toxin-like Peptides in Scorpion Venoms: Insights on Their Classification and Evolution.” *Toxicon* 107 (December): 317–26. https://doi.org/10.1016/j.toxicon.2015.06.029.

Sharma, Arun, Pallavi Kapoor, Ankur Gautam, Kumardeep Chaudhary, Rahul Kumar, Jagat Singh Chauhan, Atul Tyagi, and Gajendra P. S. Raghava. 2013. “Computational Approach for Designing Tumor Homing Peptides.” *Scientific Reports* 3: 1607. https://doi.org/10.1038/srep01607.

Shen, Hong-Bin, and Kuo-Chen Chou. 2008. “PseAAC: A Flexible Web Server for Generating Various Kinds of Protein Pseudo Amino Acid Composition.” *Analytical Biochemistry* 373 (2): 386–88. https://doi.org/10.1016/j.ab.2007.10.012.

Silverstein, Kevin A.T., William A. Moskal, Hank C. Wu, Beverly A. Underwood, Michelle A. Graham, Christopher D. Town, and Kathryn A. VandenBosch. 2007. "Small Cysteine-Rich Peptides Resembling Antimicrobial Peptides Have Been under-Predicted in Plants: *Under-Predicted Cysteine-Rich Peptides in Plants*." *The Plant Journal* 51 (2): 262–80. https://doi.org/10.1111/j.1365-313X.2007.03136.x.

Simeon, Saw, Watshara Shoombuatong, Nuttapat Anuwongcharoen, Likit Preeyanon, Virapong Prachayasittikul, Jarl E. S. Wikberg, and Chanin Nantasenamat. 2016. "OsFP: A Web Server for Predicting the Oligomeric States of Fluorescent Proteins." *Journal of Cheminformatics* 8: 72. https://doi.org/10.1186/s13321-016-0185-8.

Waghu, Faiza Hanif, Ram Shankar Barai, Pratima Gurung, and Susan Idicula-Thomas. 2016. "CAMP $_{R3}$: A Database on Sequences, Structures and Signatures of Antimicrobial Peptides: Table 1." *Nucleic Acids Research* 44 (D1): D1094–97. https://doi.org/10.1093/nar/gkv1051.

Xia, Jun-Feng, Kyungsook Han, and De-Shuang Huang. 2010. "Sequence-Based Prediction of Protein-Protein Interactions by Means of Rotation Forest and Autocorrelation Descriptor." *Protein & Peptide Letters* 17 (1): 137–45. https://doi.org/10.2174/092986610789909403.

Xiao, Xuan, Pu Wang, Wei-Zhong Lin, Jian-Hua Jia, and Kuo-Chen Chou. 2013. "IAMP-2L: A Two-Level Multi-Label Classifier for Identifying Antimicrobial Peptides and Their Functional Types." *Analytical Biochemistry* 436 (2): 168–77. https://doi.org/10.1016/j.ab.2013.01.019.

Zhang, Guang-Ya, and Bai-Shan Fang. 2008. "Predicting the Cofactors of Oxidoreductases Based on Amino Acid Composition Distribution and Chou's Amphiphilic Pseudo-Amino Acid Composition." *Journal of Theoretical Biology* 253 (2): 310–15. https://doi.org/10.1016/j.jtbi.2008.03.015.

Zhirnov, O.P., H.D. Klenk, and P.F. Wright. 2011. "Aprotinin and Similar Protease Inhibitors as Drugs against Influenza." *Antiviral Research* 92 (1): 27–36. https://doi.org/10.1016/j.antiviral.2011.07.014.

Table S4.1. Description of dataset for each of the five selected classes of cystine-stabilized peptide, their sequence similarity and number of chains training and test sets.

Dataset		UniProtKB (protein knowledgebase) search key	Sequence identity<	No. of chains	No.of chains In the training set	No. of chains in the out of sample test set
ICB	Positive	channel toxin annotation:(type:disulfid) length:[1 TO 150] NOT annotation:(type:function acetylcholine) NOT hormon (sodium OR nav OR potassium OR kv OR calcium OR cav OR ion) AND reviewed:yes	70%	697	627	70
	Negative	NOT nav NOT cav NOT kv NOT sodium NOT potassium NOT calcium NOT ion (defensin OR acetylecholine OR serine OR hormon OR enzyme OR antimicrobial OR bateria) length:[1 TO 150] annotation:(type:disulfid) AND reviewed:yes	70%	1004	903	101
AMP	Positive*	(defensin OR antimicrobial OR antibacterial OR antifungal OR antiviral) NOT ”defensin-like” length:[1 TO 150] annotation:(type:disulfid) AND reviewed:yes	70%	611	551	60
	Negative	NOT defensin NOT antimicrobial NOT antibacterial NOT antifungal NOT antiviral NOT defense (channel OR ion) length:[1 TO 150] annotation:(type:disulfid) AND reviewed:yes	70%	945	850	95
ACRI	Positive	(acetylcholine OR achr OR nachr) channel NOT nav NOT kv NOT cav NOT sodium NOT potassium NOT calcium length:[1 TO 150] annotation:(type:disulfid)	50%	105	94	11
	Negative	channel toxin (defensin OR nav OR kv OR cav OR sodium OR potassium OR calcium OR agouti OR serine OR enzyme) NOT acetylcholine NOT achr NOT nachr NOT	90%	381	342	39

		snake length:[1 TO 150] annotation:(type:disulfid)				
SPI	Positive	serine protease inhibitor annotation:(type:disulfid) length:[1 TO 150] NOT defensin NOT antimicrobial NOT antifungal NOT acetylcholine NOT nav NOT kv NOT cav NOT sodium NOT potassium NOT calcium NOT channel AND reviewed:yes	90%	328	295	33
	Negative	NOT serine NOT protease annotation:(type:disulfid) length:[1 TO 150] (defensin OR antimicrobial OR antifungal OR acetylcholine OR nav OR kv OR cav OR sodium OR potassium OR calcium OR channel) AND reviewed:yes	40%	604	543	61
HLP	Positive*	(hemolytic OR cytolytic OR cytoxic) annotation:(type:disulfid) length:[1 TO 150] annotation:(type:function hemolytic) NOT annotation:(type:function weak) NOT ”no hemolytic”	90%	146	131	15
	Negative*	annotation:(type:disulfid) length:[1 TO 150] NOT annotation:(type:function hemolytic) NOT hemolytic annotation:(type:function toxin) NOT mammalia NOT mammalian NOT cytotoxic NOT cytolytic NOT snake NOT name:”alpha mammalia” NOT name:”beta mammalia” AND reviewed:yes + (”no hemolytic” OR ”non hemolytic” OR ”not have hemolytic” OR ”lack hemolytic”) annotation:(type:disulfid) length:[1 TO 150] AND reviewed:yes	50%	222	199	23

Abbreviations : ICB = Ion channel blocker; AMP = Antimicrobial peptide; ACRI = Acetylcholine receptor inhibitor; SPI= Serine protease inhibitor; HLP = Hemolytic protein

* The AMP positive training set was manually curated after performing the sequence identity-based clustering step. The positive and negative training set of HLP were collected using composite search followed by combining two of more datasets. Also, the positive training set of HLP was manually curated to filter noise.

Table S4.2. Description of dataset for each subclass within ICB, their sequence similarity and number of chains training and test sets. NaB, KB and CaB represents the sodium, potassium and calcium channel blocker classifiers, respectively.

Dataset		Sources of Datasets	Sequence identity<	No. of chains	No.of Chains In the training set	No.of chains in the out of sample test set
NaB	Positive	Annotated as sodium channel blockers (Nab) but not potassium (Kb) or calcium (Cab) channel blockers in the Uniprot database	Nab 90%	697	341	38
	Negative	Annotated potassium (Kb) or calcium channel blockers (Cab) but not sodium channel blockers (Nab) in the Uniprot database	Kb 65% and Cab 90%	1004	271	31
KB	Positive	Annotated as potassium channel blockers (Kb) but not sodium (Nab) or calcium channel blockers (Cab) in the Uniprot database	Kab 00%	611	257	29
	Negative	Annotated sodium (Nab) or calcium channel blockers (Cab) but not potassium channel blockers (Kb) in the Uniprot database	Nab 65% and Cab 90%	945	289	33
CaB	Positive	Annotated as calcium channel blockers (Cab) but not sodium (Nab) or potassium channel blockers (Kb) in the Uniprot database	Cab 95%	105	132	15
	Negative	Annotated sodium (Nab) or potassium channel blockers (Kb) but not calcium channel blockers (Cab) in the Uniprot database	Nab 65% and Kab 65%	381	150	17

Table S4.3. Selected m-NGSG parameters for each model. Description of each parameter is discussed in detailed in *Islam et al*, 2017(S. M. A. Islam et al. 2017).

Training set		n	k	np	kp	y	c	
ICB		7	13	1	7	10	4	
AMP		4	19	1	1	3	1	
ACRI		3	1	1	1	2	1	
SPI		3	10	1	7	5	1	
HLP		1	1	1	1	3	1	
NaB		1	6	1	1	5	1	
KB		4	16	1	1	5	1	
CaB		1	3	1	1	5	1	

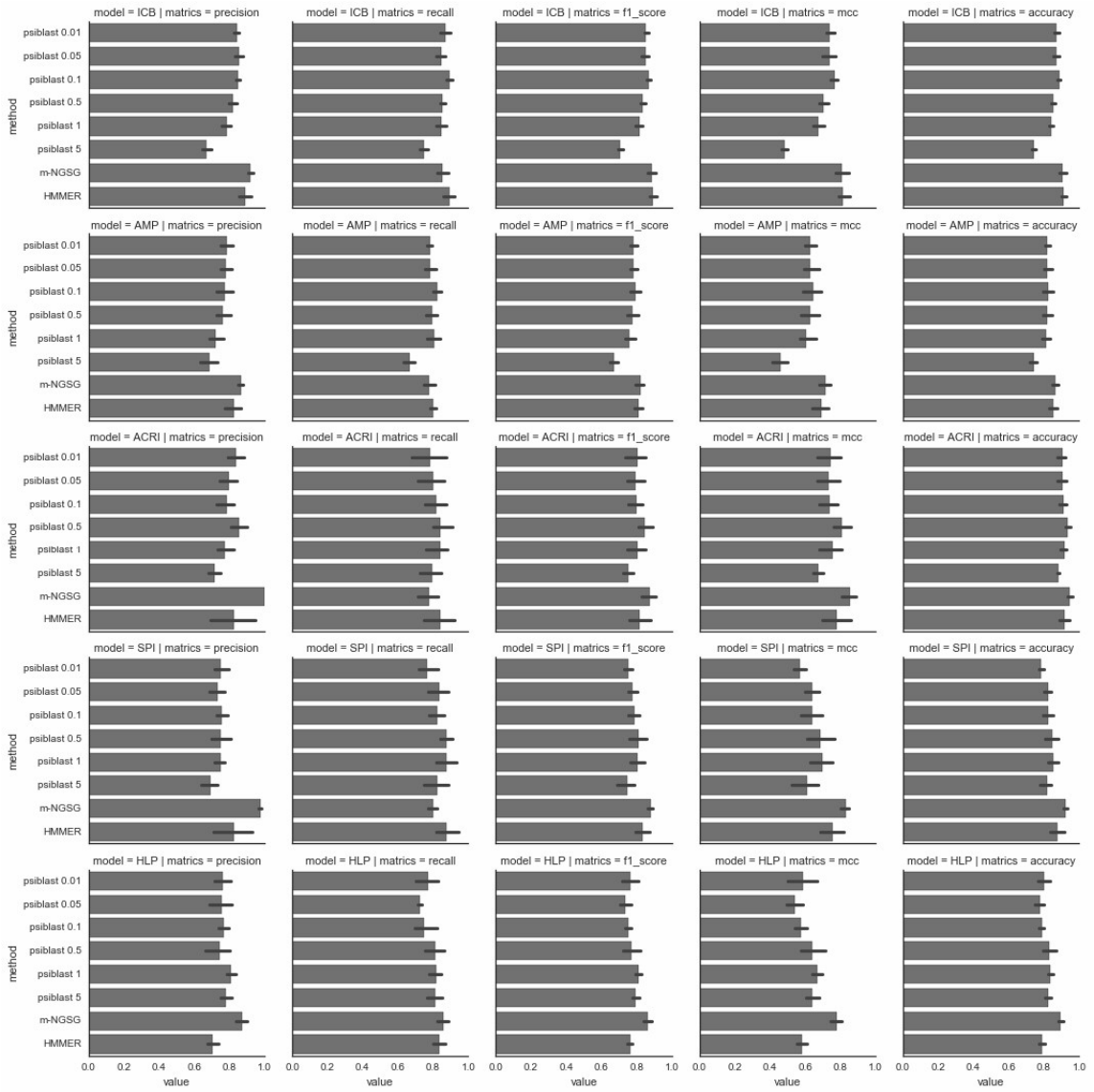

Figure S4.1. Performance comparison of each classification method for different functional-based models using five-fold cross-validation. Each row in the facet grid plot represents the function-based models while each column stands for an evaluation matric. Y-axis of each bar plot show the different methods used to build a model and X-axis shows the values of the evaluation matrices witherror bars. Except ICB dataset, m-NGSG-based models shows better F1-scores, MCC and accuracy values for all other four datasets, while PSI-BLAST shows lower values for all the data sets.

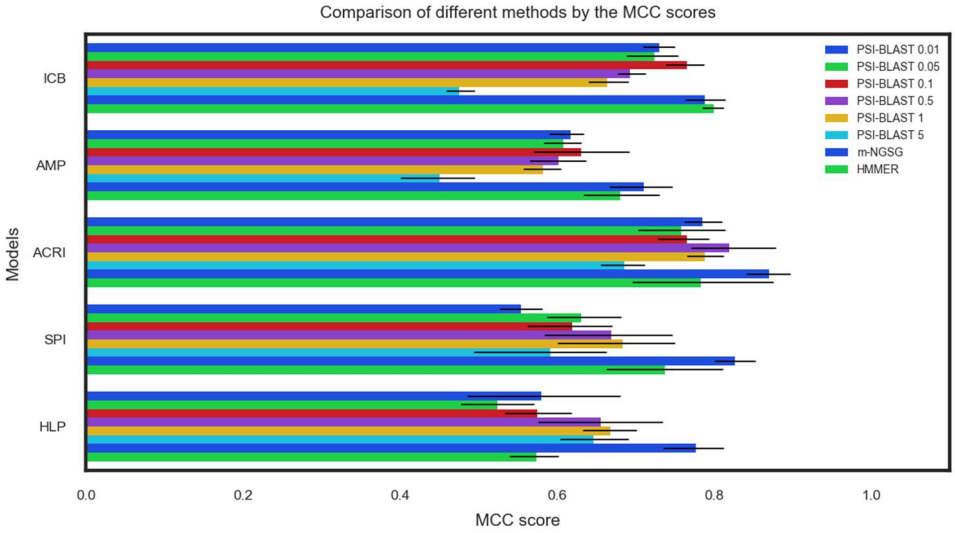

Figure S4.2. Illustrates the comparison of MCC (Mathews Correlation Coefficients) among PSI-BLAST (E-value 0.01, 0.05, 0.1, 0.5, 1and 5), m-NGSG and HMMER. Y-axis shows different function-based models, X-axis shows the MCC scores with their standard errors and each colored (in gray scale) bar plot shows the method used to build the models. Except ICB dataset, m-NGSG-based models shows better MCC scores for all other four datasets, while PSI-BLAST shows lower values for all the data sets.

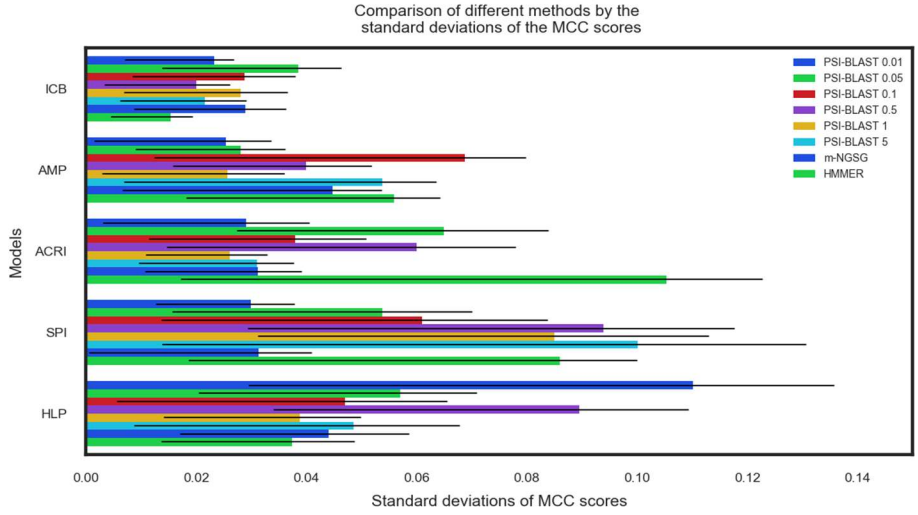

Figure S4.3. Illustrates the comparison of standard deviations of MCC (Mathews Correlation Coefficients) scores among PSI-BLAST (E-value 0.01, 0.05, 0.1, 0.5, 1and 5), m-NGSG and HMMER. Y-axis shows different function-based models, X-axis shows the standard deviations of the MCC scrores with their standard errors and each colored (in gray scale) bar plot shows the method used to build the models. This figure shows the depth of performance-consistency of each model. Higher the standard deviation, lower the performance-consistency. The m-NGSG-based models shows standard deviations of MCC scores lower than 0.05 for each model while HMMER and PSI-BLAST shows high standard deviations for a few models.

