

Gabriella Panuccio · Miguel Rocha ·
Florentino Fdez-Riverola ·
Mohd Saberi Mohamad ·
Roberto Casado-Vara *Editors*

Practical Applications
of Computational
Biology &
Bioinformatics,
14th International
Conference
(PACBB 2020)

Advances in Intelligent Systems and Computing

Volume 1240

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences,
Warsaw, Poland

Advisory Editors

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India

Rafael Bello Perez, Faculty of Mathematics, Physics and Computing,
Universidad Central de Las Villas, Santa Clara, Cuba

Emilio S. Corchado, University of Salamanca, Salamanca, Spain

Hani Hagras, School of Computer Science and Electronic Engineering,
University of Essex, Colchester, UK

László T. Kóczy, Department of Automation, Széchenyi István University,
Gyor, Hungary


Vladik Kreinovich, Department of Computer Science, University of Texas
at El Paso, El Paso, TX, USA

Chin-Teng Lin, Department of Electrical Engineering, National Chiao
Tung University, Hsinchu, Taiwan

Jie Lu, Faculty of Engineering and Information Technology,
University of Technology Sydney, Sydney, NSW, Australia

Patricia Melin, Graduate Program of Computer Science, Tijuana Institute
of Technology, Tijuana, Mexico

Nadia Nedjah, Department of Electronics Engineering, University of Rio de Janeiro,
Rio de Janeiro, Brazil

Ngoc Thanh Nguyen , Faculty of Computer Science and Management,
Wrocław University of Technology, Wrocław, Poland

Jun Wang, Department of Mechanical and Automation Engineering,
The Chinese University of Hong Kong, Shatin, Hong Kong

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing such as: computational intelligence, soft computing including neural networks, fuzzy systems, evolutionary computing and the fusion of these paradigms, social intelligence, ambient intelligence, computational neuroscience, artificial life, virtual worlds and society, cognitive science and systems, Perception and Vision, DNA and immune based systems, self-organizing and adaptive systems, e-Learning and teaching, human-centered and human-centric computing, recommender systems, intelligent control, robotics and mechatronics including human-machine teaming, knowledge-based paradigms, learning paradigms, machine ethics, intelligent data analysis, knowledge management, intelligent agents, intelligent decision making and support, intelligent network security, trust management, interactive entertainment, Web intelligence and multimedia.

The publications within “Advances in Intelligent Systems and Computing” are primarily proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

**** Indexing: The books of this series are submitted to ISI Proceedings, EI-Compendex, DBLP, SCOPUS, Google Scholar and Springerlink ****

More information about this series at <http://www.springer.com/series/11156>

Gabriella Panuccio · Miguel Rocha ·
Florentino Fdez-Riverola ·
Mohd Saberi Mohamad ·
Roberto Casado-Vara
Editors

Practical Applications
of Computational Biology
& Bioinformatics, 14th
International Conference
(PACBB 2020)

 Springer

Editors

Gabriella Panuccio
Enhanced Regenerative Medicine
Istituto Italiano di Tecnologia
Genoa, Genova, Italy

Miguel Rocha
Department de Informática
Universidade do Minho
Braga, Portugal

Florentino Fdez-Riverola
Computer Science Department
University of Vigo
Vigo, Spain

Mohd Saberi Mohamad
Institute for Artificial Intelligence and Big
Data (AIBIG)
Universiti Malaysia Kelantan, Kampus Kota
Kota Bharu, Malaysia

Roberto Casado-Vara
Biotechnology, Intelligent Systems
and Educational Technology (BISITE)
Research Group
University of Salamanca
Salamanca, Salamanca, Spain

ISSN 2194-5357

ISSN 2194-5365 (electronic)

Advances in Intelligent Systems and Computing

ISBN 978-3-030-54567-3

ISBN 978-3-030-54568-0 (eBook)

<https://doi.org/10.1007/978-3-030-54568-0>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

There are diverse sequencing techniques, and new technologies emerge continually, making it possible to obtain a large amount of multi-omics data. Bioscience is progressively turning into a kind of computer science, as it has begun to rely on computer science applications. As a result, bioinformatics and computational biology are fields that encounter new challenges as they attempt to analyze, process, assimilate, and get insight into data. To be able to overcome those challenges, it is necessary to develop new algorithms and approaches in fields such as databases, statistics, data mining, machine learning, optimization, computer science, machine learning, and artificial intelligence. A new generation of interdisciplinary researchers, with extensive background in biological and computational sciences, work on meeting those needs.

The International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB) is an annual international event dedicated to applied research and challenges in bioinformatics and computational biology. Building on the success of previous events, this volume gathers the contributions for the 14th PACBB Conference. All submissions have been thoroughly reviewed and selected by an international committee, which includes members from 21 different countries. The PACBB'20 technical program includes 21 papers of authors from many different countries (Australia, Colombia, Egypt, Germany, India, Malaysia, Portugal, Saudi Arabia, Slovakia, South Korea, Spain, Switzerland, Turkey, United Arab Emirates, UK, and USA) and different subfields in bioinformatics and computational biology. There will be special issues in JCR-ranked journals, such as *Interdisciplinary Sciences: Computational Life Sciences*, *Integrative Bioinformatics*, *Information Fusion*, *Neurocomputing*, *Sensors, Processes, and Electronics*. Therefore, this event will strongly promote the interaction among researchers from international research groups working in diverse fields. The scientific content will be innovative, and it will help improve the valuable work that is being carried out by the participants.

