

Lecture Notes in Networks and Systems 553

Florentino Fdez-Riverola ·

Miguel Rocha ·

Mohd Saberi Mohamad ·

Simona Caraiman ·

Ana Belén Gil-González *Editors*

Practical Applications of Computational Biology and Bioinformatics, 16th International Conference (PACBB 2022)

 Springer

Lecture Notes in Networks and Systems

Volume 553

Series Editor

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

Advisory Editors

Fernando Gomide, Department of Computer Engineering and Automation—DCA,
School of Electrical and Computer Engineering—FEEC, University of
Campinas—UNICAMP, São Paulo, Brazil

Okay Kaynak, Department of Electrical and Electronic Engineering,
Bogazici University, Istanbul, Turkey

Derong Liu, Department of Electrical and Computer Engineering, University of
Illinois at Chicago, Chicago, USA

Institute of Automation, Chinese Academy of Sciences, Beijing, China

Witold Pedrycz, Department of Electrical and Computer Engineering, University of
Alberta, Alberta, Canada

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

Marios M. Polycarpou, Department of Electrical and Computer Engineering,
KIOS Research Center for Intelligent Systems and Networks, University of Cyprus,
Nicosia, Cyprus

Imre J. Rudas, Óbuda University, Budapest, Hungary

Jun Wang, Department of Computer Science, City University of Hong Kong,
Kowloon, Hong Kong

The series “Lecture Notes in Networks and Systems” publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Florentino Fdez-Riverola · Miguel Rocha ·
Mohd Saberi Mohamad · Simona Caraiman ·
Ana Belén Gil-González
Editors

Practical Applications of Computational Biology and Bioinformatics, 16th International Conference (PACBB 2022)

 Springer

Editors

Florentino Fdez-Riverola
Computer Science Department
Universidad de Vigo
Vigo, Spain

Miguel Rocha
Campus de Gualtar
Universidade do Minho
Braga, Portugal

Mohd Saberi Mohamad
College of Medicine and Health Sciences
United Arab Emirates University
Al Ain, Abu Dhabi, United Arab Emirates

Simona Caraiman
Gheorghe Asachi Technical University
of Iași
Iași, Romania

Ana Belén Gil-González 
Edificio I+D+i
University of Salamanca
Salamanca, Spain

ISSN 2367-3370

ISSN 2367-3389 (electronic)

Lecture Notes in Networks and Systems

ISBN 978-3-031-17023-2

ISBN 978-3-031-17024-9 (eBook)

<https://doi.org/10.1007/978-3-031-17024-9>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2023

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

The success of bioinformatics in recent years has been prompted by research in molecular biology and molecular medicine in several initiatives. These initiatives gave rise to an exponential increase in the volume and diversification of data, including nucleotide and protein sequences and annotations, high-throughput experimental data, biomedical literature, among many others. Systems biology is a related research area that has been replacing the reductionist view that dominated biology research in the last decades, requiring the coordinated efforts of biological researchers with those related to data analysis, mathematical modelling, computer simulation and optimization.

The accumulation and exploitation of large-scale databases prompt new computational technology and for research into these issues. In this context, many widely successful computational models and tools used by biologists in these initiatives, such as clustering and classification methods for gene expression data, are based on computer science/artificial intelligence (CS/AI) techniques. In fact, these methods have been helping in tasks related to knowledge discovery, modelling and optimization tasks, aiming at the development of computational models so that the response of biological complex systems to any perturbation can be predicted. The 16th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB) aims to promote the interaction among the scientific community to discuss applications of CS/AI with an interdisciplinary character, exploring the interactions between sub-areas of CS/AI, bioinformatics, chemoinformatic and systems biology. The PACBB'22 technical programme includes ten papers of authors from many different countries (Bahrain, Canada, France, Italy, Portugal, Saudi Arabia, Spain and UK) and different subfields in bioinformatics and computational biology. All papers underwent a peer review selection: each paper was assessed by three different reviewers from an international panel composed of about 46 members from 11 countries. The quality of submissions was on average good, with an acceptance rate of approximately 60% (10 accepted papers from 15 submissions).