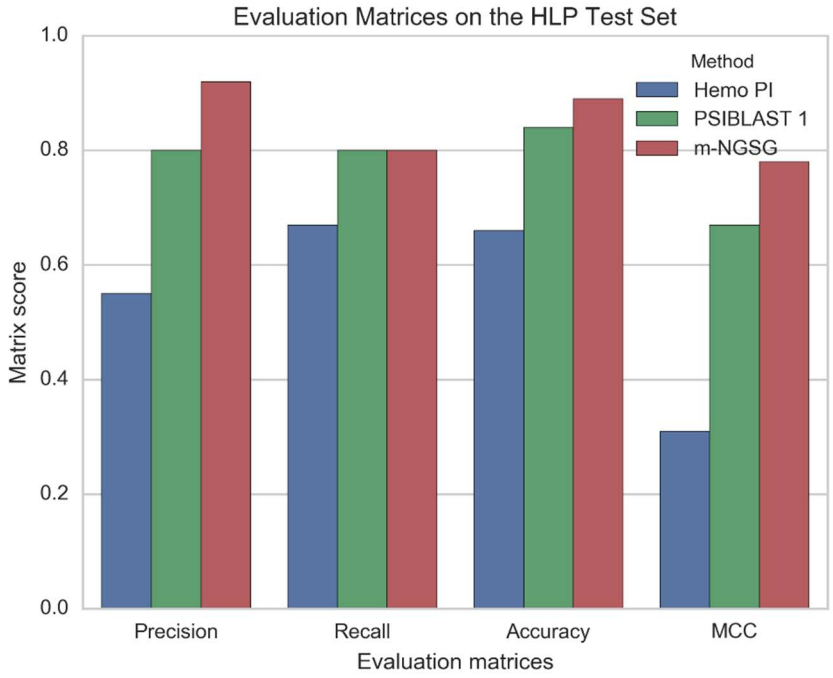

Figure S4.4. The precision, recall, accuracy and MCC values obtained applying each method on the out of sample HLP test set. The m-NGSG-based HLP model performed better than the Hemo PI in respect to each of evaluation matrices (Precision, Recall, Accuracy and MCC).

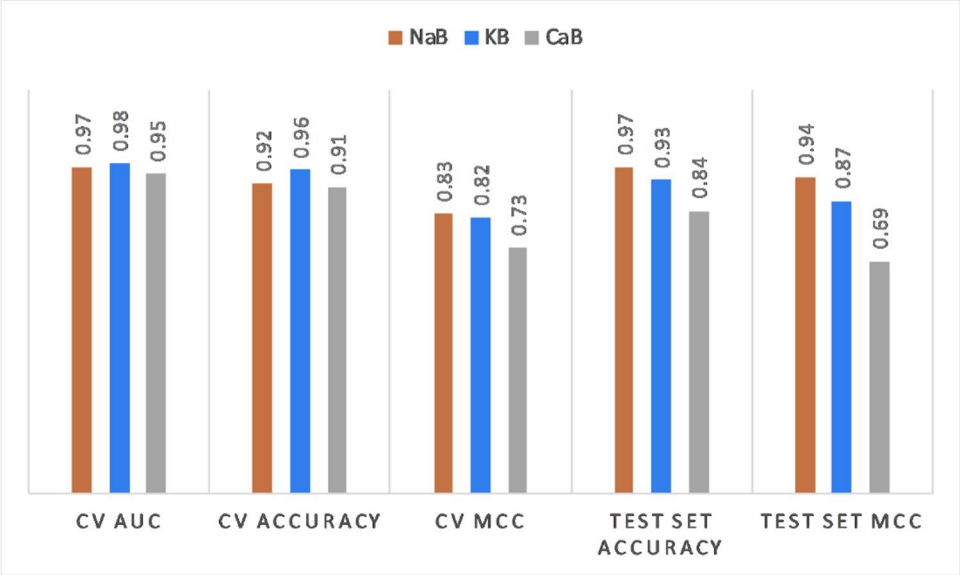

Figure S4.5. Illustrates the performance of the sub-classifiers of Ion Channel Blocker (ICB). NaB, KB and CaB represents the sodium, potassium and calcium channel blocker classifiers, respectively. CV AUC indicates the area under curve (AUC) using five-fold cross-validation; CV ACCURACY indicates the accuracy using five-fold cross-validation; CV MCC indicates the Mathews Correlation Coefficient (MCC) values using five-fold cross-validation; TEST SET ACCURACY indicates the accuracy using the out of sample test set; TEST SET MCC indicates the MCC values using the out of sample test set.

CHAPTER FIVE

The Use of a Virus-Derived Targeting Peptide to Selectively Kill Staphylococcus Bacteria with Antimicrobial Peptides

Abstract

Traditional antibiotics destroy the microbiome broadly, with consequent unintended impacts on long-term health. Targeted therapies seek to selectively eliminate a pathogen without disrupting the microbiome community. Here, we target to fungal antimicrobial peptides (AMPs), plectasin and eurocin, by genetically fusing their coding sequence to that of the host-binding protein of bacteriophage A12C, which selectively infects Staphylococcus. Surprisingly, we noted that targeting brought no change in the toxicity of the AMP when applied to two different staphylococci, S. aureus and S. epidermidis, but found a drastic decrease in toxicity against the negative controls, Enterococcus faecalis and Bacillus subtilis. Thus, the differential selectivity in this case is a loss of toxicity against the nontarget species rather than the gain of toxicity against the target species which was reported in previous studies with other types of targeting peptides. This is the first report of the use of virus-derived peptide sequences to target antimicrobial peptides. Considering the very large databank of bacteriophages and their bacterial hosts, this targeting approach should be generally applicable to a wide range of bacterial pathogens.

Introduction

Small molecule antibiotics are the standard treatment against bacterial infections, but they have three key deficits. First, antibiotics have the long discovery and development cycles typical of small molecule drugs(Charles and Grayson 2004; Norrby, Nord, and Finch 2005). Second, the broad-spectrum nature of antibiotics disrupts the gut microbiome and can lead to the rise of opportunistic pathogens (Enright et al. 2002; Thung et al. 2016). Finally, resistance against antibiotics is increasing as bacterial populations under selection pressure develop effective antibiotic-binding proteins, efflux pumps and degradative enzymes (Soto 2013). Antimicrobial peptides (AMPs) are a well-studied antibiotic alternative that can address these deficits.

The first problem with antibiotics, that of the long discovery cycle, is addressed by the sheer ubiquity of AMPs in nature. AMPs are found across bacterial, animal and plant taxa and function against bacterial, viral and/or fungal targets(Hancock and Sahl 2006). To accelerate access to these natural AMPs, our group has developed algorithms for discovering AMP ORFs from genomic data. First, we have developed an SVM-based algorithm model (Islam et al. 2015) to identify ORFs corresponding to the sequential tri-disulfide peptide (STP) structure that is typical of the larger, highly stable AMPs. Second, we have developed natural language processing-based algorithms for determining protein function (Islam et al. 2017), allowing for the screening of functional AMPs across many taxa. Once these sequences are discovered, they can be recombinantly expressed in bacterial (Y. Li 2011; C. Li et al. 2010), fungal (de Bruin et al. 2005; Cregg et al. 2009) or plant (Huafang Lai and Jake Stahnke 2013; Nadal et al. 2012) bio factories for

function confirmation and mass production, greatly speeding up the process of drug development.

The second problem of antibiotics, that of the disruption of the greater microbiome by broad spectrum activity, can be resolved by peptide targeting. Targeting has gained ascendance in cancer therapy research and studies centered around directing drug activity, including RNAi, CRISPR Cas9 and gene therapy methodologies. Targeting can be accomplished using virus delivery or by attaching small peptide targeting moieties such as pheromones and antibody fragments (e.g., scFv) (Peschen et al. 2004; B. Wang et al. 1999). There are a limited number of examples of targeting applied to AMPs. A transgene coding for an scFv targeting domain fused to an AMP resulted in a transgenic plant resistant to pathogenic fungi (Peschen et al. 2004). As a drug-based example, targeting moieties based on pheromones conjugated with synthetic AMPs provided specific inhibition of Streptococcus mutans, a dental carries agent (Eckert et al. 2006). Quorum-sensing peptide conjugates like ArgD with plectasin (a fungal AMP) were developed against methicillin resistant S. aureus (Mao et al. 2013). It intrigued us that viral-guided targeting, with potentially universal application against bacteria and fungi, has not yet been used with AMPs ((Parachin et al. 2012).

The third problem of antibiotics, that of the development of pathogen strains resistant to the antibiotic, can potentially be solved by the use of AMPs. Resistance against AMPs is rare and is slow to develop in pathogens (Maróti et al. 2011). Cationic AMPs usually target the fundamental property of the negatively charged nature of the bacterial cell outer membrane, and combined with the hydrophobic regions of the AMP,

which directly interact with the bacterial membrane(Nguyen, Haney, and Vogel 2011; G. Wang et al. 2015).

In this study, we demonstrate a high level of production of the cationic AMPs plectasin (Mygind et al. 2005) and eurocin (Oeemig et al. 2012) targeted by fusion to bacteriophage A12C coat protein display peptide with specificity for Streptococcus aureus (Yacoby et al. 2006). The SUMO E. coli expression vector was used (J. F. Li et al. 2009). This is the first reported use of viral-based targeting with AMPs. Interestingly, the targeting domain does not enhance AMP toxicity towards the target bacterial species, but instead operates by strongly decreasing toxicity against non-target bacteria.

Materials and Methods

Reagents

E. coli (BL21 and 10β) strains were purchased from New England Biolabs. The pE-SUMOstar vector used for *E. coli* expression was purchased from LifeSensors. The Ulp1 protease was expressed in *E. coli* using pFGET19_Ulp1 plasmid purchased from Addgene. The gBlock containing *E. coli*-codon optimized sequences of plectasin, eurocin, and the A12C fusion peptide were purchased from IDT. The strains of bacteria used for antimicrobial assay were obtained from S. J. Kim, Department of Chemistry and Biochemistry, Baylor University, and the Microbiology Laboratory, Department of Biology, Baylor University (See Table 5.1.)

Construction and Cloning of Plasmid

After digestion, the synthesized genes (Integrated DNA Technologies) were cloned into the pE-SUMOstar vector following the SUMO protease cleavage site (Figure 1). The

recombinant plasmids were electroporated into *E. coli* 10β cells and positively transformed colonies were selected with kanamycin and screened via PCR. The prepared plasmids were extracted and transformed into chemically competent BL21 cells for expression (Pope and Kent 1996).

Expression, Extraction and Purification of Proteins

Positive BL21 transformants were grown in 20 ml 2X YT broth (50 μg/mL kanamycin) at 37ºC overnight with shaking. The primary culture was used to inoculate a secondary culture of 500 ml 2X YT broth (50 μg/mL kanamycin). The secondary cultures were grown at 37ºC with shaking (220 rpm) to an OD600 of 0.7. This was followed by four hours of induction with 0.1 mM IPTG at 180 rpm. The cells were harvested by centrifugation at 10,000g for 1 hr at 4ºC. The bacterial pellets were resuspended with PBS buffer containing 25 mM imidazole and 0.1 mg/ml lysozyme and then frozen overnight to facilitate lysis of bacterial cell. The frozen suspensions were thawed and sonicated at 40% amplitude with a probe sonicator. The lysed and sonicated slurry was then ultracentrifuged at 80,000g for 1 hr at 4ºC and the resultant supernatant was retained. The supernatant was then subjected to nickel column chromatography using PBS with 25 mM imidazole as the binding and wash buffer and PBS with 500 mM imidazole as the elution buffer. The eluents were screened for the presence of proteins by SDS-PAGE and the positive fractions were combined for storage at 4ºC. Before using the proteins, the SUMO fusion partner was removed using added Ulp1 protease (1U per 100 μg of substrate) at 4ºC overnight under mild nutation. The extent of cleavage was confirmed by SDS-PAGE. The gel bands corresponding to the AMPs were also excised and subjected to in-gel tryptic digestion (Thermo Fisher). After the digestion with trypsin, confirmation of the proteins' identity

was performed by LC-ESI-MS (Waters) at the Baylor University Mass Spectrometry Center using samples obtained by in-gel tryptic digestions of SDS-PAGE bands of the respective proteins. The analysis of the MS data was done by BioLynx.

In Vitro Bactericidal Activity Assay

The Ulp-1 protease-cleaved proteins were tested for antimicrobial assays against four strains of bacteria: *Staphylococcus aureus, Staphylococcus epidermidis, Enterococcus faecalis* and *Bacillus subtilis*. These four strains were selected because they are gram positive and the AMPs plectasin and eurocin are specifically active against gram positive bacteria (Mygind et al. 2005; Oeemig et al. 2012). The control used for the experiment was free fusion partner SUMO protein dissolved in PBS as the vehicle. Vancomycin was used as the positive control, which was experimentally determined to be active against these bacteria. The standard protocol for a microtiter plate assay with serial dilution was used (Sarker, Nahar, and Kumarasamy 2007). Briefly, the first well of the 12-well row in the 96 well microtiter plate contained 50µl of the highest concentration of test protein/control solution with serial 2-fold dilutions leading to the last well having 2^{-11th} of the concentration as the initial well. The serial dilution was done with PBS buffer and additional 30µl of Tryptic-Soy Broth (TSB)/LB media was given to the wells before inoculating with 10µl of the bacterial culture. For inoculation, the bacteria were grown in TSB/LB media overnight and then diluted in the same media to meet the McFarland 0.5 standard. After inoculation, the plates were grown at 37ºC for 8 to 12 hours (depending on the strain). After the initial growth period, 10µl of resazurin solution (0.0015% w/v in DI water) was added. After adding resazurin, the plates were allowed to grow for 30 min to an hour before checking the progress. The results were reconfirmed by allowing the plates to grow further for a

period of 12 hours and then checked for the change in coloration of the wells. Each test and control peptide were tested against each strain of bacteria for n>5 replicates.

Results

Protein Expression and Purification

Strong expression of all four proteins, targeted and untargeted, were observed. The cleaved AMPs (with or without the targeting domain) and the SUMO fusion partner, at 4-6 kDa and ~17 kD respectively, were clearly visualized with SDS-PAGE (Figure 2). For a further confirmation, the trypsin-digested proteins were extracted from the SDS-PAGE gel bands and detected by mass spectrometry (LC-ESI-MS, Waters). Peptide identities were confirmed using the Biolynx application (Waters), which created hypothetical MS peaks by virtual trypsin digestion of the four protein sequences and matched them with the spectrum generated experimentally. The average from SDS-PAGE bands with NIH ImageJ and are provided in Table 5.2.

In Vitro Bactericidal Activity Assay

A differential toxicity was observed, with the addition of the viral A12 targeting domain driving a loss of activity against the nontarget species rather than a gain of activity against the target species. A12C-AMPs retained their toxicity against both target bacterial species but showed a dramatic decrease in toxicity against nontarget species compared to nontargeted AMP (Figure 3). This data is presented in tabular format in Supplementary Table 5.1. Purified SUMO dissolved in PBS was used as a negative control for the experiment and showed no antimicrobial activity. The targeted versions of the peptides did not confer any added advantage to the AMPs when acting on both the Staphylococci, as

both the non-targeted and targeted version of the AMPs had similar killing potential as evidenced by the box plot in Figure 3. For the nontarget *E. faecalis* and *B. subtilis*, the attachment of the A12C fusion partner increased the mean MIC values for both plectasin and eurocin to over 70 μM compared to <10 μM seen without the fusion partner ($p<0.001$; ANOVA 2-tailed test). For *S. aureus* and *S.epidermidis*, however, no significant rise in MIC values was observed upon attachment of the fusion partner for either eurocin or plectasin.

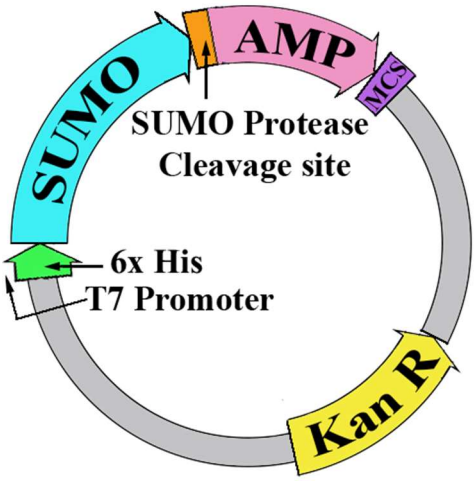

Figure 5.1: pE-SUMOstar/AMP *E. coli* vector. The SUMO protease cleavage site allowed the release of AMP (plectasin or eurocin) from the SUMO fusion partner. MCS, multiple cloning site (MCS).

Table 5.1. AMPs with and without viral targeting moiety from phage A12C.

Peptide	Sequence	Molecular Weight (in Da)
Plectasin	GFGCNGPWDEDDMQCHNHCK SIKGYKGGYCAKGGFVCKCY	4408
A12C- Plectasin	GVHMVAGPGREPTGGGHMGF GCNGPWDEDDMQCHNHCKSI KGYKGGYCAKGGFVCKCY	6137
Eurocin	GFGCPGDAYQCSEHCRALGG GRTGGYCAGPWYLGHPTCTCSF	4345
A12C-Eurocin	GVHMVAGPGREPTGGGHMGF GCPGDAYQCSEHCRALGGGR TGGYCAGPWYLGHPTCTCSF	6074
* The underlined sequence is the A12C targeting domain		

Table 5.2. Mean Yield (n=3) of targeted and nontargeted AMPs from E. coli/SUMO expression system.