This symposium is organized by the University of L'Aquila with the collaboration of the University of Malaysia Kelantan, the University of Minho, the University of Vigo, and the University of Salamanca. We would like to thank all the

contributing authors, the members of the Program Committee, the sponsors (IBM, Indra, AEPIA, APPI, AIIS, EurAI, and AIR Institute). We thank for funding support to the project: “Intelligent and sustainable mobility supported by multi-agent systems and edge computing” (Id. RTI2018-095390-B-C32), and finally, we thank the Local Organization members and the Program Committee members for their valuable work, which is essential for the success of PACBB’20.

Gabriella Panuccio
Miguel Rocha
Florentino Fdez-Riverola
Mohd Saberi Mohamad
Roberto Casado-Vara

Organization

General Co-chairs

Gabriella Panuccio	University of Genoa, Italy
Miguel Rocha	University of Minho, Portugal
Florentino Fdez-Riverola	University of Vigo, Spain
Mohd Saberi Mohamad	Universiti Malaysia Kelantan, Malaysia
Roberto Casado-Vara	University of Salamanca, Spain

Program Committee

Vera Afreixo	University of Aveiro, Portugal
Amparo Alonso-Betanzos	University of A Coruña, Spain
Rene Alquezar	Technical University of Catalonia, Spain
Manuel Álvarez Díaz	University of A Coruña, Spain
Jeferson Arango Lopez	Universidad de Caldas, Colombia
Joel Arrais	University of Coimbra, Portugal
Julio Banga	Instituto de Investigaciones Marinas (C.S.I.C.), Spain
Carlos Bastos	University of Aveiro, Portugal
Carole Bernon	IRIT/UPS, France
Lourdes Borrajo	University of Vigo, Spain
Ana Cristina Braga	University of Minho, Portugal
Boris Brimkov	Rice University, USA
Guillermo Calderon	Autonomous University of Manizales, Colombia
Rui Camacho	University of Porto, Portugal
José Antonio Castellanos Garzón	University of Salamanca, Spain
Luis Fernando Castillo	Universidad de Caldas, Colombia
José Manuel Colom	University of Zaragoza, Spain

Fernanda Brito Correia	DETI/IEETA University of Aveiro and DEIS/ISEC/Polytechnic Institute of Coimbra, Portugal
Daniela Correia	University of Minho, Portugal
Ángel Martín del Rey	University of Salamanca, Spain
Roberto Costumero	Technical University of Madrid, Spain
Francisco Couto	University of Lisbon, Faculty of Sciences, Portugal
Yingbo Cui	National University of Defense Technology, China
Masoud Daneshtalab	KTH Royal Institute of Technology in Stockholm, Sweden
Javier De Las Rivas	University of Salamanca, Spain
Sergio Deusdado	Technical Institute of Bragança, Portugal
Oscar Dias	University of Minho, Portugal
Fernando Diaz	University of Valladolid, Spain
Ramón Doallo	Univ. A Coruña, Spain
Xavier Domingo-Almenara	Rovira i Virgili University, Spain
Pedro Ferreira	Ipatimup: Institute of Molecular Pathology and Immunology of the University of Porto, Portugal
João Diogo Ferreira	University of Lisbon, Faculty of Sciences, Portugal
Nuno Filipe	University of Porto, Portugal
Mohd Firdaus-Raih	National University of Malaysia, Malaysia
Nuno A. Fonseca	University of Porto, Portugal
Dino Franklin	Federal University of Uberlandia, Spain
Alvaro Gaitan	Café de colombia, Colombia
Narmer Galeano	Universidad Catolica de Manizales, Colombia
Vanessa Maria Gervin	Hathor Group, Brazil
Rosalba Giugno	University of Verona, Italia
Josep Gómez	University Rovira i Virgili, Spain
Patricia Gonzalez	University of A Coruña, Spain
Consuelo Gonzalo-Martin	Universidad Politécnica de Madrid, Spain
David Hoksza	University Karlova, Czech Republic
Roberto Casado-Vara	University of Salamanca, Spain
Natthakan Iam-On	Mae Fah Luang University, Thailand
Gustavo Isaza	University of Caldas, Colombia
Paula Jorge	University of Minho, Portugal
Martin Krallinger	National Center for Oncological Research, Spain
Rosalía Laza	Universidad de Vigo, Spain
Thierry Lecroq	University of Rouen, France
Giovani Librelotto	Federal University of Santa Maria, Portugal
Filipe Liu	CEB, University of Minho, Portugal
Ruben Lopez-Cortes	University of Vigo, Spain