There will be special issues in JCR-ranked journals, such as Interdisciplinary sciences: mathematical biosciences and engineering, 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 (Italy) with the collaboration of the United Arab Emirates University, the University of Minho, the University of Vigo, the University of Salamanca and the Gheorghe Asachi Technical University of Iași. We would like to thank all the contributing authors, the members of the programme committee and the sponsors. 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 for their valuable work, which is essential for the success of PACBB'22.

Vigo, Spain
Braga, Portugal
Al Ain, United Arab Emirates
Iași, Romania
Salamanca, Spain

Florentino Fdez-Riverola
Miguel Rocha
Mohd Saberi Mohamad
Simona Caraiman
Ana Belén Gil-González

Acknowledgements

Sponsors

  Department of Informatics & Science and Communication Engineering

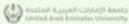
  MESVA - Department of Medical Clinics, Diagnostic and Therapeutic Sciences and Health Services

Organizers

  **Universidade de Vigo**

 United Arab Emirates University

 **Universidade de Minho**

 **UNIVERSIDAD DE SALAMANCA**

 **ISTITUTO ITALIANO DI TECNOLOGIA**

Support from National Associations

Organization

Program Committee Chairs

Mohd Saberi Mohamad, United Arab Emirates University, United Arab Emirates
Miguel Rocha, University of Minho, Portugal

Organising Committee Chairs

Florentino Fdez-Riverola, University of Vigo, Spain
Ana Belén Gil-González, University of Salamanca, Spain
Simona Caraiman, Gheorghe Asachi Technical University of Iași, Romania

Advisory Committee

Grabriella Panuccio, Istituto Italiano di Tecnologia, Italy

Local Organizing Committee

Pierpaolo Vittorini (Co-chair), University of L'aquila, Italy
Tania Di Mascio (Co-chair), University of L'aquila, Italy
Federica Caruso, University of L'Aquila, Italy
Anna Maria Angelone, University of L'Aquila, Italy

Organizing Committee

Juan M. Corchado Rodríguez, University of Salamanca, Spain; AIR Institute, 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
Liliana Durón, 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
Alfonso González Briones, University of Salamanca, Spain
Yeray Mezquita Martín, University of Salamanca, Spain
Beatriz Bellido, University of Salamanca, Spain
María Alonso, University of Salamanca, Spain
Sergio Marquez, University of Salamanca, Spain
Marta Plaza Hernández, University of Salamanca, Spain
Guillermo Hernández González, AIR Institute, Spain
Ricardo S. Alonso Rincón, University of Salamanca, Spain
Raúl López, University of Salamanca, Spain
Sergio Alonso, University of Salamanca, Spain
Andrea Gil, University of Salamanca, Spain
Javier Parra, University of Salamanca, Spain

Programme Committee

Vera Afreixo, University of Aveiro, Portugal
Manuel Álvarez Díaz, University of A Coruña, Spain
Joel P. Arrais, University of Coimbra, Portugal
Carlos Bastos, University of Aveiro, Portugal
Lourdes Borrajo, University of Vigo, Spain
Ana Cristina Braga, University of Minho, Portugal
Rui Camacho, University of Porto, Portugal
Angel Canal, Universidad de Salamanca, Spain
Yingbo Cui, National University of Defense Technology, China
Sergio Deusdado, IPB-Polytechnic Institute of Bragança, Portugal
Oscar Dias, University of Minho, Portugal
Nuno Filipe, University of Porto, Portugal
Dino Franklin, Federal University of Uberlandia, Brazil
Narmer Galeano, Universidad Catolica de Manizales, Colombia