Peptide	mg per L of cell culture	µmol per L of cell culture
Plectasin	15.7	3.6
A12C-Plectasin	26.1	4.2
Eurocin	10.2	2.4
A12C-Eurocin	19.5	3.2

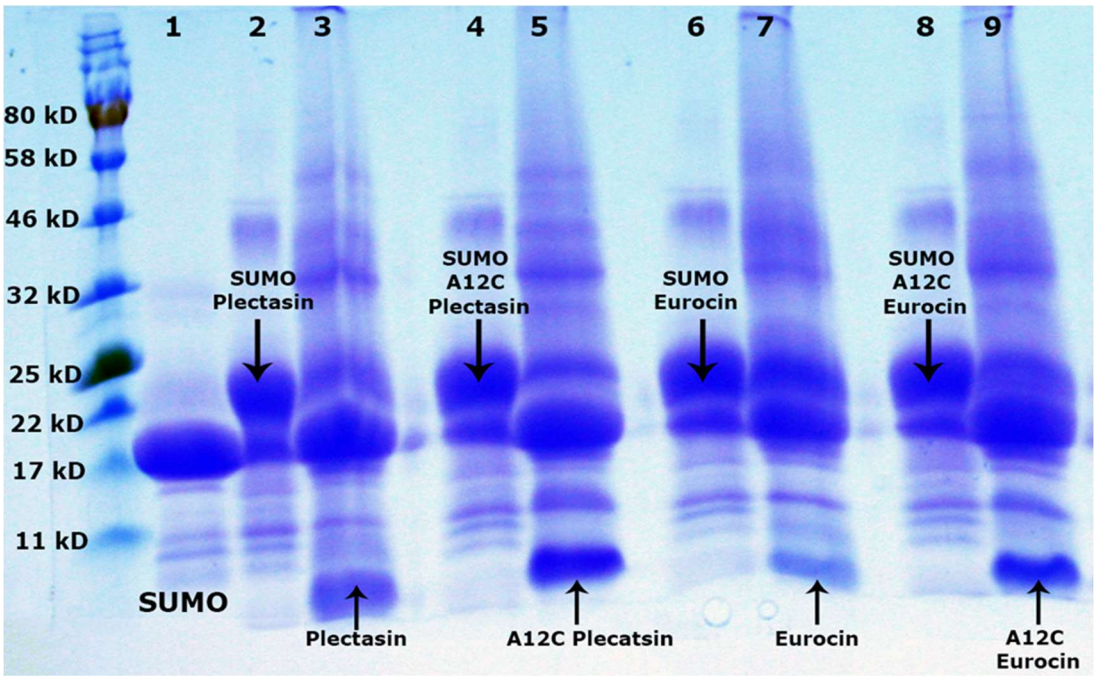

Figure 5.2. Expression of SUMO/AMP in *E. coli* and cleavage of AMP free of SUMO fusion partner. Plectasin (lane 2), A12C Plectasin (lane 4), Eurocin (lane 6), A12C Eurocin (lane 8) expressed with the SUMO fusion partner. On cleaving with SUMO protease (Ulp1), the cleaved SUMO protein can be seen at 17 kD on lanes 3, 5, 7 and 9; free SUMO protein control is in lane 1. The released AMPs, with and without targeting moieties, are in the same lanes as with the cleaved SUMO below 11 kDa.

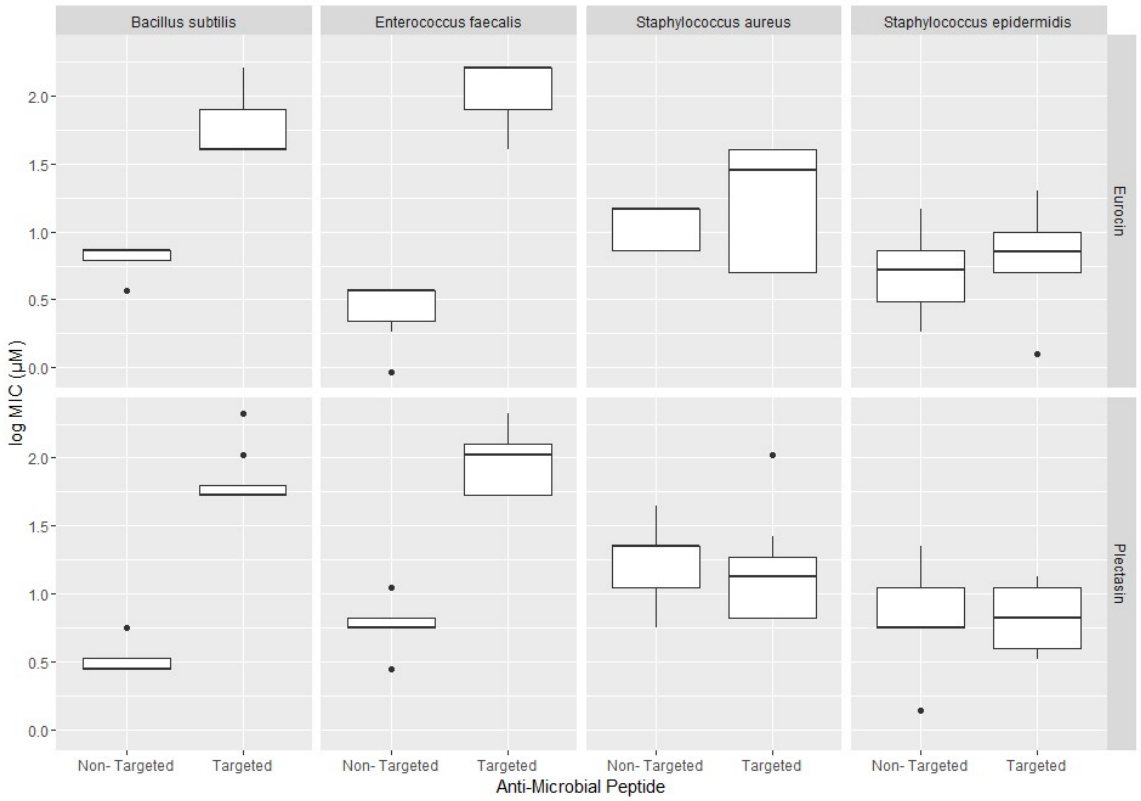

Figure 5.3. Log values for minimum inhibitory concentrations (MIC) in μM for non-targeted and targeted eurocin and plectasin against Bacillus subtilis, Enterococcus faecalis, Staphylococcus aureus and Staphylococcus epidermidis. The boxed regions represent 50% of the values while the bars represent 95%.

Discussion

With the rise of antibiotic-resistant bacterial infections, the discovery of new antimicrobial agents has become essential. AMPs are potentially less sensitive to develop resistance as they employ broadly targeted mechanism of toxicity. In addition, the advancement of sequencing technology and predictive algorithms (Islam et al. 2015; Islam, Kearney, and Baker 2017; Xiao et al. 2013) has expedited the discovery of new AMPs. This allows for data mining and the collection of large libraries of presumably well-adapted and functional native AMPs. However, as we have now gained an appreciation of the need

to preserve native microbiomes, it is seen that a limitation of AMP applications in biotechnology is their broad range of antimicrobial activity without sufficient specificity.

Eliminating pathogenic organisms without affecting the commensal microorganisms is an important property for the next generation of antibiotics. Disturbing the microflora can lead to the rise of opportunistic pathogens and decreased health outcomes generally. In the pursuit to achieve specificity in their activity, several studies have already demonstrated the development of targeted antimicrobial action against *Streptococcus mutans* (Eckert et al. 2006), *Enterococcus faecalis* (Qiu et al. 2005), and *Staphylococcus aureus* (Mao et al. 2013). In most cases, targeting moieties were derived from pheromone or quorum sensing peptides. However, an AMP fused to a targeting domain of bacteriophage origin has, to our knowledge, not been reported.

In this study we produced the specifically targeted AMPs, A12C-Plectasin and A12C-Eurocin, fused with a filamentous phage protein which has previously been shown to have a selective action against *Staphylococcus* bacteria (Yacoby et al. 2006). We observed a larger MIC for non-staphyloccoccal bacteria by the A12C AMPs compared to the non-targeted parental AMPs, while non-targeted and targeted AMPs exhibited similar MICs on both staphylococci (see Fig 3 and Supplementary Table 5.1). The result was a set of targeted AMPs with antimicrobial activity specific to *Staphylococcus* while showing no significant antimicrobial action towards non-target bacterial species.

It is challenging to express high quantities of soluble, correctly folded and biologically active AMPs in *E. coli* (Ingham and Moore 2007). Nevertheless, we were able to harvest AMPs at relatively high concentrations (see Table 5.2) using the SUMO fusion partner. We used the SUMO expression system and obtained a high concentration of the

target proteins which also displayed the expected activity following the protease cleavage and separated from their SUMO fusion partner. An equal concentration of SUMO lacked toxicity, demonstrating that the toxicity was the property of the AMP and not the fusion partner.

Continued investigation of targeting moieties for targeted AMPs is necessary to keep pace with the constantly increasing number of antibiotic-resistant bacterial infections. As an advancement, we have demonstrated a targeted AMP using the combination of a phage display protein and an AMP for the first time. This study not only demonstrated the viability of using a viral protein as a targeting moiety, but also showed the toxicity of the AMP towards the target pathogen was equal to that of its non-targeted counterpart. Most pathogenic bacteria are vulnerable to a specific phage. These phages are, therefore, an abundant and widely applicable source of targeting peptides (Elbreki et al. 2014; Matsuzaki et al. 2005; Viertel, Ritter, and Horz 2014) directing AMPs against specific bacterial pathogens.

References

Bruin, Eric C. de, Erwin H. Duitman, Arjo L. de Boer, Marten Veenhuis, Ineke G. A. Bos, and C. Erik Hack. 2005. "Pharmaceutical Proteins From Methylotrophic Yeasts." In *Therapeutic Proteins*, by C. Mark Smales and David C. James, 308:065–076. New Jersey: Humana Press. https://doi.org/10.1385/1-59259-922-2:065.

Charles, Patrick G. P., and M. Lindsay Grayson. 2004. "The Dearth of New Antibiotic Development: Why We Should Be Worried and What We Can Do about It." *The Medical Journal of Australia* 181 (10): 549–53.

Cregg, James M., Ilya Tolstorukov, Anasua Kusari, Jay Sunga, Knut Madden, and Thomas Chappell. 2009. "Chapter 13 Expression in the Yeast Pichia Pastoris." In *Methods in Enzymology*, 463:169–89. Elsevier. https://doi.org/10.1016/S0076-6879(09)63013-5.

Eckert, R., J. He, D. K. Yarbrough, F. Qi, M. H. Anderson, and W. Shi. 2006. “Targeted Killing of Streptococcus Mutans by a Pheromone-Guided ‘Smart’ Antimicrobial Peptide.” *Antimicrobial Agents and Chemotherapy* 50 (11): 3651–57. https://doi.org/10.1128/AAC.00622-06.

Elbreki, Mohamed, R. Paul Ross, Colin Hill, Jim O’Mahony, Olivia McAuliffe, and Aidan Coffey. 2014. “Bacteriophages and Their Derivatives as Biotherapeutic Agents in Disease Prevention and Treatment.” *Journal of Viruses* 2014: 1–20. https://doi.org/10.1155/2014/382539.

Enright, Mark C., D. Ashley Robinson, Gaynor Randle, Edward J. Feil, Hajo Grundmann, and Brian G. Spratt. 2002. “The Evolutionary History of Methicillin-Resistant Staphylococcus Aureus (MRSA).” *Proceedings of the National Academy of Sciences of the United States of America* 99 (11): 7687–92. https://doi.org/10.1073/pnas.122108599.

Hancock, Robert E. W., and Hans-Georg Sahl. 2006. “Antimicrobial and Host-Defense Peptides as New Anti-Infective Therapeutic Strategies.” *Nature Biotechnology* 24 (12): 1551–57. https://doi.org/10.1038/nbt1267.

Huafang Lai, Qiang Chen, and Jonathan Hurtado Jake Stahnke. 2013. “Agroinfiltration as an Effective and Scalable Strategy of Gene Delivery for Production of Pharmaceutical Proteins.” *Advanced Techniques in Biology & Medicine* 01 (01). https://doi.org/10.4172/atbm.1000103.

Ingham, Aaron B., and Robert J. Moore. 2007. “Recombinant Production of Antimicrobial Peptides in Heterologous Microbial Systems.” *Biotechnology and Applied Biochemistry* 47 (1): 1. https://doi.org/10.1042/BA20060207.

Islam, S M Ashiqul, Benjamin J Heil, Christopher Michel Kearney, and Erich J Baker. 2017. “Protein Classification Using Modified N-Grams and Skip-Grams.” *Bioinformatics*, December. https://doi.org/10.1093/bioinformatics/btx823.

Islam, S M Ashiqul, Christopher Michel Kearney, and Erich J. Baker. 2017. “CSPred: A Machine-Learning-Based Compound Model to Identify the Functional Activities of Biologically-Stable Toxins.” In , 2254–55. IEEE. https://doi.org/10.1109/BIBM.2017.8218014.

Islam, S. M. Ashiqul, Tanvir Sajed, Christopher Michel Kearney, and Erich J Baker. 2015. “PredSTP: A Highly Accurate SVM Based Model to Predict Sequential Cystine Stabilized Peptides.” *BMC Bioinformatics* 16 (1). https://doi.org/10.1186/s12859-015-0633-x.

Li, Chun, Hans-Matti Blencke, Victoria Paulsen, Tor Haug, and Klara Stensvåg. 2010. “Powerful Workhorses for Antimicrobial Peptide Expression and Characterization.” *Bioengineered Bugs* 1 (3): 217–20. https://doi.org/10.4161/bbug.1.3.11721.

Li, Jian Feng, Jie Zhang, Ren Song, Jia Xin Zhang, Yang Shen, and Shuang Quan Zhang. 2009. “Production of a Cytotoxic Cationic Antibacterial Peptide in Escherichia Coli Using SUMO Fusion Partner.” *Applied Microbiology and Biotechnology* 84 (2): 383–88. https://doi.org/10.1007/s00253-009-2109-2.

Li, Yifeng. 2011. “Recombinant Production of Antimicrobial Peptides in Escherichia Coli: A Review.” *Protein Expression and Purification* 80 (2): 260–67. https://doi.org/10.1016/j.pep.2011.08.001.

Mao, Ruoyu, Da Teng, Xiumin Wang, Di Xi, Yong Zhang, Xiaoyuan Hu, Yalin Yang, and Jianhua Wang. 2013. “Design, Expression, and Characterization of a Novel Targeted Plectasin against Methicillin-Resistant Staphylococcus Aureus.” *Applied Microbiology and Biotechnology* 97 (9): 3991–4002. https://doi.org/10.1007/s00253-012-4508-z.

Maróti, Gergely, Attila Kereszt, Éva Kondorosi, and Peter Mergaert. 2011. “Natural Roles of Antimicrobial Peptides in Microbes, Plants and Animals.” *Research in Microbiology* 162 (4): 363–74. https://doi.org/10.1016/j.resmic.2011.02.005.

Matsuzaki, Shigenobu, Mohammad Rashel, Jumpei Uchiyama, Shingo Sakurai, Takako Ujihara, Masayuki Kuroda, Masahiko Ikeuchi, et al. 2005. “Bacteriophage Therapy: A Revitalized Therapy against Bacterial Infectious Diseases.” *Journal of Infection and Chemotherapy: Official Journal of the Japan Society of Chemotherapy* 11 (5): 211–19. https://doi.org/10.1007/s10156-005-0408-9.

Mygind, Per H., Rikke L. Fischer, Kirk M. Schnorr, Mogens T. Hansen, Carsten P. Sönksen, Svend Ludvigsen, Dorotea Raventós, et al. 2005. “Plectasin Is a Peptide Antibiotic with Therapeutic Potential from a Saprophytic Fungus.” *Nature* 437 (7061): 975–80. https://doi.org/10.1038/nature04051.

Nadal, Anna, Maria Montero, Nuri Company, Esther Badosa, Joaquima Messeguer, Laura Montesinos, Emilio Montesinos, and Maria Pla. 2012. “Constitutive Expression of Transgenes Encoding Derivatives of the Synthetic Antimicrobial Peptide BP100: Impact on Rice Host Plant Fitness.” *BMC Plant Biology* 12 (1): 159. https://doi.org/10.1186/1471-2229-12-159.

Nguyen, Leonard T., Evan F. Haney, and Hans J. Vogel. 2011. “The Expanding Scope of Antimicrobial Peptide Structures and Their Modes of Action.” *Trends in Biotechnology* 29 (9): 464–72. https://doi.org/10.1016/j.tibtech.2011.05.001.

Norrby, S Ragnar, Carl Erik Nord, and Roger Finch. 2005. “Lack of Development of New Antimicrobial Drugs: A Potential Serious Threat to Public Health.” *The Lancet Infectious Diseases* 5 (2): 115–19. https://doi.org/10.1016/S1473-3099(05)01283-1.

Oeemig, Jesper S., Carina Lynggaard, Daniel H. Knudsen, Frederik T. Hansen, Kent D. Nørgaard, Tanja Schneider, Brian S. Vad, et al. 2012. "Eurocin, a New Fungal Defensin: Structure, Lipid Binding, and Its Mode of Action." *The Journal of Biological Chemistry* 287 (50): 42361–72. https://doi.org/10.1074/jbc.M112.382028.

Parachin, Nádia Skorupa, Kelly Cristina Mulder, Antônio Américo Barbosa Viana, Simoni Campos Dias, and Octávio Luiz Franco. 2012. "Expression Systems for Heterologous Production of Antimicrobial Peptides." *Peptides* 38 (2): 446–56. https://doi.org/10.1016/j.peptides.2012.09.020.