Hugo López-Fernández	University of Vigo, Spain
Eva Lorenzo Iglesias	University of Vigo, Portugal
Analia Lourenco	University of Vigo, Spain
Sara Madeira	University of Lisbon, Faculty of Sciences, Portugal
Marcelo Maraschin	Federal University of Santa Catarina, Brazil
Marcos Martinez-Romero	Stanford University, USA
Sérgio Matos	IEETA, Universidade de Aveiro, Portugal
Mohd Saberi Mohamad	University Teknologi Malaysia, Spain
Loris Nanni	University of Padua, Italy
José Luis Oliveira	University of Aveiro, Portugal
Maria Olivia Pereira	University of Minho, Centre of Biological Engineering, Portugal
Alexandre Perera Lluna	Technical University of Catalonia, Spain
Martin Pérez Pérez	University of Vigo, SING group, Spain
Gael Pérez Rodríguez	University of Vigo, SING group, Spain
Cindy Perscheid	Hasso-Plattner-Institut, Denmark
Armando Pinho	University of Aveiro, Portugal
Ignacio Ponzoni	National South University, Argentina
Antonio Prestes Garcia	Universidad Politécnica de Madrid, Spain
Heri Ramampiaro	Norwegian University of Science and Technology, Norway
Juan Ranea	University of Malaga, Spain
Miguel Reboiro-Jato	University of Vigo, Spain
Jose Ignacio Requeno	University of Zaragoza, Spain
João Manuel Rodrigues	DETI/IEETA, University of Aveiro, Portugal
Alejandro Rodriguez	Universidad Politécnica de Madrid, Spain
Alfonso Rodriguez-Paton	Universidad Politécnica de Madrid, Spain
Miriam Rubio Camarillo	National Center for Oncological Research, Spain
Gustavo Santos-Garcia	Universidad de Salamanca, Spain
Pedro Sernadela	University of Aveiro, Portugal
Amin Shoukry	Egypt Japan Univ of Science and Technology, Egypt
Naresh Singhal	University of Auckland, New Zealand
Ana Margarida Sousa	University of Minho, Portugal
Niclas Ståhl	University of Skovde, Sweden
Carolyn Talcott	SRI International, USA
Mehmet Tan	TOBB University of Economics and Technology, Turkey
Rita Margarida Teixeira	ESTG - IPL, Portugal
Ascenso	
Mark Thompson	LUMC, Netherland
Antonio J. Tomeu-Hardasmal	University of Cadiz, Spain
Alicia Troncoso	Universidad Pablo de Olavide, Spain
Turki Turki	New Jersey Institute of Technology, USA

Eduardo Valente	IPCB, Portugal
Alfredo Vellido	Technical University of Catalonia, Spain
Jorge Vieira	University of Porto, Portugal
Alejandro F. Villaverde	Instituto de Investigaciones Marinas (C.S.I.C.), Spain
Pierpaolo Vittorini	University of L'Aquila - Department of Life, Health, and Environmental Sciences, Italy

Organizing Committee

Juan M. Corchado Rodríguez	University of Salamanca, Spain AIR Institute, Spain
Roberto Casado Vara	University of Salamanca, Spain
Fernando De la Prieta	University of Salamanca, Spain
Sara Rodríguez González	University of Salamanca, Spain
Javier Prieto Tejedor	University of Salamanca, Spain AIR Institute, Spain
Pablo Chamoso Santos	University of Salamanca, Spain
Belén Pérez Lancho	University of Salamanca, Spain
Ana Belén Gil González	University of Salamanca, Spain
Ana De Luis Reboredo	University of Salamanca, Spain
Angélica González Arrieta	University of Salamanca, Spain
Emilio S. Corchado Rodríguez	University of Salamanca, Spain
Ángel Martín del Rey	University of Salamanca, Spain
Ángel Luis Sánchez Lázaro	University of Salamanca, Spain
Alfonso González Briones	University Complutense of Madrid, Spain
Yeray Mezquita Martín	University of Salamanca, Spain
Enrique Goyenechea	University of Salamanca, Spain AIR Institute, Spain
Javier J. Martín Limorti	University of Salamanca, Spain
Alberto Rivas Camacho	University of Salamanca, Spain
Ines Sitton Candanedo	University of Salamanca, Spain
Elena Hernández Nieves	University of Salamanca, Spain
Beatriz Bellido	University of Salamanca, Spain
María Alonso	University of Salamanca, Spain
Diego Valdeolmillos	AIR Institute, Spain
Sergio Marquez	University of Salamanca, Spain
Jorge Herrera	University of Salamanca, Spain
Marta Plaza Hernández	University of Salamanca, Spain
David García Retuerta	University of Salamanca, Spain
Guillermo Hernández González	AIR Institute, Spain

Luis Carlos Martínez de Iturrate	University of Salamanca, Spain AIR Institute, Spain
Ricardo S. Alonso Rincón	University of Salamanca, Spain
Javier Parra	University of Salamanca, Spain
Niloufar Shoeibi	University of Salamanca, Spain
Zakieh Alizadeh-Sani	University of Salamanca, Spain





Local Organizing Committee

Pierpaolo Vittorini	University of L’Aquila, Italy
Tania Di Mascio	University of L’Aquila, Italy
Giovanni De Gasperis	University of L’Aquila, Italy
Federica Caruso	University of L’Aquila, Italy
Alessandra Galassi	University of L’Aquila, Italy

PACBB 2020 Sponsors

Sponsors	Organizers
 	   

Support from National Associations

Contents

Identification of Antimicrobial Peptides from Macroalgae with Machine Learning	1
Michela Caprani, Orla Slattery, Joan O’Keeffe, and John Healy	
A Health-Related Study from Food Online Reviews. The Case of Gluten-Free Foods	12
Martín Pérez-Pérez, Anália Lourenço, Gilberto Igrejas, and Florentino Fdez-Riverola	
The Activity of Bioinformatics Developers and Users in Stack Overflow	23
Roi Pérez-López, Guillermo Blanco, Florentino Fdez-Riverola, and Anália Lourenço	
ProPythia: A Python Automated Platform for the Classification of Proteins Using Machine Learning	32
Ana Marta Sequeira, Diana Lousa, and Miguel Rocha	
Inferences on <i>Mycobacterium Leprae</i> Host Immune Response Escape and Antibiotic Resistance Using Genomic Data and GenomeFastScreen	42
Hugo López-Fernández, Cristina P. Vieira, Florentino Fdez-Riverola, Miguel Reboiro-Jato, and Jorge Vieira	
Compi Hub: A Public Repository for Sharing and Discovering Compi Pipelines	51
Alba Nogueira-Rodríguez, Hugo López-Fernández, Osvaldo Graña-Castro, Miguel Reboiro-Jato, and Daniel Glez-Peña	
DeepACPpred: A Novel Hybrid CNN-RNN Architecture for Predicting Anti-Cancer Peptides	60
Nathaniel Lane and Indika Kahanda	

Preventing Cardiovascular Disease Development Establishing Cardiac Well-Being Indexes	70
Ana Duarte and Orlando Belo	
Fuzzy Matching for Cellular Signaling Networks in a Choroidal Melanoma Model	80
Adrián Riesco, Beatriz Santos-Buitrago, Merrill Knapp, Gustavo Santos-García, Emiliano Hernández Galilea, and Carolyn Talcott	
Towards A More Effective Bidirectional LSTM-Based Learning Model for Human-Bacterium Protein-Protein Interactions	91
Huaming Chen, Jun Shen, Lei Wang, and Yaochu Jin	
Machine Learning for Depression Screening in Online Communities	102
Alina Trifan, Rui Antunes, and José Luís Oliveira	
Towards Triclustering-Based Classification of Three-Way Clinical Data: A Case Study on Predicting Non-invasive Ventilation in ALS	112
Diogo Soares, Rui Henriques, Marta Gromicho, Susana Pinto, Mamede de Carvalho, and Sara C. Madeira	
Searching RNA Substructures with Arbitrary Pseudoknots	123
Michela Quadrini	
An Application of Ontological Engineering for Design and Specification of Ontocancro	134
Jéssica A. Bonini, Matheus D. Da Silva, Rafael Pereira, Bruno A. Mozzaquatro, Ricardo G. Martini, and Giovani R. Librelotto	
Evaluation of the Effect of Cell Parameters on the Number of Microtubule Merotelic Attachments in Metaphase Using a Three-Dimensional Computer Model	144
Maxim A. Krivov, Fazoil I. Ataulakhanov, and Pavel S. Ivanov	
Reconciliation of Regulatory Data: The Regulatory Networks of <i>Escherichia coli</i> and <i>Bacillus subtilis</i>	155
Diogo Lima, Fernando Cruz, Miguel Rocha, and Oscar Dias	
A Hybrid of Bat Algorithm and Minimization of Metabolic Adjustment for Succinate and Lactate Production	166
Mei Yen Man, Mohd Saberi Mohamad, Yee Wen Choon, and Mohd Arfian Ismail	
Robustness of Pathway Enrichment Analysis to Transcriptome-Wide Gene Expression Platform	176
Joanna Zyla, Kinga Leszczorz, and Joanna Polanska	

Hypoglycemia Prevention Using an Embedded Model Control with a Safety Scheme: In-silico Test 186
Fabian Leon-Vargas, Andres L. Jutinico, and Andres Molano-Jimenez

Bidirectional-Pass Algorithm for Interictal Event Detection 197
David García-Retuerta, Angel Canal-Alonso, Roberto Casado-Vara, Angel Martin-del Rey, Gabriella Panuccio, and Juan M. Corchado

Towards the Reconstruction of the Genome-Scale Metabolic Model of *Lactobacillus acidophilus* La-14 205
Emanuel Cunha, Ahmad Zeidan, and Oscar Dias

Author Index 215



Identification of Antimicrobial Peptides from Macroalgae with Machine Learning

Michela Caprani¹(✉), Orla Slattery¹, Joan O’Keeffe¹, and John Healy²

¹ Marine and Freshwater Research Centre (MFRC),
Galway-Mayo Institute of Technology, Galway, Ireland

michela.caprani@research.gmit.ie, {orla.slattery,joan.okeeffe}@gmit.ie

² Department of Computer Science and Applied Physics,
Galway-Mayo Institute of Technology, Galway, Ireland

john.healy@gmit.ie

Abstract. Antimicrobial peptides (AMPs) are essential components of innate host defense showing a broad spectrum of activity against bacteria, viruses, fungi, and multi-resistant pathogens. Despite their diverse nature, with high sequence similarities in distantly related mammals, invertebrate and plant species, their presence and functional roles in marine macroalgae remain largely unexplored. In recent years, computational tools have successfully predicted and identified encoded AMPs sourced from ubiquitous dual-functioning proteins, including histones and ribosomes, in various aquatic species. In this paper, a computational design is presented that uses machine learning classifiers, artificial neural networks and random forests, to identify putative AMPs in macroalgae. 42,213 protein sequences from five macroalgae were processed by the classifiers which identified 24 putative AMPs. While initial testing with AMP databases positively identifies these sequences as AMPs, an absolute determination cannot be made without *in vitro* extraction and purification techniques. If confirmed, these AMPs will be the first-ever identified in macroalgae.

Keywords: Antimicrobial peptides · Macroalgae · Pseudo Amino Acid Composition (PseAAC) · Machine learning classifiers

1 Introduction

Since the introduction of antibiotics, the development of microbial resistance to conventional antibiotics has progressed, prompting complications for the treatment of infectious disease. Antimicrobial peptides (AMPs, host defense peptides or innate immune peptides) are recognized as an alternative therapeutic agent to address the emergence of resistant strains [1]. AMPs are gene-coded short amino acid sequences (<50 amino acids) that carry a net cationic charge (+2 to +9), with an amphipathic structure [2]. AMPs are further classified by their diverse sequence

composition and secondary structure including α -helical, β -sheet, or extended linear peptides [3]. These features exhibit antimicrobial activity by selectively disrupting microbial membranes, causing cellular death by loss of electrochemical gradient, leakage of contents and disruption of metabolic processes [4].