Peschen, Dieter, He-Ping Li, Rainer Fischer, Fritz Kreuzaler, and Yu-Cai Liao. 2004. "Fusion Proteins Comprising a Fusarium-Specific Antibody Linked to Antifungal Peptides Protect Plants against a Fungal Pathogen." *Nature Biotechnology* 22 (6): 732–38. https://doi.org/10.1038/nbt970.

Pope, B., and H. M. Kent. 1996. "High Efficiency 5 Min Transformation of Escherichia Coli." *Nucleic Acids Research* 24 (3): 536–37. https://doi.org/10.1093/nar/24.3.536.

Qiu, X.-Q., J. Zhang, H. Wang, and G. Y. Wu. 2005. "A Novel Engineered Peptide, a Narrow-Spectrum Antibiotic, Is Effective against Vancomycin-Resistant Enterococcus Faecalis." *Antimicrobial Agents and Chemotherapy* 49 (3): 1184–89. https://doi.org/10.1128/AAC.49.3.1184-1189.2005.

Sarker, Satyajit D., Lutfun Nahar, and Yashodharan Kumarasamy. 2007. "Microtitre Plate-Based Antibacterial Assay Incorporating Resazurin as an Indicator of Cell Growth, and Its Application in the in Vitro Antibacterial Screening of Phytochemicals." *Methods* 42 (4): 321–24. https://doi.org/10.1016/j.ymeth.2007.01.006.

Soto, Sara M. 2013. "Role of Efflux Pumps in the Antibiotic Resistance of Bacteria Embedded in a Biofilm." *Virulence* 4 (3): 223–29. https://doi.org/10.4161/viru.23724.

Thung, I., H. Aramin, V. Vavinskaya, S. Gupta, J. Y. Park, S. E. Crowe, and M. A. Valasek. 2016. "Review Article: The Global Emergence of *Helicobacter Pylori* Antibiotic Resistance." *Alimentary Pharmacology & Therapeutics* 43 (4): 514–33. https://doi.org/10.1111/apt.13497.

Viertel, Tania Mareike, Klaus Ritter, and Hans-Peter Horz. 2014. "Viruses versus Bacteria-Novel Approaches to Phage Therapy as a Tool against Multidrug-Resistant Pathogens." *The Journal of Antimicrobial Chemotherapy* 69 (9): 2326–36. https://doi.org/10.1093/jac/dku173.

Wang, B., Y.-B. Chen, O. Ayalon, J. Bender, and A. Garen. 1999. “Human Single-Chain Fv Immunoconjugates Targeted to a Melanoma-Associated Chondroitin Sulfate Proteoglycan Mediate Specific Lysis of Human Melanoma Cells by Natural Killer Cells and Complement.” *Proceedings of the National Academy of Sciences* 96 (4): 1627–32. https://doi.org/10.1073/pnas.96.4.1627.

Wang, Guangshun, Biswajit Mishra, Kyle Lau, Tamara Lushnikova, Radha Golla, and Xiuqing Wang. 2015. “Antimicrobial Peptides in 2014.” *Pharmaceuticals* 8 (4): 123–50. https://doi.org/10.3390/ph8010123.

Xiao, Xuan, Pu Wang, Wei-Zhong Lin, Jian-Hua Jia, and Kuo-Chen Chou. 2013. “IAMP-2L: A Two-Level Multi-Label Classifier for Identifying Antimicrobial Peptides and Their Functional Types.” *Analytical Biochemistry* 436 (2): 168–77. https://doi.org/10.1016/j.ab.2013.01.019.

Yacoby, I., M. Shamis, H. Bar, D. Shabat, and I. Benhar. 2006. “Targeting Antibacterial Agents by Using Drug-Carrying Filamentous Bacteriophages.” *Antimicrobial Agents and Chemotherapy* 50 (6): 2087–97. https://doi.org/10.1128/AAC.00169-06.

Supplementary Data

Table S5.1. MIC and log MIC values of the 4 AMPs against the 4 bacteria

Bacteria	**Mean (SD) of MIC (μM) (n≥6)**			**Mean (SD) of log MIC (μM) (n≥6)**		
	Plectasin	**A12C-Plectasin**	**P-value**	**Plectasin**	**A12C-Plectasin**	**P-value**
Bacillus subtilis	**3.49 (1.21)**	**79.65 (53.1)**	**0.0020**	**0.52 (0.489)**	**1.83 (0.21)**	**1.133E-09**
***Enterococcus faecalis* OG1RF**	**6.62 (2.76)**	**112.83 (61.91)**	**0.0005**	**0.784 (0.18)**	**1.988 (0.235)**	**3.769E-08**
***Staphylococcus aureus* SA113**	**19.21 (10.87)**	**25.60 (33.55)**	**0.6238**	**1.21 (0.25)**	**1.16 (0.40)**	**0.7873**
Staphylococcus epidermidis	**8.96 (6.32)**	**7.74 (4.14)**	**0.7167**	**0.83 (0.35)**	**0.82 (0.24)**	**0.9558**
Bacteria	**Mean (SD) of MIC (μM) (n≥6)**			**Mean (SD) of log MIC (μM) (n≥6)**		
	Eurocin	**A12C-Eurocin**	**P-value**	**Eurocin**	**A12C-Eurocin**	**P-value**
Bacillus subtilis	**6.44 (1.59)**	**65.16 (39.78)**	**0.0016**	**0.79 (0.13)**	**1.75 (0.21)**	**7.341E-08**
***Enterococcus faecalis* OG1RF**	**3.03 (1.01)**	**124.31 (45.54)**	**2.512E-07**	**0.44 (0.2)**	**2.05 (0.20)**	**1.563E-12**
***Staphylococcus aureus* SA113**	**11.964 (3.56)**	**24.06 (16.59)**	**0.0744**	**1.05 (0.14)**	**1.212 (0.42)**	**0.3632**
Staphylococcus epidermidis	**5.98 (3.98)**	**8.56 (5.98)**	**0.3882**	**0.67 (0.29)**	**0.80 (0.37)**	**0.5442**

CHAPTER SIX

Conclusion

The objective of the computational part of this study was to construct machine learning-based models for the prediction of the structural and functional characteristics of the cystine-stabilized proteins. In chapter three, I developed an SVM-based model to predict STPs from the primary sequences of unknown peptides. There, I manually selected some promising characteristics from the primary sequences the peptides which could be helpful to classify STPs from nonSTPs. Eventually, I was able the build the PredSTP model that can classify between STPs and nonSTPs with a >94% accuracy and implements an automated method to find cystine stabilized toxins containing a compact arrangement of the tri-disulfide domain with minimal sequence identity. Therefore, this approach provides useful directions for enhancement of theoretical and experimental research to find new antimicrobial peptides, insecticides, and other stable peptide drug candidates by shortening the discovery time of potential bioactive peptides. Further research may benefit from a model that classifies all cystine stabilized peptide toxins (inhibitor or antimicrobial) into the different subgroups based on source, mode of action, and target organisms.

While PredSTP can predict STPs which could be used as stable toxin peptides, this model cannot classify the functional characteristics of the peptide. Therefore, there was a need to construct model/models to classify the functional characteristics cystine-stabilized peptides. One necessary step to create to machine learning-based model is feature generation from the data points. Though in PredSTP, the feature vectors were manually generated from the primary

sequences of the peptide, this process is not efficient for constructing multiple models. Hence, I developed the m-NGSG framework that is described in Chapter Three to automate the feature generation step. The meta-comparison results outlined in this study following the construct of the m-NGSG framework illustrate this is a useful fully automated feature generation method. This framework will benefit the machine learning-based protein classification community, particularity those interested in classification based on primary protein sequence. It is expected that m-NGSG will significantly reduce the workload for the feature generation step regardless of protein characteristics and sequence size. Moreover, by analyzing the feature importance, the distinctive part of the sequence (motif) in a protein class can be revealed, which is often difficult to discover using multiple sequence alignment.

Now that the m-NGSG framework is built, the generation of novel models to categorize cystine-stabilized peptide sequences based on a particular biological function becomes straightforward. Consequently, I developed CSPred which is a combination of five different models and can classify the cystine-stabilized peptides based on their function. CSPred can be used by any researcher for data mining of genomic data to rapidly accrue a set of candidate cystine-stabilized peptide sequences with the desired chemical functionality. The short and simple process is detailed in the Availability section in Chapter Four. This analytical capability has several practical outcomes. First, such genes could be upregulated in the source organism to avoid the use of foreign transgenes, such as upregulating native antimicrobial peptides in crop plants to fight plant disease. Second, a bank of channel blockers, for example, could be constructed from data mining a variety of genomes across several taxa to test an array of divergent sequences for functionality after expression of the peptides in the wet lab. Novel or more powerful activities might be found from such surveys as well as from hybrids created from

this bank of sequences. Finally, some fundamental questions may be asked once a fuller survey of cystine-stabilized peptides is completed for a given organism.

In the wet lab part of the study, I mainly worked on expression and functional assay of the cystine-stabilized antimicrobial peptide. In Chapter Five, I demonstrated production and purification system of the cystine-stabilized antimicrobial peptides with a targeted domain that came from a viral origin. Since most pathogenic bacteria are vulnerable to a specific phage, this approach should be broadly applicable as a control against bacterial pathogens. Paired with the methods described in my dissertation for datamining antimicrobial cystine-stabilized peptides, I have developed a complete system for developing targeted control measures against pathogenic bacteria for all host taxa.

REFERENCES

Adams, Michael E. 2004. "Agatoxins: Ion Channel Specific Toxins from the American Funnel Web Spider, Agelenopsis Aperta." *Toxicon* 43 (5): 509–25. https://doi.org/10.1016/j.toxicon.2004.02.004.

Akondi, Kalyana B., Markus Muttenthaler, Sébastien Dutertre, Quentin Kaas, David J. Craik, Richard J. Lewis, and Paul F. Alewood. 2014. "Discovery, Synthesis, and Structure–Activity Relationships of Conotoxins." *Chemical Reviews* 114 (11): 5815–47. https://doi.org/10.1021/cr400401e.

Aloush, Valerie, Shiri Navon-Venezia, Yardena Seigman-Igra, Shaltiel Cabili, and Yehuda Carmeli. 2006. "Multidrug-Resistant Pseudomonas Aeruginosa: Risk Factors and Clinical Impact." *Antimicrobial Agents and Chemotherapy* 50 (1): 43–48. https://doi.org/10.1128/AAC.50.1.43-48.2006.

Altschul, S. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25 (17): 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

Altschul, S F, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25 (17): 3389–3402.

Asgari, Ehsaneddin, and Mohammad R. K. Mofrad. 2015. "Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics." *PloS One* 10 (11): e0141287. https://doi.org/10.1371/journal.pone.0141287.

Avrutina, Olga, Hans-Ulrich Schmoldt, Dusica Gabrijelcic-Geiger, Dung Le Nguyen, Christian P. Sommerhoff, Ulf Diederichsen, and Harald Kolmar. 2005. "Trypsin Inhibition by Macrocyclic and Open-Chain Variants of the Squash Inhibitor MCoTI-II." *Biological Chemistry* 386 (12): 1301–6. https://doi.org/10.1515/BC.2005.148.

Beilen, J. B. van, M. Neuenschwander, T. H. M. Smits, C. Roth, S. B. Balada, and B. Witholt. 2002. "Rubredoxins Involved in Alkane Oxidation." *Journal of Bacteriology* 184 (6): 1722–32. https://doi.org/10.1128/JB.184.6.1722-1732.2002.

Bende, Niraj S., Sławomir Dziemborowicz, Mehdi Mobli, Volker Herzig, John Gilchrist, Jordan Wagner, Graham M. Nicholson, Glenn F. King, and Frank Bosmans. 2014. "A Distinct Sodium Channel Voltage-Sensor Locus Determines Insect Selectivity of the Spider Toxin Dc1a." *Nature Communications* 5: 4350. https://doi.org/10.1038/ncomms5350.

Bock, J R, and D A Gough. 2001a. “Predicting Protein--Protein Interactions from Primary Structure.” *Bioinformatics (Oxford, England)* 17 (5): 455–60.

Bock, J. R., and D. A. Gough. 2001b. “Predicting Protein--Protein Interactions from Primary Structure.” *Bioinformatics (Oxford, England)* 17 (5): 455–60.

Boeckmann, B. 2003. “The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL in 2003.” *Nucleic Acids Research* 31 (1): 365–70. https://doi.org/10.1093/nar/gkg095.

Bonham-Carter, Oliver, Joe Steele, and Dhundy Bastola. 2014. “Alignment-Free Genetic Sequence Comparisons: A Review of Recent Approaches by Word Analysis.” *Briefings in Bioinformatics* 15 (6): 890–905. https://doi.org/10.1093/bib/bbt052.

Bonning, Bryony C, Narinder Pal, Sijun Liu, Zhaohui Wang, S Sivakumar, Philip M Dixon, Glenn F King, and W Allen Miller. 2013. “Toxin Delivery by the Coat Protein of an Aphid-Vectored Plant Virus Provides Plant Resistance to Aphids.” *Nature Biotechnology* 32 (1): 102–5. https://doi.org/10.1038/nbt.2753.

Bosnjak, I., V. Bojovic, T. Segvic-Bubic, and A. Bielen. 2014. “Occurrence of Protein Disulfide Bonds in Different Domains of Life: A Comparison of Proteins from the Protein Data Bank.” *Protein Engineering Design and Selection* 27 (3): 65–72. https://doi.org/10.1093/protein/gzt063.

Bourinet, Emmanuel, and Gerald W. Zamponi. 2016. “Block of Voltage-Gated Calcium Channels by Peptide Toxins.” *Neuropharmacology*, October. https://doi.org/10.1016/j.neuropharm.2016.10.016.

Braunstein, A., N. Papo, and Y. Shai. 2004. “In Vitro Activity and Potency of an Intravenously Injected Antimicrobial Peptide and Its DL Amino Acid Analog in Mice Infected with Bacteria.” *Antimicrobial Agents and Chemotherapy* 48 (8): 3127–29. https://doi.org/10.1128/AAC.48.8.3127-3129.2004.

Brooke, B D, R H Hunt, F Chandre, P Carnevale, and M Coetzee. 2002. “Stable Chromosomal Inversion Polymorphisms and Insecticide Resistance in the Malaria Vector Mosquito Anopheles Gambiae (Diptera: Culicidae).” *Journal of Medical Entomology* 39 (4): 568–73.

Bruin, Eric C. de, Erwin H. Duitman, Arjo L. de Boer, Marten Veenhuis, Ineke G. A. Bos, and C. Erik Hack. 2005. “Pharmaceutical Proteins From Methylotrophic Yeasts.” In *Therapeutic Proteins*, by C. Mark Smales and David C. James, 308:065–076. New Jersey: Humana Press. https://doi.org/10.1385/1-59259-922-2:065.

Bulet, P, C Hetru, J L Dimarcq, and D Hoffmann. 1999. “Antimicrobial Peptides in Insects; Structure and Function.” *Developmental and Comparative Immunology* 23 (4–5): 329–44.

Butt, Tauseef R., Suzanne C. Edavettal, John P. Hall, and Michael R. Mattern. 2005. "SUMO Fusion Technology for Difficult-to-Express Proteins." *Protein Expression and Purification* 43 (1): 1–9. https://doi.org/10.1016/j.pep.2005.03.016.

Cai, C Z, L Y Han, Z L Ji, X Chen, and Y Z Chen. 2003a. "SVM-Prot: Web-Based Support Vector Machine Software for Functional Classification of a Protein from Its Primary Sequence." *Nucleic Acids Research* 31 (13): 3692–97.

Cai, C. Z., L. Y. Han, Z. L. Ji, X. Chen, and Y. Z. Chen. 2003b. "SVM-Prot: Web-Based Support Vector Machine Software for Functional Classification of a Protein from Its Primary Sequence." *Nucleic Acids Research* 31 (13): 3692–97.

Cai, Y. D., X. J. Liu, X. Xu, and G. P. Zhou. 2001. "Support Vector Machines for Predicting Protein Structural Class." *BMC Bioinformatics* 2: 3.

Cao, Jiuwen, and Lianglin Xiong. 2014. "Protein Sequence Classification with Improved Extreme Learning Machine Algorithms." *BioMed Research International* 2014: 103054. https://doi.org/10.1155/2014/103054.

Carlini, Célia R, and Maria Fátima Grossi-de-Sá. 2002. "Plant Toxic Proteins with Insecticidal Properties. A Review on Their Potentialities as Bioinsecticides." *Toxicon: Official Journal of the International Society on Toxinology* 40 (11): 1515–39.

Cavnar, William B., and M. Trenkle John. 1994. "N-Gram-Based Text Categorization." *Ann Arbor Mi* 48113 (2): 161–75.

Cereghino, Geoff P Lin, Joan Lin Cereghino, Christine Ilgen, and James M Cregg. 2002. "Production of Recombinant Proteins in Fermenter Cultures of the Yeast Pichia Pastoris." *Current Opinion in Biotechnology* 13 (4): 329–32.

Chan, Yau Sang, Randy Chi Fai Cheung, Lixin Xia, Jack Ho Wong, Tzi Bun Ng, and Wai Yee Chan. 2016. "Snake Venom Toxins: Toxicity and Medicinal Applications." *Applied Microbiology and Biotechnology* 100 (14): 6165–81. https://doi.org/10.1007/s00253-016-7610-9.

Chang, Darby, Yu-Tang Syu, and Po-Chang Lin. 2010. "Predicting the Protein-Protein Interactions Using Primary Structures with Predicted Protein Surface." *BMC Bioinformatics* 11 (Suppl 1): S3. https://doi.org/10.1186/1471-2105-11-S1-S3.

Charles, Patrick G. P., and M. Lindsay Grayson. 2004. "The Dearth of New Antibiotic Development: Why We Should Be Worried and What We Can Do about It." *The Medical Journal of Australia* 181 (10): 549–53.

Chaudhary, Kumardeep, Ritesh Kumar, Sandeep Singh, Abhishek Tuknait, Ankur Gautam, Deepika Mathur, Priya Anand, Grish C. Varshney, and Gajendra P. S. Raghava. 2016a. “A Web Server and Mobile App for Computing Hemolytic Potency of Peptides.” *Scientific Reports* 6 (March): 22843. https://doi.org/10.1038/srep22843.

———. 2016b. “A Web Server and Mobile App for Computing Hemolytic Potency of Peptides.” *Scientific Reports* 6 (1). https://doi.org/10.1038/srep22843.

Cheek, Sara, S. Sri Krishna, and Nick V. Grishin. 2006. “Structural Classification of Small, Disulfide-Rich Protein Domains.” *Journal of Molecular Biology* 359 (1): 215–37. https://doi.org/10.1016/j.jmb.2006.03.017.

Chevallier, M R, and M Aigle. 1979. “Qualitative Detection of Penicillinase Produced by Yeast Strains Carrying Chimeric Yeast-Coli Plasmids.” *FEBS Letters* 108 (1): 179–80.

Chou, Kuo-Chen. 2005. “Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes.” *Bioinformatics (Oxford, England)* 21 (1): 10–19. https://doi.org/10.1093/bioinformatics/bth466.

Cipáková, Ingrid, and Eva Hostinová. 2005. “Production of the Human-Beta-Defensin Using Saccharomyces Cerevisiae as a Host.” *Protein and Peptide Letters* 12 (6): 551–54.

Circo, Raffaella, Barbara Skerlavaj, Renato Gennaro, Antonio Amoroso, and Margherita Zanetti. 2002. “Structural and Functional Characterization of HBD-1(Ser35), a Peptide Deduced from a DEFB1 Polymorphism.” *Biochemical and Biophysical Research Communications* 293 (1): 586–92. https://doi.org/10.1016/S0006-291X(02)00267-X.

Clark, Wyatt T., and Predrag Radivojac. 2013. “VECTOR QUANTIZATION KERNELS FOR THE CLASSIFICATION OF PROTEIN SEQUENCES AND STRUCTURES.” In , 316–27. WORLD SCIENTIFIC. https://doi.org/10.1142/9789814583220_0031.

Cohen-Inbar, Or, and Menashe Zaaroor. 2016. “Glioblastoma Multiforme Targeted Therapy: The Chlorotoxin Story.” *Journal of Clinical Neuroscience* 33 (November): 52–58. https://doi.org/10.1016/j.jocn.2016.04.012.

Colgrave, Michelle L., Andrew C. Kotze, Yen-Hua Huang, John O’Grady, Shane M. Simonsen, and David J. Craik. 2008. “Cyclotides: Natural, Circular Plant Peptides That Possess Significant Activity against Gastrointestinal Nematode Parasites of Sheep.” *Biochemistry* 47 (20): 5581–89. https://doi.org/10.1021/bi800223y.

Colgrave, Michelle L., Andrew C. Kotze, David C. Ireland, Conan K. Wang, and David J. Craik. 2008. "The Anthelmintic Activity of the Cyclotides: Natural Variants with Enhanced Activity." *Chembiochem: A European Journal of Chemical Biology* 9 (12): 1939–45. https://doi.org/10.1002/cbic.200800174.

Colgrave, Michelle L., Andrew C. Kotze, Steven Kopp, James S. McCarthy, Glen T. Coleman, and David J. Craik. 2009. "Anthelmintic Activity of Cyclotides: In Vitro Studies with Canine and Human Hookworms." *Acta Tropica* 109 (2): 163–66. https://doi.org/10.1016/j.actatropica.2008.11.003.

Conibear, Anne C., Alexander Bochen, K. Johan Rosengren, Petar Stupar, Conan Wang, Horst Kessler, and David J. Craik. 2014. "The Cyclic Cystine Ladder of Theta-Defensins as a Stable, Bifunctional Scaffold: A Proof-of-Concept Study Using the Integrin-Binding RGD Motif." *ChemBioChem* 15 (3): 451–59. https://doi.org/10.1002/cbic.201300568.

Conibear, Anne C., K. Johan Rosengren, Norelle L. Daly, Sónia Troeira Henriques, and David J. Craik. 2013. "The Cyclic Cystine Ladder in θ-Defensins Is Important for Structure and Stability, but Not Antibacterial Activity." *The Journal of Biological Chemistry* 288 (15): 10830–40. https://doi.org/10.1074/jbc.M113.451047.

Conibear, Anne C., K. Johan Rosengren, Peta J. Harvey, and David J. Craik. 2012. "Structural Characterization of the Cyclic Cystine Ladder Motif of θ-Defensins." *Biochemistry* 51 (48): 9718–26. https://doi.org/10.1021/bi301363a.

Craik, David J., Shane Simonsen, and Norelle L. Daly. 2002. "The Cyclotides: Novel Macrocyclic Peptides as Scaffolds in Drug Design." *Current Opinion in Drug Discovery & Development* 5 (2): 251–60.

Cregg, James M., Ilya Tolstorukov, Anasua Kusari, Jay Sunga, Knut Madden, and Thomas Chappell. 2009. "Chapter 13 Expression in the Yeast Pichia Pastoris." In *Methods in Enzymology*, 463:169–89. Elsevier. https://doi.org/10.1016/S0076-6879(09)63013-5.

Cui, Hang, Vibhu Mittal, and Mayur Datar. n.d. "Comparative Experiments on Sentiment Classification for Online Product Reviews." In .

Daly, Rachel, and Milton T W Hearn. 2006. "Expression of the Human Activin Type I and II Receptor Extracellular Domains in Pichia Pastoris." *Protein Expression and Purification* 46 (2): 456–67. https://doi.org/10.1016/j.pep.2005.10.001.

Delaunois, Bertrand, Sylvain Cordelier, Alexandra Conreux, Christophe Clément, and Philippe Jeandet. 2009. "Molecular Engineering of Resveratrol in Plants." *Plant Biotechnology Journal* 7 (1): 2–12. https://doi.org/10.1111/j.1467-7652.2008.00377.x.

Ding, Hui, En-Ze Deng, Lu-Feng Yuan, Li Liu, Hao Lin, Wei Chen, and Kuo-Chen Chou. 2014. "ICTX-Type: A Sequence-Based Predictor for Identifying the Types of Conotoxins in Targeting Ion Channels." *BioMed Research International* 2014: 1–10. https://doi.org/10.1155/2014/286419.

Ding, Hui, Peng-Mian Feng, Wei Chen, and Hao Lin. 2014. "Identification of Bacteriophage Virion Proteins by the ANOVA Feature Selection and Analysis." *Mol. BioSyst.* 10 (8): 2229–35. https://doi.org/10.1039/C4MB00316K.

Ding, Hui, Hao Lin, Wei Chen, Zi-Qiang Li, Feng-Biao Guo, Jian Huang, and Nini Rao. 2014. "Prediction of Protein Structural Classes Based on Feature Selection Technique." *Interdisciplinary Sciences: Computational Life Sciences* 6 (3): 235–40. https://doi.org/10.1007/s12539-013-0205-6.

Du, Pufeng, Shengjiao Cao, and Yanda Li. 2009a. "SubChlo: Predicting Protein Subchloroplast Locations with Pseudo-Amino Acid Composition and the Evidence-Theoretic K-Nearest Neighbor (ET-KNN) Algorithm." *Journal of Theoretical Biology* 261 (2): 330–35. https://doi.org/10.1016/j.jtbi.2009.08.004.

———. 2009b. "SubChlo: Predicting Protein Subchloroplast Locations with Pseudo-Amino Acid Composition and the Evidence-Theoretic K-Nearest Neighbor (ET-KNN) Algorithm." *Journal of Theoretical Biology* 261 (2): 330–35. https://doi.org/10.1016/j.jtbi.2009.08.004.

Du, Pufeng, Shuwang Gu, and Yasen Jiao. 2014. "PseAAC-General: Fast Building Various Modes of General Form of Chou's Pseudo-Amino Acid Composition for Large-Scale Protein Datasets." *International Journal of Molecular Sciences* 15 (3): 3495–3506. https://doi.org/10.3390/ijms15033495.

Dubchak, I., I. Muchnik, S. R. Holbrook, and S. H. Kim. 1995. "Prediction of Protein Folding Class Using Global Description of Amino Acid Sequence." *Proceedings of the National Academy of Sciences of the United States of America* 92 (19): 8700–8704.

Dutertre, Sébastien, Annette Nicke, and Victor I. Tsetlin. 2017. "Nicotinic Acetylcholine Receptor Inhibitors Derived from Snake and Snail Venoms." *Neuropharmacology*, June. https://doi.org/10.1016/j.neuropharm.2017.06.011.

Eckert, R., J. He, D. K. Yarbrough, F. Qi, M. H. Anderson, and W. Shi. 2006. "Targeted Killing of Streptococcus Mutans by a Pheromone-Guided 'Smart' Antimicrobial Peptide." *Antimicrobial Agents and Chemotherapy* 50 (11): 3651–57. https://doi.org/10.1128/AAC.00622-06.

Edgar, R. C. 2004. "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput." *Nucleic Acids Research* 32 (5): 1792–97. https://doi.org/10.1093/nar/gkh340.

Elbreki, Mohamed, R. Paul Ross, Colin Hill, Jim O'Mahony, Olivia McAuliffe, and Aidan Coffey. 2014. "Bacteriophages and Their Derivatives as Biotherapeutic Agents in Disease Prevention and Treatment." *Journal of Viruses* 2014: 1–20. https://doi.org/10.1155/2014/382539.

Emeleus, H. J. 1959. *Advances in Inorganic Chemistry*. New York: Academic Press.

Enright, Mark C., D. Ashley Robinson, Gaynor Randle, Edward J. Feil, Hajo Grundmann, and Brian G. Spratt. 2002. "The Evolutionary History of Methicillin-Resistant Staphylococcus Aureus (MRSA)." *Proceedings of the National Academy of Sciences of the United States of America* 99 (11): 7687–92. https://doi.org/10.1073/pnas.122108599.

Ettayapuram Ramaprasad, Azhagiya Singam, Sandeep Singh, Raghava Gajendra P S, and Subramanian Venkatesan. 2015. "AntiAngioPred: A Server for Prediction of Anti-Angiogenic Peptides." *PloS One* 10 (9): e0136990. https://doi.org/10.1371/journal.pone.0136990.

Fan, Yong-Xian, Jiangning Song, Xiangzeng Kong, and Hong-Bin Shen. 2011. "PredCSF: An Integrated Feature-Based Approach for Predicting Conotoxin Superfamily." *Protein & Peptide Letters* 18 (3): 261–67. https://doi.org/10.2174/092986611794578341.

Finn, R. D., J. Clements, and S. R. Eddy. 2011. "HMMER Web Server: Interactive Sequence Similarity Searching." *Nucleic Acids Research* 39 (suppl): W29–37. https://doi.org/10.1093/nar/gkr367.

Frazão, Bárbara, Vitor Vasconcelos, and Agostinho Antunes. 2012. "Sea Anemone (Cnidaria, Anthozoa, Actiniaria) Toxins: An Overview." *Marine Drugs* 10 (12): 1812–51. https://doi.org/10.3390/md10081812.

Fry, Bryan G., Kim Roelants, Donald E. Champagne, Holger Scheib, Joel D. A. Tyndall, Glenn F. King, Timo J. Nevalainen, et al. 2009. "The Toxicogenomic Multiverse: Convergent Recruitment of Proteins into Animal Venoms." *Annual Review of Genomics and Human Genetics* 10: 483–511. https://doi.org/10.1146/annurev.genom.9.081307.164356.

Gao, Bingmiao, Chao Peng, Jiaan Yang, Yunhai Yi, Junqing Zhang, and Qiong Shi. 2017. "Cone Snails: A Big Store of Conotoxins for Novel Drug Discovery." *Toxins* 9 (12). https://doi.org/10.3390/toxins9120397.

García-Olmedo, F., A. Molina, A. Segura, and M. Moreno. 1995. "The Defensive Role of Nonspecific Lipid-Transfer Proteins in Plants." *Trends in Microbiology* 3 (2): 72–74.

Garg, Aarti, Manoj Bhasin, and Gajendra P. S. Raghava. 2005. "Support Vector Machine-Based Method for Subcellular Localization of Human Proteins Using Amino Acid Compositions, Their Order, and Similarity Search." *The Journal of Biological Chemistry* 280 (15): 14427–32. https://doi.org/10.1074/jbc.M411789200.

Gelly, J.-C. 2004. "The KNOTTIN Website and Database: A New Information System Dedicated to the Knottin Scaffold." *Nucleic Acids Research* 32 (90001): 156D–159. https://doi.org/10.1093/nar/gkh015.

Gelly, Jean-Christophe, Jérôme Gracy, Quentin Kaas, Dung Le-Nguyen, Annie Heitz, and Laurent Chiche. 2004. "The KNOTTIN Website and Database: A New Information System Dedicated to the Knottin Scaffold." *Nucleic Acids Research* 32 (Database issue): D156-159. https://doi.org/10.1093/nar/gkh015.

Ghiassi, M., J. Skinner, and D. Zimbra. 2013. "Twitter Brand Sentiment Analysis: A Hybrid System Using n-Gram Analysis and Dynamic Artificial Neural Network." *Expert Systems with Applications* 40 (16): 6266–82. https://doi.org/10.1016/j.eswa.2013.05.057.

Góngora-Benítez, Miriam, Judit Tulla-Puche, and Fernando Albericio. 2014. "Multifaceted Roles of Disulfide Bonds. Peptides as Therapeutics." *Chemical Reviews* 114 (2): 901–26. https://doi.org/10.1021/cr400031z.

González, C., G. M. Langdon, M. Bruix, A. Gálvez, E. Valdivia, M. Maqueda, and M. Rico. 2000. "Bacteriocin AS-48, a Microbial Cyclic Polypeptide Structurally and Functionally Related to Mammalian NK-Lysin." *Proceedings of the National Academy of Sciences of the United States of America* 97 (21): 11221–26. https://doi.org/10.1073/pnas.210301097.

Gordon, Y. Jerold, Eric G. Romanowski, and Alison M. McDermott. 2005. "A Review of Antimicrobial Peptides and Their Therapeutic Potential as Anti-Infective Drugs." *Current Eye Research* 30 (7): 505–15. https://doi.org/10.1080/02713680590968637.

Gould, Andrew, Yanbin Ji, Teshome L. Aboye, and Julio A. Camarero. 2011. "Cyclotides, a Novel Ultrastable Polypeptide Scaffold for Drug Discovery." *Current Pharmaceutical Design* 17 (38): 4294–4307.

Gracy, J., D. Le-Nguyen, J.-C. Gelly, Q. Kaas, A. Heitz, and L. Chiche. 2007a. "KNOTTIN: The Knottin or Inhibitor Cystine Knot Scaffold in 2007." *Nucleic Acids Research* 36 (Database): D314–19. https://doi.org/10.1093/nar/gkm939.

———. 2007b. "KNOTTIN: The Knottin or Inhibitor Cystine Knot Scaffold in 2007." *Nucleic Acids Research* 36 (Database): D314–19. https://doi.org/10.1093/nar/gkm939.

Gracy, Jérôme, and Laurent Chiche. 2011. “Structure and Modeling of Knottins, a Promising Molecular Scaffold for Drug Discovery.” *Current Pharmaceutical Design* 17 (38): 4337–50.

Gracy, Jérôme, Dung Le-Nguyen, Jean-Christophe Gelly, Quentin Kaas, Annie Heitz, and Laurent Chiche. 2008a. “KNOTTIN: The Knottin or Inhibitor Cystine Knot Scaffold in 2007.” *Nucleic Acids Research* 36 (Database issue): D314-319. https://doi.org/10.1093/nar/gkm939.

———. 2008b. “KNOTTIN: The Knottin or Inhibitor Cystine Knot Scaffold in 2007.” *Nucleic Acids Research* 36 (Database issue): D314-319. https://doi.org/10.1093/nar/gkm939.

Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. 2006a. “A Closer Look at Skip-Gram Modelling.” In , 1–4. sn.

———. 2006b. “A Closer Look at Skip-Gram Modelling.” In *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, 1–4.

Hammami, R., J. Ben Hamida, G. Vergoten, and I. Fliss. 2009. “PhytAMP: A Database Dedicated to Antimicrobial Plant Peptides.” *Nucleic Acids Research* 37 (Database): D963–68. https://doi.org/10.1093/nar/gkn655.

Hancock, Robert E. W., and Hans-Georg Sahl. 2006. “Antimicrobial and Host-Defense Peptides as New Anti-Infective Therapeutic Strategies.” *Nature Biotechnology* 24 (12): 1551–57. https://doi.org/10.1038/nbt1267.

Hemingway, J, and H Ranson. 2000a. “Insecticide Resistance in Insect Vectors of Human Disease.” *Annual Review of Entomology* 45: 371–91. https://doi.org/10.1146/annurev.ento.45.1.371.

———. 2000b. “Insecticide Resistance in Insect Vectors of Human Disease.” *Annual Review of Entomology* 45: 371–91. https://doi.org/10.1146/annurev.ento.45.1.371.