The marine environment is known to be one of the richest sources of AMPs, identified throughout aquatic life. Additional to AMPs, ubiquitous proteins not previously associated with immunity namely, histone and ribosomal families have shown potent antimicrobial activities from fish, crustacean and mollusks species [5–8]. For decades, marine macroalgae including, Chlorophyta (green), Rhodophyta (red) and Phaeophyta (brown) have been established as sustainable candidate raw materials for generating novel bioactive compounds. Macroalgae are found in intertidal regions that are ubiquitously exposed to diverse chronic stressors such as osmotic stress, excess levels of ultraviolet radiation, salinity, and invasive microbial or pathogenic species, thereby, assisting in the stimulation of innate immune responses [9, 10]. These abiotic factors have contributed to macroalgae development of secondary metabolites, primarily from fatty acids, polysaccharides, or phenolic compounds. Despite their allelopathy and rich protein composition (10% to 50% dry weight), the discovery of innate antimicrobial protein and peptides activity remains limited.

Computational approaches have the capacity to identify encoded AMPs, or antimicrobial compounds from marine macroalgae. Conventionally, *in vitro* purification techniques such as, enzymatic hydrolysis and chromatography approaches have been successful in the isolation of AMPs. However, such methods are costly, laborious and time-consuming. The identification of unannotated AMPs by bioinformatic tools has increased the distribution and evolution of antimicrobial drug discovery by recognising suitable candidates prior to experimental proceedings. Supervised machine learning classifiers such as Support Vector Machines (SVM), Hidden Markov Models (HMM), Random Forests (RF) and Artificial Neural Networks (ANN) have been shown to be effective in mining and identifying AMPs from complex biological data [11]. Computational models can further characterize AMPs by physiochemical descriptors. The Pseudo Amino Acid Composition model (PseAAC) introduced by Chou [12], integrates information effectively through discrete numerical vector analysis by analysing the physiochemical and biochemical properties of protein and peptide sequences.

Challenges in genomic sequencing, decoding protein sequences and analysing genetic features such as protein-coding genes have hindered the understanding of biosynthesis of antimicrobial compounds in macroalgae [13]. Although whole genomes of macroalgae remain functionally unannotated, researchers have decoded the majority of protein sequences from different phyla types including, (red) *Chondrus crispus*, *Palmaria palmata*, *Gracilariopsis lemaneiformis*, and (brown) *Ectocarpus siliculosus* and *Saccharina japonica*. This paper describes an effective method for the identification of AMPs from annotated macroalgae protein sequences by exploiting machine learning classifiers, including random forests and artificial neural networks. The technique presented in this paper was successfully used to identify AMPs for each of the above macroalgae. If confirmed, these AMPs will be the first ever identified in a macroalgae.

2 Methods

Protein sequences from five macroalgae were downloaded from the National Centre of Bioinformatic Information (NCBI) database. These annotated seaweed protein sequences included the brown macroalgae, *E.siliculosus* (24,202 sequences) and *S.japonica* (838 sequences), and the red macroalgae, *C.crispus* (15,320 sequences), *P.palmata* (937 sequences) and *G.lemaneiformis* (886 sequences). These sequences were augmented with proteins obtained from the Uniprot database including 13,492 uncharacterized protein sequences and 37 histone sequences.

2.1 Training and Testing Datasets

The training and testing datasets used in this study were extracted from sequences taken from the Antimicrobial Peptide Database (APD) [14], the AntiBP Server [15] and Uniprot. The positive AMP dataset consisted of 2,338 APD sequences and 1,209 from the AntiBP Server. The latter set consists of processed and clipped APD protein sequences and provided the training data with a different representation of some of the positive AMPs.

The negative dataset of 6,332 sequences consisted of 1,207 taken from the AntiBP Server and 6,332 from Uniprot. Using the approach described by Veltri *et al.* [16] and the UniProt Consortium [17], the negative Uniprot dataset was filtered for sequences with experimentally validated cytoplasmic localization and excluded any with antimicrobial, antibiotic, antiviral or antifungal properties. The combined positive and negative AMP dataset contained a total of 11,086 sequences.

2.2 Classifier Configuration

The Statistical Machine Intelligence and Learning Engine (Smile v1.5.2) [18] was utilised to create neural network and random forest classifiers for the identification of AMPs in the macroalgae protein sequences.

The neural network topology consisted of an input layer of 200 nodes, a single hidden layer and an output layer of two nodes. Although deep neural networks have become increasingly popular in bioinformatics [19], one and two hidden layers have been shown to be sufficient for any continuous and discontinuous function respectively [20]. We used the formula $N_h = \frac{|D|}{\alpha \times (N_i + N_o)}$ from Hagan *et al.* [21] to compute the number of nodes, N_h , in the hidden layer, where $|D|$ is the size of the training dataset, N_i and N_o are the number of nodes in the input and output layers respectively, and α is a scaling value in the range [0..2]. This formula helps prevent overfitting by limiting the number of free parameters in the neural network architecture to a small portion of the dimensions that exist in the training dataset.

We capped the number of decision trees in the random forest at 128, as this recommended figure represents the upper limit before diminishing returns reported by Oshiro *et al.* [22].

The training and testing of the classifiers and the subsequent analysis of macroalgae protein sequences was undertaken on an OSX 10.13.2 platform, with a 1.8 GHz Intel Core i7 processor, 16 GB of RAM and an instance of the OpenJDK 12.0.1 64-bit Java Virtual Machine.