Henriques, Sónia Troeira, and David J. Craik. 2010. “Cyclotides as Templates in Drug Design.” *Drug Discovery Today* 15 (1–2): 57–64. https://doi.org/10.1016/j.drudis.2009.10.007.

Hernandez-Garcia, Carlos M, Robert A Bouchard, Paul J Rushton, Michelle L Jones, Xianfeng Chen, Michael P Timko, and John J Finer. 2010. “High Level Transgenic Expression of Soybean (Glycine Max) GmERF and Gmubi Gene Promoters Isolated by a Novel Promoter Analysis Pipeline.” *BMC Plant Biology* 10 (1): 237. https://doi.org/10.1186/1471-2229-10-237.

Herzig, V., D. L. A. Wood, F. Newell, P.-A. Chaumeil, Q. Kaas, G. J. Binford, G. M. Nicholson, D. Gorse, and G. F. King. 2011. "ArachnoServer 2.0, an Updated Online Resource for Spider Toxin Sequences and Structures." *Nucleic Acids Research* 39 (Database): D653–57. https://doi.org/10.1093/nar/gkq1058.

Herzig, Volker, and Glenn King. 2015. "The Cystine Knot Is Responsible for the Exceptional Stability of the Insecticidal Spider Toxin ω-Hexatoxin-Hv1a." *Toxins* 7 (10): 4366–80. https://doi.org/10.3390/toxins7104366.

Herzig, Volker, David L A Wood, Felicity Newell, Pierre-Alain Chaumeil, Quentin Kaas, Greta J Binford, Graham M Nicholson, Dominique Gorse, and Glenn F King. 2011. "ArachnoServer 2.0, an Updated Online Resource for Spider Toxin Sequences and Structures." *Nucleic Acids Research* 39 (Database issue): D653-657. https://doi.org/10.1093/nar/gkq1058.

Hiramatsu, K, N Aritaka, H Hanaki, S Kawasaki, Y Hosoda, S Hori, Y Fukuchi, and I Kobayashi. 1997. "Dissemination in Japanese Hospitals of Strains of Staphylococcus Aureus Heterogeneously Resistant to Vancomycin." *Lancet* 350 (9092): 1670–73. https://doi.org/10.1016/S0140-6736(97)07324-8.

Horwege, Sebastian, Sebastian Lindner, Marcus Boden, Klas Hatje, Martin Kollmar, Chris-André Leimeister, and Burkhard Morgenstern. 2014. "Spaced Words and Kmacs: Fast Alignment-Free Sequence Comparison Based on Inexact Word Matches." *Nucleic Acids Research* 42 (Web Server issue): W7-11. https://doi.org/10.1093/nar/gku398.

Hu, Minqing, and Bing Liu. 2004. "Mining and Summarizing Customer Reviews." In , 168. ACM Press. https://doi.org/10.1145/1014052.1014073.

Hua, S., and Z. Sun. 2001. "Support Vector Machine Approach for Protein Subcellular Localization Prediction." *Bioinformatics* 17 (8): 721–28. https://doi.org/10.1093/bioinformatics/17.8.721.

Hua, Sujun, and Zhirong Sun. 2001. "A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach." *Journal of Molecular Biology* 308 (2): 397–407. https://doi.org/10.1006/jmbi.2001.4580.

Huafang Lai, Qiang Chen, and Jonathan Hurtado Jake Stahnke. 2013. "Agroinfiltration as an Effective and Scalable Strategy of Gene Delivery for Production of Pharmaceutical Proteins." *Advanced Techniques in Biology & Medicine* 01 (01). https://doi.org/10.4172/atbm.1000103.

Huang, Y., B. Niu, Y. Gao, L. Fu, and W. Li. 2010. "CD-HIT Suite: A Web Server for Clustering and Comparing Biological Sequences." *Bioinformatics* 26 (5): 680–82. https://doi.org/10.1093/bioinformatics/btq003.

Huang, Ying, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li. 2010. "CD-HIT Suite: A Web Server for Clustering and Comparing Biological Sequences." *Bioinformatics* 26 (5): 680–82. https://doi.org/10.1093/bioinformatics/btq003.

Ingham, Aaron B., and Robert J. Moore. 2007. "Recombinant Production of Antimicrobial Peptides in Heterologous Microbial Systems." *Biotechnology and Applied Biochemistry* 47 (1): 1. https://doi.org/10.1042/BA20060207.

Islam, S M Ashiqul, Benjamin J Heil, Christopher Michel Kearney, and Erich J Baker. 2017a. "Protein Classification Using Modified N-Grams and Skip-Grams." *Bioinformatics*, December. https://doi.org/10.1093/bioinformatics/btx823.

———. 2017b. "Protein Classification Using Modified N-Grams and Skip-Grams." *Bioinformatics*, December. https://doi.org/10.1093/bioinformatics/btx823.

Islam, S M Ashiqul, Christopher Michel Kearney, and Erich J. Baker. 2017. "CSPred: A Machine-Learning-Based Compound Model to Identify the Functional Activities of Biologically-Stable Toxins." In , 2254–55. IEEE. https://doi.org/10.1109/BIBM.2017.8218014.

Islam, S. M. Ashiqul, Tanvir Sajed, Christopher Michel Kearney, and Erich J. Baker. 2015a. "PredSTP: A Highly Accurate SVM Based Model to Predict Sequential Cystine Stabilized Peptides." *BMC Bioinformatics* 16 (July): 210. https://doi.org/10.1186/s12859-015-0633-x.

———. 2015b. "PredSTP: A Highly Accurate SVM Based Model to Predict Sequential Cystine Stabilized Peptides." *BMC Bioinformatics* 16 (July): 210. https://doi.org/10.1186/s12859-015-0633-x.

Islam, S. M. Ashiqul, Tanvir Sajed, Christopher Michel Kearney, and Erich J Baker. 2015c. "PredSTP: A Highly Accurate SVM Based Model to Predict Sequential Cystine Stabilized Peptides." *BMC Bioinformatics* 16 (1). https://doi.org/10.1186/s12859-015-0633-x.

Islam, SM Ashiqul, Christopher Michel Kearney, Ankan Choudhury, and Erich J. Baker. 2017. "Protein Classification Using Modified *N-Gram* and *Skip-Gram* Models: Extended Abstract." In , 586–586. ACM Press. https://doi.org/10.1145/3107411.3108193.

Jennings, C., J. West, C. Waine, D. Craik, and M. Anderson. 2001. "Biosynthesis and Insecticidal Properties of Plant Cyclotides: The Cyclic Knotted Proteins from Oldenlandia Affinis." *Proceedings of the National Academy of Sciences of the United States of America* 98 (19): 10614–19. https://doi.org/10.1073/pnas.191366898.

Jennings, Cameron V., K. Johan Rosengren, Norelle L. Daly, Manuel Plan, Jackie Stevens, Martin J. Scanlon, Clement Waine, David G. Norman, Marilyn A. Anderson, and David J. Craik. 2005a. "Isolation, Solution Structure, and Insecticidal Activity of Kalata B2, a Circular Protein with a Twist: Do Möbius Strips Exist in Nature?" *Biochemistry* 44 (3): 851–60. https://doi.org/10.1021/bi047837h.

———. 2005b. "Isolation, Solution Structure, and Insecticidal Activity of Kalata B2, a Circular Protein with a Twist: Do Möbius Strips Exist in Nature?" *Biochemistry* 44 (3): 851–60. https://doi.org/10.1021/bi047837h.

Jia, Jianhua, Zi Liu, Xuan Xiao, Bingxiang Liu, and Kuo-Chen Chou. 2015. "IPPI-Esml: An Ensemble Classifier for Identifying the Interactions of Proteins by Incorporating Their Physicochemical Properties and Wavelet Transforms into PseAAC." *Journal of Theoretical Biology* 377 (July): 47–56. https://doi.org/10.1016/j.jtbi.2015.04.011.

Kaas, Q., R. Yu, A.-H. Jin, S. Dutertre, and D. J. Craik. 2012. "ConoServer: Updated Content, Knowledge, and Discovery Tools in the Conopeptide Database." *Nucleic Acids Research* 40 (D1): D325–30. https://doi.org/10.1093/nar/gkr886.

Kaas, Quentin, Rilei Yu, Ai-Hua Jin, Sébastien Dutertre, and David J Craik. 2012. "ConoServer: Updated Content, Knowledge, and Discovery Tools in the Conopeptide Database." *Nucleic Acids Research* 40 (Database issue): D325-330. https://doi.org/10.1093/nar/gkr886.

Kawashima, S. 2000. "AAindex: Amino Acid Index Database." *Nucleic Acids Research* 28 (1): 374–374. https://doi.org/10.1093/nar/28.1.374.

Keˇselj, Vlado, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. "N-GRAM-BASED AUTHOR PROFILES FOR AUTHORSHIP ATTRIBUTION." In *Pacific Association for Computational Linguistics*.

Kedarisetti, Pradyumna, Marcin J. Mizianty, Quentin Kaas, David J. Craik, and Lukasz Kurgan. 2014a. "Prediction and Characterization of Cyclic Proteins from Sequences in Three Domains of Life." *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1844 (1): 181–90. https://doi.org/10.1016/j.bbapap.2013.05.002.

———. 2014b. "Prediction and Characterization of Cyclic Proteins from Sequences in Three Domains of Life." *Biochimica Et Biophysica Acta* 1844 (1 Pt B): 181–90. https://doi.org/10.1016/j.bbapap.2013.05.002.

———. 2014c. "Prediction and Characterization of Cyclic Proteins from Sequences in Three Domains of Life." *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1844 (1): 181–90. https://doi.org/10.1016/j.bbapap.2013.05.002.

———. 2014d. “Prediction and Characterization of Cyclic Proteins from Sequences in Three Domains of Life.” *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1844 (1): 181–90. https://doi.org/10.1016/j.bbapap.2013.05.002.

King, Glenn F., and Margaret C. Hardy. 2013. “Spider-Venom Peptides: Structure, Pharmacology, and Potential for Control of Insect Pests.” *Annual Review of Entomology* 58 (1): 475–96. https://doi.org/10.1146/annurev-ento-120811-153650.

Koehbach, Johannes, Alfred F. Attah, Andreas Berger, Roland Hellinger, Toni M. Kutchan, Eric J. Carpenter, Megan Rolf, et al. 2013. “Cyclotide Discovery in Gentianales Revisited-Identification and Characterization of Cyclic Cystine-Knot Peptides and Their Phylogenetic Distribution in Rubiaceae Plants: Cyclotide Discovery in Gentianales Revisited.” *Biopolymers* 100 (5): 438–52. https://doi.org/10.1002/bip.22328.

Koehbach, Johannes, Margaret O’Brien, Markus Muttenthaler, Marion Miazzo, Muharrem Akcan, Alysha G. Elliott, Norelle L. Daly, et al. 2013. “Oxytocic 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* 110 (52): 21183–88. https://doi.org/10.1073/pnas.1311183110.

Kuzmenkov, A. I., E. V. Grishin, and A. A. Vassilevski. 2015. “Diversity of Potassium Channel Ligands: Focus on Scorpion Toxins.” *Biochemistry (Moscow)* 80 (13): 1764–99. https://doi.org/10.1134/S0006297915130118.

Layer, P., and V. Stanghellini. 2014. “Review Article: Linaclotide for the Management of Irritable Bowel Syndrome with Constipation.” *Alimentary Pharmacology & Therapeutics* 39 (4): 371–84. https://doi.org/10.1111/apt.12604.

Lehrer, R I, A K Lichtenstein, and T Ganz. 1993. “Defensins: Antimicrobial and Cytotoxic Peptides of Mammalian Cells.” *Annual Review of Immunology* 11: 105–28. https://doi.org/10.1146/annurev.iy.11.040193.000541.

Leslie, Christina, Eleazar Eskin, and William Stafford Noble. 2001. “THE SPECTRUM KERNEL: A STRING KERNEL FOR SVM PROTEIN CLASSIFICATION.” In , 564–75. WORLD SCIENTIFIC. https://doi.org/10.1142/9789812799623_0053.

Leslie, Christina S., Eleazar Eskin, Adiel Cohen, Jason Weston, and William Stafford Noble. 2004. “Mismatch String Kernels for Discriminative Protein Classification.” *Bioinformatics (Oxford, England)* 20 (4): 467–76. https://doi.org/10.1093/bioinformatics/btg431.

Lewis, Richard J., and Maria L. Garcia. 2003a. “Therapeutic Potential of Venom Peptides.” *Nature Reviews Drug Discovery* 2 (10): 790–802. https://doi.org/10.1038/nrd1197.

———. 2003b. “Therapeutic Potential of Venom Peptides.” *Nature Reviews Drug Discovery* 2 (10): 790–802. https://doi.org/10.1038/nrd1197.

Li, Chun, Hans-Matti Blencke, Victoria Paulsen, Tor Haug, and Klara Stensvåg. 2010. “Powerful Workhorses for Antimicrobial Peptide Expression and Characterization.” *Bioengineered Bugs* 1 (3): 217–20. https://doi.org/10.4161/bbug.1.3.11721.

Li, Jian Feng, Jie Zhang, Ren Song, Jia Xin Zhang, Yang Shen, and Shuang Quan Zhang. 2009. “Production of a Cytotoxic Cationic Antibacterial Peptide in Escherichia Coli Using SUMO Fusion Partner.” *Applied Microbiology and Biotechnology* 84 (2): 383–88. https://doi.org/10.1007/s00253-009-2109-2.

Li, W., and A. Godzik. 2006. “Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.” *Bioinformatics* 22 (13): 1658–59. https://doi.org/10.1093/bioinformatics/btl158.

Li, Yifeng. 2011. “Recombinant Production of Antimicrobial Peptides in Escherichia Coli: A Review.” *Protein Expression and Purification* 80 (2): 260–67. https://doi.org/10.1016/j.pep.2011.08.001.

Liang, Song-Ping, and Xin Pan. 1995. “A Lectin-like Peptide Isolated from the Venom of the Chinese Bird Spider Selenocosmia Huwena.” *Toxicon* 33 (7): 875–82. https://doi.org/10.1016/0041-0101(95)00033-I.

Lin, Hao, Hui Ding, Feng-Biao Guo, An-Ying Zhang, and Jian Huang. 2008. “Predicting Subcellular Localization of Mycobacterial Proteins by Using Chous Pseudo Amino Acid Composition.” *Protein & Peptide Letters* 15 (7): 739–44. https://doi.org/10.2174/092986608785133681.

Lin, Hao, and Qian-Zhong Li. 2007. “Predicting Conotoxin Superfamily and Family by Using Pseudo Amino Acid Composition and Modified Mahalanobis Discriminant.” *Biochemical and Biophysical Research Communications* 354 (2): 548–51. https://doi.org/10.1016/j.bbrc.2007.01.011.

Liu, Bin, Jinghao Xu, Quan Zou, Ruifeng Xu, Xiaolong Wang, and Qingcai Chen. 2014. “Using Distances between Top-n-Gram and Residue Pairs for Protein Remote Homology Detection.” *BMC Bioinformatics* 15 (Suppl 2): S3. https://doi.org/10.1186/1471-2105-15-S2-S3.

Lu, S., J. Van Eck, X. Zhou, A. B. Lopez, D. M. O’Halloran, K. M. Cosman, B. J. Conlin, et al. 2006. “The Cauliflower Or Gene Encodes a DnaJ Cysteine-Rich Domain-Containing Protein That Mediates High Levels of -Carotene Accumulation.” *THE PLANT CELL ONLINE* 18 (12): 3594–3605. https://doi.org/10.1105/tpc.106.046417.

Malagón, David, Bonnie Botterill, Darren J. Gray, Erica Lovas, Mary Duke, Christian Gray, Steven R. Kopp, et al. 2013. "Anthelminthic Activity of the Cyclotides (Kalata B1 and B2) against Schistosome Parasites." *Biopolymers* 100 (5): 461–70. https://doi.org/10.1002/bip.22229.

Mao, Ruoyu, Da Teng, Xiumin Wang, Di Xi, Yong Zhang, Xiaoyuan Hu, Yalin Yang, and Jianhua Wang. 2013. "Design, Expression, and Characterization of a Novel Targeted Plectasin against Methicillin-Resistant Staphylococcus Aureus." *Applied Microbiology and Biotechnology* 97 (9): 3991–4002. https://doi.org/10.1007/s00253-012-4508-z.

Maróti, Gergely, Attila Kereszt, Éva Kondorosi, and Peter Mergaert. 2011. "Natural Roles of Antimicrobial Peptides in Microbes, Plants and Animals." *Research in Microbiology* 162 (4): 363–74. https://doi.org/10.1016/j.resmic.2011.02.005.

Marr, Alexandra K., William J. Gooderham, and Robert Ew Hancock. 2006. "Antibacterial Peptides for Therapeutic Use: Obstacles and Realistic Outlook." *Current Opinion in Pharmacology* 6 (5): 468–72. https://doi.org/10.1016/j.coph.2006.04.006.

Matsumura, M., G. Signor, and B. W. Matthews. 1989. "Substantial Increase of Protein Stability by Multiple Disulphide Bonds." *Nature* 342 (6247): 291–93. https://doi.org/10.1038/342291a0.

Matsuzaki, Shigenobu, Mohammad Rashel, Jumpei Uchiyama, Shingo Sakurai, Takako Ujihara, Masayuki Kuroda, Masahiko Ikeuchi, et al. 2005. "Bacteriophage Therapy: A Revitalized Therapy against Bacterial Infectious Diseases." *Journal of Infection and Chemotherapy: Official Journal of the Japan Society of Chemotherapy* 11 (5): 211–19. https://doi.org/10.1007/s10156-005-0408-9.

Mattanovich, Diethard, Paola Branduardi, Laura Dato, Brigitte Gasser, Michael Sauer, and Danilo Porro. 2012. "Recombinant Protein Production in Yeasts." In *Recombinant Gene Expression*, edited by Argelia Lorence, 824:329–58. Totowa, NJ: Humana Press. http://link.springer.com/10.1007/978-1-61779-433-9_17.

Mobli, Mehdi, Eivind A. B. Undheim, and Lachlan D. Rash. 2017. "Modulation of Ion Channels by Cysteine-Rich Peptides: From Sequence to Structure." *Advances in Pharmacology (San Diego, Calif.)* 79: 199–223. https://doi.org/10.1016/bs.apha.2017.03.001.

Mohabatkar, Hassan, Majid Mohammad Beigi, Kolsoum Abdolahi, and Sasan Mohsenzadeh. 2013. "Prediction of Allergenic Proteins by Means of the Concept of Chou's Pseudo Amino Acid Composition and a Machine Learning Approach." *Medicinal Chemistry (Shariqah (United Arab Emirates))* 9 (1): 133–37.

Mondal, Sukanta, Rajasekaran Bhavna, Rajasekaran Mohan Babu, and Suryanarayanarao Ramakumar. 2006. "Pseudo Amino Acid Composition and Multi-Class Support Vector Machines Approach for Conotoxin Superfamily Classification." *Journal of Theoretical Biology* 243 (2): 252–60. https://doi.org/10.1016/j.jtbi.2006.06.014.

Monroc, Sylvie, Esther Badosa, Lidia Feliu, Marta Planas, Emili Montesinos, and Eduard Bardají. 2006. "De Novo Designed Cyclic Cationic Peptides as Inhibitors of Plant Pathogenic Bacteria." *Peptides* 27 (11): 2567–74. https://doi.org/10.1016/j.peptides.2006.04.019.

Mourão, Caroline, and Elisabeth Schwartz. 2013. "Protease Inhibitors from Marine Venomous Animals and Their Counterparts in Terrestrial Venomous Animals." *Marine Drugs* 11 (6): 2069–2112. https://doi.org/10.3390/md11062069.

Muggleton, S, R D King, and M J Sternberg. 1992. "Protein Secondary Structure Prediction Using Logic-Based Machine Learning." *Protein Engineering* 5 (7): 647–57.

Mulvenna, Jason P, Conan Wang, and David J Craik. 2006. "CyBase: A Database of Cyclic Protein Sequence and Structure." *Nucleic Acids Research* 34 (Database issue): D192-194. https://doi.org/10.1093/nar/gkj005.

Munasinghe, Nehan, and MacDonald Christie. 2015. "Conotoxins That Could Provide Analgesia through Voltage Gated Sodium Channel Inhibition." *Toxins* 7 (12): 5386–5407. https://doi.org/10.3390/toxins7124890.

Murphy, M. Paul, and Harry LeVine. 2010. "Alzheimer's Disease and the Amyloid-β Peptide." Edited by Mark A. Lovell. *Journal of Alzheimer's Disease* 19 (1): 311–23. https://doi.org/10.3233/JAD-2010-1221.

Mygind, Per H., Rikke L. Fischer, Kirk M. Schnorr, Mogens T. Hansen, Carsten P. Sönksen, Svend Ludvigsen, Dorotea Raventós, et al. 2005. "Plectasin Is a Peptide Antibiotic with Therapeutic Potential from a Saprophytic Fungus." *Nature* 437 (7061): 975–80. https://doi.org/10.1038/nature04051.

Mylne, Joshua S., Lai Yue Chan, Aurelie H. Chanson, Norelle L. Daly, Hanno Schaefer, Timothy L. Bailey, Philip Nguyencong, Laura Cascales, and David J. Craik. 2012. "Cyclic Peptides Arising by Evolutionary Parallelism via Asparaginyl-Endopeptidase-Mediated Biosynthesis." *The Plant Cell* 24 (7): 2765–78. https://doi.org/10.1105/tpc.112.099085.

Nadal, Anna, Maria Montero, Nuri Company, Esther Badosa, Joaquima Messeguer, Laura Montesinos, Emilio Montesinos, and Maria Pla. 2012. "Constitutive Expression of Transgenes Encoding Derivatives of the Synthetic Antimicrobial Peptide BP100: Impact on Rice Host Plant Fitness." *BMC Plant Biology* 12 (1): 159. https://doi.org/10.1186/1471-2229-12-159.

Nguyen, Giang Kien Truc, Wei Han Lim, Phuong Quoc Thuc Nguyen, and James P. Tam. 2012. "Novel Cyclotides and Uncyclotides with Highly Shortened Precursors from Chassalia Chartacea and Effects of Methionine Oxidation on Bioactivities." *The Journal of Biological Chemistry* 287 (21): 17598–607. https://doi.org/10.1074/jbc.M111.338970.

Nguyen, Giang Kien Truc, Sen Zhang, Ngan Thi Kim Nguyen, Phuong Quoc Thuc Nguyen, Ming Sheau Chiu, Antony Hardjojo, and James P. Tam. 2011. "Discovery and Characterization of Novel Cyclotides Originated from Chimeric Precursors Consisting of Albumin-1 Chain a and Cyclotide Domains in the Fabaceae Family." *The Journal of Biological Chemistry* 286 (27): 24275–87. https://doi.org/10.1074/jbc.M111.229922.

Nguyen, Leonard T., Evan F. Haney, and Hans J. Vogel. 2011a. "The Expanding Scope of Antimicrobial Peptide Structures and Their Modes of Action." *Trends in Biotechnology* 29 (9): 464–72. https://doi.org/10.1016/j.tibtech.2011.05.001.

———. 2011b. "The Expanding Scope of Antimicrobial Peptide Structures and Their Modes of Action." *Trends in Biotechnology* 29 (9): 464–72. https://doi.org/10.1016/j.tibtech.2011.05.001.

Nielsen, H, S Brunak, and G von Heijne. 1999. "Machine Learning Approaches for the Prediction of Signal Peptides and Other Protein Sorting Signals." *Protein Engineering* 12 (1): 3–9.

Nielsen, Henrik. 2017. "Predicting Secretory Proteins with SignalP." *Methods in Molecular Biology (Clifton, N.J.)* 1611: 59–73. https://doi.org/10.1007/978-1-4939-7015-5_6.

Niemann, Hartmut H., Hans-Ulrich Schmoldt, Alexander Wentzel, Harald Kolmar, and Dirk W. Heinz. 2006. "Barnase Fusion as a Tool to Determine the Crystal Structure of the Small Disulfide-Rich Protein McoEeTI." *Journal of Molecular Biology* 356 (1): 1–8. https://doi.org/10.1016/j.jmb.2005.11.005.

Norrby, S Ragnar, Carl Erik Nord, and Roger Finch. 2005. "Lack of Development of New Antimicrobial Drugs: A Potential Serious Threat to Public Health." *The Lancet Infectious Diseases* 5 (2): 115–19. https://doi.org/10.1016/S1473-3099(05)01283-1.

Norton, Raymond S., and K. George Chandy. 2017. "Venom-Derived Peptide Inhibitors of Voltage-Gated Potassium Channels." *Neuropharmacology*, July. https://doi.org/10.1016/j.neuropharm.2017.07.002.

Oeemig, Jesper S., Carina Lynggaard, Daniel H. Knudsen, Frederik T. Hansen, Kent D. Nørgaard, Tanja Schneider, Brian S. Vad, et al. 2012. "Eurocin, a New Fungal Defensin: Structure, Lipid Binding, and Its Mode of Action." *The Journal of Biological Chemistry* 287 (50): 42361–72. https://doi.org/10.1074/jbc.M112.382028.

Olivera, B. M., W. R. Gray, R. Zeikus, J. M. McIntosh, J. Varga, J. Rivier, V. de Santos, and L. J. Cruz. 1985. “Peptide Neurotoxins from Fish-Hunting Cone Snails.” *Science (New York, N.Y.)* 230 (4732): 1338–43.

Ortiz, Ernesto, Georgina B. Gurrola, Elisabeth Ferroni Schwartz, and Lourival D. Possani. 2015. “Scorpion Venom Components as Potential Candidates for Drug Development.” *Toxicon* 93 (January): 125–35. https://doi.org/10.1016/j.toxicon.2014.11.233.

Ovchinnikova, Tatiana V., Sergey V. Balandin, Galina M. Aleshina, Andrey A. Tagaev, Yulia F. Leonova, Eugeny D. Krasnodembsky, Alexander V. Men’shenin, and Vladimir N. Kokryakov. 2006. “Aurelin, a Novel Antimicrobial Peptide from Jellyfish Aurelia Aurita with Structural Features of Defensins and Channel-Blocking Toxins.” *Biochemical and Biophysical Research Communications* 348 (2): 514–23. https://doi.org/10.1016/j.bbrc.2006.07.078.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. “Thumbs up?: Sentiment Classification Using Machine Learning Techniques.” In , 10:79–86. Association for Computational Linguistics. https://doi.org/10.3115/1118693.1118704.

Parachin, Nádia Skorupa, Kelly Cristina Mulder, Antônio Américo Barbosa Viana, Simoni Campos Dias, and Octávio Luiz Franco. 2012a. “Expression Systems for Heterologous Production of Antimicrobial Peptides.” *Peptides* 38 (2): 446–56. https://doi.org/10.1016/j.peptides.2012.09.020.

———. 2012b. “Expression Systems for Heterologous Production of Antimicrobial Peptides.” *Peptides* 38 (2): 446–56. https://doi.org/10.1016/j.peptides.2012.09.020.

Peschen, Dieter, He-Ping Li, Rainer Fischer, Fritz Kreuzaler, and Yu-Cai Liao. 2004. “Fusion Proteins Comprising a Fusarium-Specific Antibody Linked to Antifungal Peptides Protect Plants against a Fungal Pathogen.” *Nature Biotechnology* 22 (6): 732–38. https://doi.org/10.1038/nbt970.

Peters, Jenny, and Eva Stoger. 2011. “Transgenic Crops for the Production of Recombinant Vaccines and Anti-Microbial Antibodies.” *Human Vaccines* 7 (3): 367–74.

Pope, B., and H. M. Kent. 1996. “High Efficiency 5 Min Transformation of Escherichia Coli.” *Nucleic Acids Research* 24 (3): 536–37. https://doi.org/10.1093/nar/24.3.536.

Porto, William F., Valéria A. Souza, Diego O. Nolasco, and Octávio L. Franco. 2012. “In Silico Identification of Novel Hevein-like Peptide Precursors.” *Peptides* 38 (1): 127–36. https://doi.org/10.1016/j.peptides.2012.07.025.

Possani, Lourival D., Baltazar Becerril, Muriel Delepierre, and Jan Tytgat. 1999. "Scorpion Toxins Specific for Na+-Channels." *European Journal of Biochemistry* 264 (2): 287–300. https://doi.org/10.1046/j.1432-1327.1999.00625.x.

Postic, Guillaume, Jérôme Gracy, Charlotte Périn, Laurent Chiche, and Jean-Christophe Gelly. 2018. "KNOTTIN: The Database of Inhibitor Cystine Knot Scaffold after 10 Years, toward a Systematic Structure Modeling." *Nucleic Acids Research* 46 (D1): D454–58. https://doi.org/10.1093/nar/gkx1084.

Poth, Aaron G., Michelle L. Colgrave, Russell E. Lyons, Norelle L. Daly, and David J. Craik. 2011. "Discovery of an Unusual Biosynthetic Origin for Circular Proteins in Legumes." *Proceedings of the National Academy of Sciences of the United States of America* 108 (25): 10127–32. https://doi.org/10.1073/pnas.1103660108.

Poth, Aaron G., Joshua S. Mylne, Julia Grassl, Russell E. Lyons, A. Harvey Millar, Michelle L. Colgrave, and David J. Craik. 2012. "Cyclotides Associate with Leaf Vasculature and Are the Products of a Novel Precursor in Petunia (Solanaceae)." *The Journal of Biological Chemistry* 287 (32): 27033–46. https://doi.org/10.1074/jbc.M112.370841.

Pour-El, A. 1978. "Functionality and Protein Structure; Based on a Symposium." In .

Powers, David Martin. 2011. "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation." *Journal of Machine Learning Technologies* 2 (1): 37–63.

Qiu, Wang-Ren, Bi-Qian Sun, Xuan Xiao, Zhao-Chun Xu, and Kuo-Chen Chou. 2016. "IHyd-PseCp: Identify Hydroxyproline and Hydroxylysine in Proteins by Incorporating Sequence-Coupled Effects into General PseAAC." *Oncotarget* 7 (28): 44310–21. https://doi.org/10.18632/oncotarget.10027.

Qiu, X.-Q., J. Zhang, H. Wang, and G. Y. Wu. 2005. "A Novel Engineered Peptide, a Narrow-Spectrum Antibiotic, Is Effective against Vancomycin-Resistant Enterococcus Faecalis." *Antimicrobial Agents and Chemotherapy* 49 (3): 1184–89. https://doi.org/10.1128/AAC.49.3.1184-1189.2005.

Reddy, K. V. R., R. D. Yedery, and C. Aranha. 2004. "Antimicrobial Peptides: Premises and Promises." *International Journal of Antimicrobial Agents* 24 (6): 536–47. https://doi.org/10.1016/j.ijantimicag.2004.09.005.

Rees, D. C., and W. N. Lipscomb. 1982. "Refined Crystal Structure of the Potato Inhibitor Complex of Carboxypeptidase A at 2.5 A Resolution." *Journal of Molecular Biology* 160 (3): 475–98.

Rosengren, K. Johan, Norelle L. Daly, Manuel R. Plan, Clement Waine, and David J. Craik. 2003. "Twists, Knots, and Rings in Proteins. Structural Definition of the Cyclotide Framework." *The Journal of Biological Chemistry* 278 (10): 8606–16. https://doi.org/10.1074/jbc.M211147200.

Roth, Amy F., Ying Feng, Linyi Chen, and Nicholas G. Davis. 2002. “The Yeast DHHC Cysteine-Rich Domain Protein Akr1p Is a Palmitoyl Transferase.” *The Journal of Cell Biology* 159 (1): 23–28. https://doi.org/10.1083/jcb.200206120.

Ryan, M. C., and L. J. Sandell. 1990. “Differential Expression of a Cysteine-Rich Domain in the Amino-Terminal Propeptide of Type II (Cartilage) Procollagen by Alternative Splicing of MRNA.” *The Journal of Biological Chemistry* 265 (18): 10334–39.

Saether, Olav, David J. Craik, Iain D. Campbell, Knut Sletten, Jessie Juul, and David G. Norman. 1995. “Elucidation of the Primary and Three-Dimensional Structure of the Uterotonic Polypeptide Kalata B1.” *Biochemistry* 34 (13): 4147–58. https://doi.org/10.1021/bi00013a002.

Saez, Natalie J., Sebastian Senff, Jonas E. Jensen, Sing Yan Er, Volker Herzig, Lachlan D. Rash, and Glenn F. King. 2010. “Spider-Venom Peptides as Therapeutics.” *Toxins* 2 (12): 2851–71. https://doi.org/10.3390/toxins2122851.

Santibáñez-López, Carlos E., and Lourival D. Possani. 2015. “Overview of the Knottin Scorpion Toxin-like Peptides in Scorpion Venoms: Insights on Their Classification and Evolution.” *Toxicon* 107 (December): 317–26. https://doi.org/10.1016/j.toxicon.2015.06.029.

Sarker, Satyajit D., Lutfun Nahar, and Yashodharan Kumarasamy. 2007. “Microtitre Plate-Based Antibacterial Assay Incorporating Resazurin as an Indicator of Cell Growth, and Its Application in the in Vitro Antibacterial Screening of Phytochemicals.” *Methods* 42 (4): 321–24. https://doi.org/10.1016/j.ymeth.2007.01.006.

Schoeman, H, M A Vivier, M Du Toit, L M Dicks, and I S Pretorius. 1999. “The Development of Bactericidal Yeast Strains by Expressing the Pediococcus Acidilactici Pediocin Gene (PedA) in Saccharomyces Cerevisiae.” *Yeast (Chichester, England)* 15 (8): 647–56. https://doi.org/10.1002/(SICI)1097-0061(19990615)15:8<647::AID-YEA409>3.0.CO;2-5.

Schroeder, Bjoern O., Zhihong Wu, Sabine Nuding, Sandra Groscurth, Moritz Marcinowski, Julia Beisner, Johannes Buchner, Martin Schaller, Eduard F. Stange, and Jan Wehkamp. 2011. “Reduction of Disulphide Bonds Unmasks Potent Antimicrobial Activity of Human β-Defensin 1.” *Nature* 469 (7330): 419–23. https://doi.org/10.1038/nature09674.

Segura, Ana, Manuel Moreno, Francisco Madueño, Antonio Molina, and Francisco García-Olmedo. 1999. “Snakin-1, a Peptide from Potato That Is Active Against Plant Pathogens.” *Molecular Plant-Microbe Interactions* 12 (1): 16–23. https://doi.org/10.1094/MPMI.1999.12.1.16.

Sels, Jan, Janick Mathys, Barbara M. A. De Coninck, Bruno P. A. Cammue, and Miguel F. C. De Bolle. 2008. “Plant Pathogenesis-Related (PR) Proteins: A Focus on PR Peptides.” *Plant Physiology and Biochemistry: PPB* 46 (11): 941–50. https://doi.org/10.1016/j.plaphy.2008.06.011.

Sharma, Arun, Pallavi Kapoor, Ankur Gautam, Kumardeep Chaudhary, Rahul Kumar, Jagat Singh Chauhan, Atul Tyagi, and Gajendra P. S. Raghava. 2013a. “Computational Approach for Designing Tumor Homing Peptides.” *Scientific Reports* 3: 1607. https://doi.org/10.1038/srep01607.