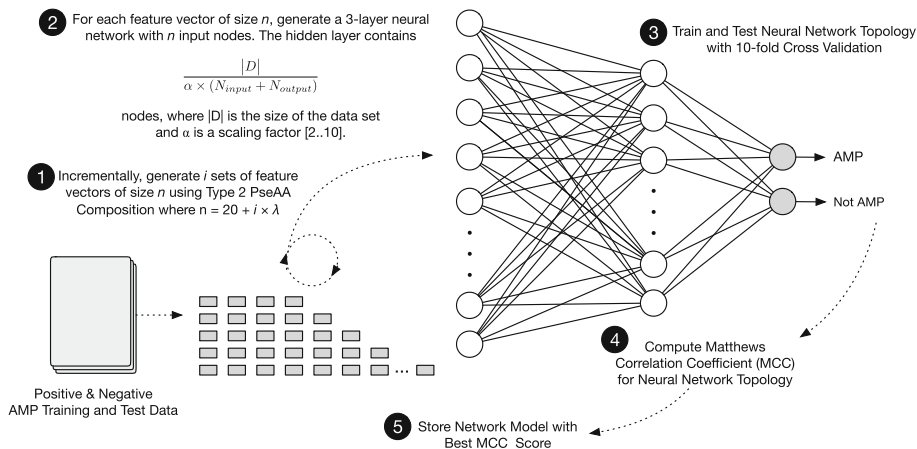


Fig. 1. Feature selection and training of the neural network. A similar approach was used to train the random forest.

2.3 Incremental Feature Selection

Selecting a feature set that is simultaneously informative and discriminating is essential for training a classifier with a high degree of accuracy. We selected 36 features from the AAindex [23,24] that are highly correlated with AMPs. The AAindex contains a matrix of published physicochemical and biochemical properties of amino acids and amino acid pairs that can be easily mapped to an input vector for a classifier.

In order to exploit their amphiphilic and other salient properties, the variable-length protein sequences used in this study were transformed into fixed-sized feature vectors from their Type 2 Pseudo Amino Acid Composition (PseAAC) [12] with $\lambda = 5$ and a weight $w = 0.05$. For an amino acid feature set of size n , a protein sequence will contain $20 + n \times \lambda$ Type 2 PseAAC features, regardless of the number of amino acid residues in the sequence. In addition to exploiting the amphiphilic relationships between its amino acid elements, PseAAC enables protein sequences of different sizes to be translated into the fixed-length vectors required by most machine learning classifiers. In this absence of this, a more complex recursive neural network would be required instead of the multilayer perceptron used for this research.

Employing a similar approach to that described by Wang *et al.* [14] and Li *et al.* [25], starting with just 4 features, the feature set was iteratively expanded. For each feature set expansion, a neural network and random forest topology was

generated to match the feature set size and then trained, tested and scored. AMP properties in the ranked feature list were added one-by-one in order to determine an optimal feature set, i.e. a new feature set was constructed when one new feature was added. Consequently, a total of 31 feature sets were created, with the minimal set of 4 features generating a PseAAC vector of length $20 + 4 \times 5 = 40$ and the upper limit of 36 features generating a $20 + 36 \times 5 = 200$ length input vector. Both the neural network and random forest were trained and tested using 10-fold cross validation.

The accuracy, sensitivity and specificity of each trained neural network and random forest configuration was calculated using the Matthews Correlation Coefficient (MCC) shown below:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (1)$$

where TP, TN, FP and FN denote true positive, true negative, false positive and false negative respectively. The MCC metric has a range of $[-1...1]$, where -1 indicates an incorrect binary classifier and +1 an entirely correct classifier. The MCC is regarded as a more balanced measure where there exists a significant discrepancy between the cardinality of each class in a dataset [26].

3 Results and Discussion

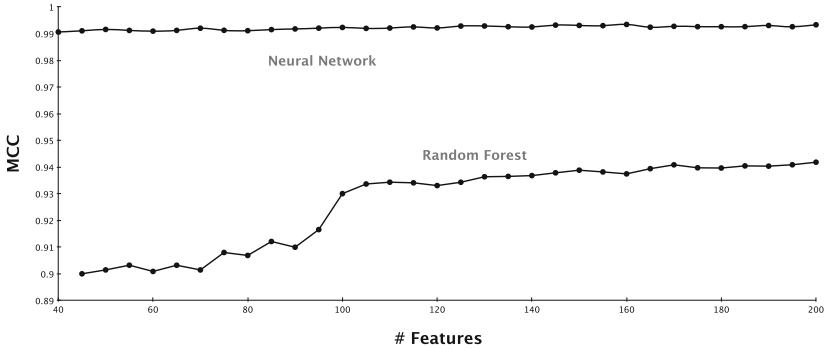


Fig. 2. MCC scores v/s Feature number

Figure 2 depicts the MCC score for the neural network and random forest classifiers as each AMP feature taken from the AAindex [23,24] was added to the feature set. In contrast with the random forest classifier, the overall performance of the artificial neural network was consistently good in terms of accuracy, sensitivity, and specificity with a smaller number of features. The model reached

Table 1. Homology of 3 peptides from the 230 Mbps *P.palmata*.

<i>Palmaria palmata</i> : 3 AMP sequences	AMP Homology	Type	AMP DB
MKVRASVKKMCDKCRVIRRHRRVMVICINPKHKQRQG	50%	Defensin (<i>A. aurita</i>)	CAMP
MVEPLLSGIVLGLIPITIIAGLLVAAYIQYRRGNQFGL	35%	β -defensin (<i>B. Taurus</i>)	CAMP
MPAIQQLVRLPRQKAVKKTSPALKACPQRRGVCTRV YTTTPKPKNSALRKYVESITSG	32%	Ubiquicidin (<i>H. sapiens</i>)	APD

the maximum value of 0.9933 when 36 features were added to the neural network. The random forest model also produced its optimal MCC value of 0.9418 after all 36 features were added. The high MCC scores confirmed the decision to restrict the training and testing dataset sets to amino acid features that are highly correlated to AMPs.

Tables 1, 2, 3, 4 and 5 show the AMPs identified by the classifiers after protein sequences from *P.palmata*, *C.crispus*, *E.siliculosus*, *S.japonica* and *G.lemanefromis* were processed. Each of the putative AMPs listed were positively identified by both the neural network and random forest. Exploration of AMPs from marine algae is highlighted as an important alternative to mammal and invertebrate antimicrobial compounds. The results reported in this study indicate that the classification method can be very effective at identifying AMPs from selected macroalgae species. The iterative feature selection technique and construction of the Type 2 PseAAC feature vector, coupled with simplified neural network and random forest topologies, identified 24 putative AMPs from the set of macroalgae. Red algae species *C.crispus* was the most abundant source, containing 12 AMPs. The model identified 3 AMPs in each of the species, *P.palmata*, *E. siliculosus*, *S. japonica* and *G.lemanefromis*.

In order to determine the origin of the AMPs identified from the macroalgae protein data, each sequence was subject to a BLAST [28] search. This application determined that the AMPs were sourced from ribosomal subunits and uncharacterized protein regions. Three AMPs from the *S. japonica* species were identified as the 50S ribosomal proteins L19, L34 and L36. The 3 AMPs identified from *P. palmata* were characterized as the 50S ribosomal proteins L14, L34 and L36. Furthermore, the 3 AMPs from *G. lemanefromis* were classified as 30S and 50S ribosomal subunits S5, L34, and L36. While ribosomal proteins are known to possess dual-functioning properties in assembly and protein translation, it is also evident that such protein types putatively play an essential role in innate host defense. Previous studies have successfully identified and purified several AMPs derived from ribosomal protein subunits, displaying a range of antimicrobial capacities in various pathogens and bacterial species [5, 8, 27].

Furthermore, to determine homology with known AMPs, the macroalgal AMP sequences were compared with the Collection of Antimicrobial Peptide (CAMP) [29] and Antimicrobial Peptide Database [14] using a sequence alignment search. The AMPs from each macroalgal species contained partial homology with defensins and histone-derived AMPs. For example, the peptide sequence

Table 2. Homology of 12 peptides from the 106.4 Mbps *C. crispus*.

<i>Chondrus crispus</i> AMP sequences	AMP Homology	Type	AMP DB
MSKHARPCCLKGGPEA	35%	H6-Histone (<i>O. mykiss</i>)	APD
MVLRRIILTVVFRSRVCATTRCLLQICVTIRLLLL	32%	Hyposin-H3 (<i>P. hypochondrialis</i>)	APD
MERHMGDLDNSMPRSTRKTLPENGSILTSTMTTCGTN	30%	Cryptdin-1 (<i>M. musculus</i>)	APD
MSIEIPTGATKSSNFWCRSKNRNQISRIWFGWSLFDY P	33%	Defensin (<i>P. hamadryas</i>)	CAMP
MPSPNSANVGVLHRAALMSRALCTSRTGSGAGREHKR QKRT	35%	Tenecin (defensin) (<i>T. molitor</i>)	1 APD
MWGRIIALHGNGHVRAKFRNQLPNSIGKGVVRVMLY PSTI	31%	β -defensin 13 (<i>B. Taurus</i>)	APD
MHAVVGILDRRETLVISRQIPHRCTFGGKPFSTNRT CTTGLSRVIKQRLSEPN	34%	β -defensin 11 (<i>B. Taurus</i>)	APD
MVPSLPTSRIVKKIATEPQSIIVEGRSLCVGMLAATGT IVQCRRMISKNPACHNCL	33%	AdDLP (defensin) (<i>A. dehalogenans</i>)	CAMP
MSVCNDKQCQSYIGYFCKFVTFEFGRCAPVDAVLAPCR KHPQLPLCKNLCTCHLAKATQLYEAPLCRLLS	31%	Gallinacin 6 (β - defensin) (chicken)	CAMP
MVIVNRGCCSDLNRPRWHSGACSSLYLPLSESLSLP LLVAVCRLLSLVFSQRGRFASTLGVANCCCGVGT SCV	32%	gcDefb1 (β - defensin) (<i>C. Idella</i>)	CAMP
MQSNSLPQRLPHVINAVMFAIQGLTAALGPGLCSSTS CKGYFILPGKYGKYTGHEYHIGFTFTHKRVLSRSIQAC DVPCSRKTTTNTGQNDTNAQRF	29%	Apl-AvBD16 (β -defensin) (<i>A. platyrhynchos</i>)	CAMP
MWGRIIAPHGNGHVRAKFRNQLPNSIGKGVVRVMLY PNAI	41%	Buforin I Histone H2A (<i>B. gargarizans</i>)	APD

from *P. palmata* had a 50% identity to *Aurelia aurita* (Moon jellyfish) defensin seen in Fig. 3. Moreover, AMPs from other the algal species showed a 29–35% identity to peptides derived from the defensin family. These results show that, analogous to most kingdoms, macroalgae have likely accumulated defensins to play a key role in their immune defense against their extrinsic surroundings [30]. In addition to defensins, histone-derived AMP homology was observed in a number of sequences. The results determined that the AMPs from *C. crispus* possess a 35% sequence similarity to histone H6 from *Oncorhynchus mykiss* (Rainbow trout), a 32% similarity to Hyposin-H3 from *Phyllomedusa hypochondrialis* (Leaf frog) and 41% similarity to histone H2A Buforin from *Bufo bufo gargarizans* (Asiatic frog). Previous studies have shown that histone-derived AMPs isolated from both marine and terrestrial species have potent antimicrobial activity [7, 31, 32].