———. 2013b. “Computational Approach for Designing Tumor Homing Peptides.” *Scientific Reports* 3: 1607. https://doi.org/10.1038/srep01607.

Shatzman, A R, and M Rosenberg. 1987. “Expression, Identification, and Characterization of Recombinant Gene Products in Escherichia Coli.” *Methods in Enzymology* 152: 661–73.

Shen, Hong-Bin, and Kuo-Chen Chou. 2008. “PseAAC: A Flexible Web Server for Generating Various Kinds of Protein Pseudo Amino Acid Composition.” *Analytical Biochemistry* 373 (2): 386–88. https://doi.org/10.1016/j.ab.2007.10.012.

Sigrist, Christian J. A., Lorenzo Cerutti, Edouard de Castro, Petra S. Langendijk-Genevaux, Virginie Bulliard, Amos Bairoch, and Nicolas Hulo. 2010. “PROSITE, a Protein Domain Database for Functional Characterization and Annotation.” *Nucleic Acids Research* 38 (suppl_1): D161–66. https://doi.org/10.1093/nar/gkp885.

Silverstein, Kevin A. T., William A. Moskal, Hank C. Wu, Beverly A. Underwood, Michelle A. Graham, Christopher D. Town, and Kathryn A. VandenBosch. 2007a. “Small Cysteine-Rich Peptides Resembling Antimicrobial Peptides Have Been under-Predicted in Plants.” *The Plant Journal: For Cell and Molecular Biology* 51 (2): 262–80. https://doi.org/10.1111/j.1365-313X.2007.03136.x.

Silverstein, Kevin A.T., William A. Moskal, Hank C. Wu, Beverly A. Underwood, Michelle A. Graham, Christopher D. Town, and Kathryn A. VandenBosch. 2007b. “Small Cysteine-Rich Peptides Resembling Antimicrobial Peptides Have Been under-Predicted in Plants: *Under-Predicted Cysteine-Rich Peptides in Plants*.” *The Plant Journal* 51 (2): 262–80. https://doi.org/10.1111/j.1365-313X.2007.03136.x.

Simeon, Saw, Watshara Shoombuatong, Nuttapat Anuwongcharoen, Likit Preeyanon, Virapong Prachayasittikul, Jarl E. S. Wikberg, and Chanin Nantasenamat. 2016a. “OsFP: A Web Server for Predicting the Oligomeric States of Fluorescent Proteins.” *Journal of Cheminformatics* 8: 72. https://doi.org/10.1186/s13321-016-0185-8.

———. 2016b. “OsFP: A Web Server for Predicting the Oligomeric States of Fluorescent Proteins.” *Journal of Cheminformatics* 8 (1). https://doi.org/10.1186/s13321-016-0185-8.

Simonsen, Shane M., Lillian Sando, David C. Ireland, Michelle L. Colgrave, Rekha Bharathi, Ulf Göransson, and David J. Craik. 2005. “A Continent of Plant Defense Peptide Diversity: Cyclotides in Australian Hybanthus (Violaceae).” *The Plant Cell* 17 (11): 3176–89. https://doi.org/10.1105/tpc.105.034678.

Singh, Pradip Kumar, Vipul Solanki, Shalley Sharma, Krishan Gopal Thakur, Beena Krishnan, and Suresh Korpole. 2015. “The Intramolecular Disulfide-Stapled Structure of Laterosporulin, a Class IId Bacteriocin, Conceals a Human Defensin-like Structural Module.” *The FEBS Journal* 282 (2): 203–14. https://doi.org/10.1111/febs.13129.

Sinha, Mau, Rashmi Prabha Singh, Gajraj Singh Kushwaha, Naseer Iqbal, Avinash Singh, Sanket Kaushik, Punit Kaur, Sujata Sharma, and Tej P. Singh. 2014. “Current Overview of Allergens of Plant Pathogenesis Related Protein Families.” *TheScientificWorldJournal* 2014: 543195. https://doi.org/10.1155/2014/543195.

Sit, Clarissa S., Ryan T. McKay, Colin Hill, R. Paul Ross, and John C. Vederas. 2011. “The 3D Structure of Thuricin CD, a Two-Component Bacteriocin with Cysteine Sulfur to α-Carbon Cross-Links.” *Journal of the American Chemical Society* 133 (20): 7680–83. https://doi.org/10.1021/ja201802f.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. “Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank.” In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–42.

Soto, Sara M. 2013. “Role of Efflux Pumps in the Antibiotic Resistance of Bacteria Embedded in a Biofilm.” *Virulence* 4 (3): 223–29. https://doi.org/10.4161/viru.23724.

Stintzi, A., T. Heitz, V. Prasad, S. Wiedemann-Merdinoglu, S. Kauffmann, P. Geoffroy, M. Legrand, and B. Fritig. 1993. “Plant ‘pathogenesis-Related’ Proteins and Their Role in Defense against Pathogens.” *Biochimie* 75 (8): 687–706.

Stotz, Henrik U., James Thomson, and Yueju Wang. 2009. “Plant Defensins: Defense, Development and Application.” *Plant Signaling & Behavior* 4 (11): 1010–12. https://doi.org/10.4161/psb.4.11.9755.

Talmadge, K, J Kaufman, and W Gilbert. 1980. “Bacteria Mature Preproinsulin to Proinsulin.” *Proceedings of the National Academy of Sciences of the United States of America* 77 (7): 3988–92.

Tam, James, Shujing Wang, Ka Wong, and Wei Tan. 2015. “Antimicrobial Peptides from Plants.” *Pharmaceuticals* 8 (4): 711–57. https://doi.org/10.3390/ph8040711.

Tan, Chade-Meng, Yuan-Fang Wang, and Chan-Do Lee. 2002. “The Use of Bigrams to Enhance Text Categorization.” *Information Processing & Management* 38 (4): 529–46. https://doi.org/10.1016/S0306-4573(01)00045-0.

Tan, Paul T. J., K. N. Srinivasan, Seng Hong Seah, Judice L. Y. Koh, Tin Wee Tan, Shoba Ranganathan, and Vladimir Brusic. 2005. “Accurate Prediction of Scorpion Toxin Functional Properties from Primary Structures.” *Journal of Molecular Graphics & Modelling* 24 (1): 17–24. https://doi.org/10.1016/j.jmgm.2005.01.003.

Tan, Paul T. J., Anitha Veeramani, Kellathur N. Srinivasan, Shoba Ranganathan, and Vladimir Brusic. 2006. “SCORPION2: A Database for Structure-Function Analysis of Scorpion Toxins.” *Toxicon: Official Journal of the International Society on Toxinology* 47 (3): 356–63. https://doi.org/10.1016/j.toxicon.2005.12.001.

Tang, Hua, Wei Chen, and Hao Lin. 2016. “Identification of Immunoglobulins Using Chou’s Pseudo Amino Acid Composition with Feature Selection Technique.” *Molecular BioSystems* 12 (4): 1269–75. https://doi.org/10.1039/C5MB00883B.

Teeter, Martha M., Jonathan A. Mazer, and James J. L’Italien. 1981. “Primary Structure of the Hydrophobic Plant Protein Crambin.” *Biochemistry* 20 (19): 5437–43. https://doi.org/10.1021/bi00522a013.

Teichert, Florian, Jonas Minning, Ugo Bastolla, and Markus Porto. 2010. “High Quality Protein Sequence Alignment by Combining Structural Profile Prediction and Profile Alignment Using SABER-TOOTH.” *BMC Bioinformatics* 11 (May): 251. https://doi.org/10.1186/1471-2105-11-251.

Thung, I., H. Aramin, V. Vavinskaya, S. Gupta, J. Y. Park, S. E. Crowe, and M. A. Valasek. 2016. “Review Article: The Global Emergence of *Helicobacter Pylori* Antibiotic Resistance.” *Alimentary Pharmacology & Therapeutics* 43 (4): 514–33. https://doi.org/10.1111/apt.13497.

Tiwari, Arvind Kumar. 2016. “Prediction of G-Protein Coupled Receptors and Their Subfamilies by Incorporating Various Sequence Features into Chou’s General PseAAC.” *Computer Methods and Programs in Biomedicine* 134 (October): 197–213. https://doi.org/10.1016/j.cmpb.2016.07.004.

Tugyi, Regina, Gábor Mezö, Erzsébet Fellinger, David Andreu, and Ferenc Hudecz. 2005. “The Effect of Cyclization on the Enzymatic Degradation of Herpes Simplex Virus Glycoprotein D Derived Epitope Peptide.” *Journal of Peptide Science: An Official Publication of the European Peptide Society* 11 (10): 642–49. https://doi.org/10.1002/psc.669.

Verma, Ruchi, and Ulrich Melcher. 2012. “A Support Vector Machine Based Method to Distinguish Proteobacterial Proteins from Eukaryotic Plant Proteins.” *BMC Bioinformatics* 13 Suppl 15: S9. https://doi.org/10.1186/1471-2105-13-S15-S9.

Viertel, Tania Mareike, Klaus Ritter, and Hans-Peter Horz. 2014. “Viruses versus Bacteria-Novel Approaches to Phage Therapy as a Tool against Multidrug-Resistant Pathogens.” *The Journal of Antimicrobial Chemotherapy* 69 (9): 2326–36. https://doi.org/10.1093/jac/dku173.

Vinga, Susana, and Jonas Almeida. 2003. “Alignment-Free Sequence Comparison-a Review.” *Bioinformatics (Oxford, England)* 19 (4): 513–23.

Waghu, Faiza Hanif, Ram Shankar Barai, Pratima Gurung, and Susan Idicula-Thomas. 2016. “CAMP $_{R3}$: A Database on Sequences, Structures and Signatures of Antimicrobial Peptides: Table 1.” *Nucleic Acids Research* 44 (D1): D1094–97. https://doi.org/10.1093/nar/gkv1051.

Wang, B., Y.-B. Chen, O. Ayalon, J. Bender, and A. Garen. 1999. “Human Single-Chain Fv Immunoconjugates Targeted to a Melanoma-Associated Chondroitin Sulfate Proteoglycan Mediate Specific Lysis of Human Melanoma Cells by Natural Killer Cells and Complement.” *Proceedings of the National Academy of Sciences* 96 (4): 1627–32. https://doi.org/10.1073/pnas.96.4.1627.

Wang, C. K. L., Q. Kaas, L. Chiche, and D. J. Craik. 2007. “CyBase: A Database of Cyclic Protein Sequences and Structures, with Applications in Protein Discovery and Engineering.” *Nucleic Acids Research* 36 (Database): D206–10. https://doi.org/10.1093/nar/gkm953.

Wang, Conan K L, Quentin Kaas, Laurent Chiche, and David J Craik. 2008. “CyBase: A Database of Cyclic Protein Sequences and Structures, with Applications in Protein Discovery and Engineering.” *Nucleic Acids Research* 36 (Database issue): D206-210. https://doi.org/10.1093/nar/gkm953.

Wang, Guangshun, Biswajit Mishra, Kyle Lau, Tamara Lushnikova, Radha Golla, and Xiuqing Wang. 2015. “Antimicrobial Peptides in 2014.” *Pharmaceuticals* 8 (4): 123–50. https://doi.org/10.3390/ph8010123.

Whetstone, Paul A, and Bruce D Hammock. 2007. “Delivery Methods for Peptide and Protein Toxins in Insect Control.” *Toxicon: Official Journal of the International Society on Toxinology* 49 (4): 576–96. https://doi.org/10.1016/j.toxicon.2006.11.009.

Wu, Yun, Yufei Zheng, and Hua Tang. 2016. “Identifying the Types of Ion Channel-Targeted Conotoxins by Incorporating New Properties of Residues into Pseudo Amino Acid Composition.” *BioMed Research International* 2016: 3981478. https://doi.org/10.1155/2016/3981478.

Xia, Jun-Feng, Kyungsook Han, and De-Shuang Huang. 2010. “Sequence-Based Prediction of Protein-Protein Interactions by Means of Rotation Forest and Autocorrelation Descriptor.” *Protein & Peptide Letters* 17 (1): 137–45. https://doi.org/10.2174/092986610789909403.

Xianfang, Wang, Wang Junmei, Wang Xiaolei, and Zhang Yue. 2017. “Predicting the Types of Ion Channel-Targeted Conotoxins Based on AVC-SVM Model.” *BioMed Research International* 2017: 2929807. https://doi.org/10.1155/2017/2929807.

Xiao, Xuan, Pu Wang, Wei-Zhong Lin, Jian-Hua Jia, and Kuo-Chen Chou. 2013a. “IAMP-2L: A Two-Level Multi-Label Classifier for Identifying Antimicrobial Peptides and Their Functional Types.” *Analytical Biochemistry* 436 (2): 168–77. https://doi.org/10.1016/j.ab.2013.01.019.

———. 2013b. “IAMP-2L: A Two-Level Multi-Label Classifier for Identifying Antimicrobial Peptides and Their Functional Types.” *Analytical Biochemistry* 436 (2): 168–77. https://doi.org/10.1016/j.ab.2013.01.019.

Xu, Yan, Jun Ding, Ling-Yun Wu, and Kuo-Chen Chou. 2013. “ISNO-PseAAC: Predict Cysteine S-Nitrosylation Sites in Proteins by Incorporating Position Specific Amino Acid Propensity into Pseudo Amino Acid Composition.” *PloS One* 8 (2): e55844. https://doi.org/10.1371/journal.pone.0055844.

Yacoby, I., M. Shamis, H. Bar, D. Shabat, and I. Benhar. 2006. “Targeting Antibacterial Agents by Using Drug-Carrying Filamentous Bacteriophages.” *Antimicrobial Agents and Chemotherapy* 50 (6): 2087–97. https://doi.org/10.1128/AAC.00169-06.

Ye, Mingyu, Keith K. Khoo, Shaoqiong Xu, Mi Zhou, Nonlawat Boonyalai, Matthew A. Perugini, Xiaoxia Shao, et al. 2012. “A Helical Conotoxin from Conus Imperialis Has a Novel Cysteine Framework and Defines a New Superfamily.” *The Journal of Biological Chemistry* 287 (18): 14973–83. https://doi.org/10.1074/jbc.M111.334615.

Yu, Chin-Sheng, Yu-Ching Chen, Chih-Hao Lu, and Jenn-Kang Hwang. 2006. “Prediction of Protein Subcellular Localization.” *Proteins: Structure, Function, and Bioinformatics* 64 (3): 643–51. https://doi.org/10.1002/prot.21018.

Yuan, Lu-Feng, Chen Ding, Shou-Hui Guo, Hui Ding, Wei Chen, and Hao Lin. 2013. “Prediction of the Types of Ion Channel-Targeted Conotoxins Based on Radial Basis Function Network.” *Toxicology in Vitro: An International Journal Published in Association with BIBRA* 27 (2): 852–56. https://doi.org/10.1016/j.tiv.2012.12.024.

Zeng, Zhiqiang, Hua Shi, Yun Wu, and Zhiling Hong. 2015. “Survey of Natural Language Processing Techniques in Bioinformatics.” *Computational and Mathematical Methods in Medicine* 2015: 1–10. https://doi.org/10.1155/2015/674296.

Zhang, Guang-Ya, and Bai-Shan Fang. 2008. “Predicting the Cofactors of Oxidoreductases Based on Amino Acid Composition Distribution and Chou’s Amphiphilic Pseudo-Amino Acid Composition.” *Journal of Theoretical Biology* 253 (2): 310–15. https://doi.org/10.1016/j.jtbi.2008.03.015.

Zhang, Jun, Zhengshuang Hua, Zebo Huang, QiZhu Chen, Qingyun Long, David J. Craik, Alan J. M. Baker, Wensheng Shu, and Bin Liao. 2015. “Two Blast-Independent Tools, CyPerl and CyExcel, for Harvesting Hundreds of Novel Cyclotides and Analogues from Plant Genomes and Protein Databases.” *Planta* 241 (4): 929–40. https://doi.org/10.1007/s00425-014-2229-5.

Zhao, Xiao-Wei, Zhi-Qiang Ma, and Ming-Hao Yin. 2012. “Predicting Protein-Protein Interactions by Combing Various Sequence- Derived Features into the General Form of Chou’s Pseudo Amino Acid Composition.” *Protein & Peptide Letters* 19 (5): 492–500. https://doi.org/10.2174/092986612800191080.

Zhirnov, O.P., H.D. Klenk, and P.F. Wright. 2011. “Aprotinin and Similar Protease Inhibitors as Drugs against Influenza.” *Antiviral Research* 92 (1): 27–36. https://doi.org/10.1016/j.antiviral.2011.07.014.

Zhu, Shunyi, Herve Darbon, Karin Dyason, Fons Verdonck, and Jan Tytgat. 2003. “Evolutionary Origin of Inhibitor Cystine Knot Peptides.” *The FASEB Journal* 17 (12): 1765–67. https://doi.org/10.1096/fj.02-1044fje.

Zhu, Shunyi, Steve Peigneur, Bin Gao, Lan Luo, Di Jin, Yong Zhao, and Jan Tytgat. 2011a. “Molecular Diversity and Functional Evolution of Scorpion Potassium Channel Toxins.” *Molecular & Cellular Proteomics: MCP* 10 (2): M110.002832. https://doi.org/10.1074/mcp.M110.002832.

———. 2011b. “Molecular Diversity and Functional Evolution of Scorpion Potassium Channel Toxins.” *Molecular & Cellular Proteomics: MCP* 10 (2): M110.002832. https://doi.org/10.1074/mcp.M110.002832.