Table 3. Homology of 3 peptides from the 198.4 Mbps *E.siliculosus*.

<i>Ectocarpus siliculosus</i> AMP sequences	AMP homology	Type	AMP DB
MGGFYGWQLSACWWRSGCAPATW	32%	β -defensin 3 (<i>C.floridanus</i>)	APD
MRTAWRNTCAPPERSRPWLPGSGRTVTHPVARRRCAGL SEISWKHPVRARSWCALGRTQTTS	36%	Penaeidin-3a (<i>P.vannamei</i>)	APD
MVLYRQAANTVERWMGIRARTHMRCAVLATAAFVKAN TWIRDYHRPSAHVRQKYPNGNHGGS	29%	WB Piscidin 5 (<i>M.chrysops</i>)	CAMP

Table 4. Homology of 3 peptides from the 551.5 Mbps *S.japonica*.

<i>Saccharina japonica</i> AMP sequences	AMP homology	Type	AMP DB
VPALLAFRLGKTLYS	48%	H4-(86–100) Histone (<i>Rat</i>)	ADP
MKVRASVKKMCEKCRIRRHGRVQVICTNLKHKQRQG	33%	β -defensin 10 (<i>B.Taurus</i>)	CAMP
MTKRTLGGTNRKVIASVGFARMKTKQGCKVINRRR KKRKNLSI	21%	β -defensin (<i>S.salar</i>)	CAMP

Table 5. Homology of 3 peptides from the 91.2 Mbps *G.lemaneaformis*.

<i>Gracilariopsis lemaneaformis</i> AMP sequences	AMP homology	Type	AMP DB
MKVRASVKKKCDKCRIRRHRRKVIHCENAKHKQRQG	35%	Cryptdin-5 α -defensin (<i>M.musculus</i>)	APD
MSQGIKNGTNRKQIKKSGFRARMSTYSGRKIINLRRR KRRKKIVL	31%	rhesus θ - defensin-1 (<i>R.Macaque</i>)	APD
MKSVITVISAADAAGRFPTSSDLESVQGNIQRAAAR LEAAEKLDNHEAVVKEAG	33%	Apl-AvBD16 β - defensin (<i>A.platyrrhynchos</i>)	APD

Score = 20.0 bits (40), Expect = 2.3
Identities = 7/14 (50%), Positives = 10/14 (71%), Gaps = 0/14 (0%)

```
Query 1  MKVRASVKKMCDKC 14
          +K+RA+ KK C C
Sbjct 27  VKLRANCKKTCGLC 40
```

Fig. 3. Sequence alignment of identified *P.palmata* AMP with 50% homology with *Aurelia aurita* (Moon jellyfish) defensin.

In summary, the computational results suggest macroalgal immunity involves the use of ribosome and histone-derived AMPs and indicates their potential use of defensin-like AMPs. A range of *in vitro* methods such as chemical and enzymatic extractions, as well as chromatographic techniques namely, ion exchange, and High-Performance Liquid Chromatography (HPLC), have previously been utilized to isolate and characterize such AMP types [8, 33, 34]. Future work in relation to this study will utilize the sequence information to make informed decisions regarding the isolation and further characterization of these putative AMPs from the various macroalgal species. Therefore, the *in silico* model holds a strong potential to become a useful tool to identify novel AMPs prior to experimental processes.

4 Conclusion

In this study, we implemented a machine learning approach using artificial neural network and random forest models for the identification of AMPs from five macroalgae species. The approach identified 24 putative AMPs from the collected macroalgae protein sequences. The AMPs were derived from ribosomal subunits and uncharacterized protein, sharing regions of similarity to defensin and histone AMP families. It is possible that the identified AMPs may play a significant role in macroalgal host defense. If isolated by *in vitro* applications, these will be the first-ever identified AMPs from macroalgae. Consequently, this method can then be applied to organisms where AMP identity remains unknown.

Acknowledgements. This work is supported by a grant from the Enterprise Partnership Scheme, the Ireland Research Council (IRC) and This is Seaweed (<https://thisisseaweed.com>).

References

1. Zasloff, M.: Antimicrobial peptides of multicellular organisms. *Nature* **415**(6870), 389–395 (2002)
2. Hancock, R.E., Lehrer, R.: Cationic peptides: a new source of antibiotics. *Trends Biotechnol.* **16**(2), 82–88 (1998)
3. Powers, J.P.S., Hancock, R.E.: The relationship between peptide structure and antibacterial activity. *Peptides* **24**(11), 1681–1691 (2003)
4. Bahar, A.A., Ren, D.: Antimicrobial peptides. *Pharmaceuticals* **6**(12), 1543–1575 (2013)
5. Fernandes, J.M., Smith, V.J.: A novel antimicrobial function for a ribosomal peptide from rainbow trout skin. *Biochem. Biophys. Res. Commun.* **296**(1), 167–171 (2002)
6. Fernandes, J.M., Kemp, G.D., Molle, M.G., Smith, V.J.: Anti-microbial properties of histone H2A from skin secretions of rainbow trout, *Oncorhynchus mykiss*. *Biochem. J.* **368**(2), 611–620 (2002)
7. Patat, S.A., Carnegie, R.B., Kingsbury, C., Gross, P.S., Chapman, R., Schey, K.L.: Antimicrobial activity of histones from hemocytes of the Pacific white shrimp. *Eur. J. Biochem.* **271**(23–24), 4825–4833 (2004)