Rosalba Giugno, University of Verona, Italy
Gustavo Isaza, University of Caldas, Colombia
Paula Jorge, IBB - CEB Centre of Biological Engineering, Portugal
Rosalia Laza, Universidad de Vigo, Spain
Thierry Lecroq, University of Rouen, France
Giovani Librelotto, Universidade Federal de Santa Maria, Brazil
Filipe Liu, Data Science and Learning Division, Argonne National Laboratory, USA
Hugo López Fernández, Instituto de Investigación e Inovação em Saúde (i3S), Spain
Eva Lorenzo Iglesias, University of Vigo, Spain
Gonçalo Marques, Polytechnic of Coimbra, Portugal
Mohd Saberi Mohamad, United Arab Emirates University, United Arab Emirates
Loris Nanni, University of Padua, Italy
José Luis Oliveira, University of Aveiro, Portugal
Vitor Pereira, University of Minho, Portugal
Armando Pinho, University of Aveiro, Portugal
Ignacio Ponzoni, CONICET, Argentina
Miguel Reboiro-Jato, University of Vigo, Spain
Jose Ignacio Requeno, University of Zaragoza, Spain
João Manuel Rodrigues, DETI/IEETA, University of Aveiro, Portugal
Gustavo Santos García, Universidad de Salamanca, Spain
Ana Margarida Sousa, University of Minho, Portugal
Niclas Ståhl, University of Skövde, Sweden
Carolyn Talcott, SRI International, USA
Rita Margarida Teixeira Ascenso, ESTG–IPL, Portugal
Antonio J. Tomeu-Hardasmal, University of Cadiz, Spain
Eduardo Valente, IPCB, Portugal
Alejandro F. Villaverde, Instituto de Investigaciones Marinas (CSIC), Spain
Pierpaolo Vittorini, University of L'Aquila, Italy

Contents

TooT-BERT-T: A BERT Approach on Discriminating Transport Proteins from Non-transport Proteins	1
Hamed Ghazikhani and Gregory Butler	
Machine Learning and Deep Learning Techniques for Epileptic Seizures Prediction: A Brief Review	13
Marco Hernández, Ángel Canal-Alonso, Fernando de la Prieta, Sara Rodríguez, Javier Prieto, and Juan Manuel Corchado	
The Covid-19 Decision Support System (C19DSS) – A Mobile App	23
Pierpaolo Vittorini, Nicolò Casano, Gaia Sinatti, Silvano Junior Santini, and Clara Balsano	
Towards a Flexible and Portable Workflow for Analyzing miRNA-Seq Neuropsychiatric Data: An Initial Replicability Assessment	31
Daniel Pérez-Rodríguez, Mateo Pérez-Rodríguez, Roberto C. Agís-Balboa, and Hugo López-Fernández	
The NAD Interactome, Identification of Putative New NAD-Binding Proteins	43
Sara Duarte-Pereira, Sérgio Matos, José Luís Oliveira, and Raquel M. Silva	
Multiple Instance Learning Based on Mol2vec Molecular Substructure Embeddings for Discovery of NDM-1 Inhibitors	55
Thomas Papastergiou, Jérôme Azé, Sandra Bringay, Maxime Louet, Pascal Poncelet, and Laurent Gavara	
Towards Improving Bio-Image Segmentation Quality Through Ensemble Post-processing of Deep Learning and Classical 3D Segmentation Pipelines	67
Anuradha Kar	

Exploring *Xylella fastidiosa*'s Metabolic Traits Using a GSM Model of the Phytopathogenic Bacterium 79
Alexandre Oliveira, Emanuel Cunha, Miguel Silva, Cristiana Faria, and Oscar Dias

Genomic Regions with Atypical Concentration of Inverted Repeats 89
Carlos A. C. Bastos, Vera Afreixo, João M. O. S. Rodrigues, and Armando J. Pinho

EvoPPI 2: A Web and Local Platform for the Comparison of Protein–Protein Interaction Data from Multiple Sources from the Same and Distinct Species 101
Miguel Reboiro-Jato, Jorge Vieira, Sara Rocha, André D. Sousa, Hugo López-Fernández, and Cristina P. Vieira

Author Index 111

TooT-BERT-T: A BERT Approach on Discriminating Transport Proteins from Non-transport Proteins



Hamed Ghazikhani and Gregory Butler

Abstract Transmembrane transport proteins (transporters) serve a crucial role for the transport of hydrophilic molecules across hydrophobic membranes in every living cell. The structures and functions of many membrane proteins are unknown due to the enormous effort required to characterize them. This article proposes TooT-BERT-T, a technique that employs the BERT representation to analyze and discriminate between transporters and non-transporters using a Logistic Regression classifier. Additionally, we evaluate frozen and fine-tuned representations from two different BERT models. Compared to state-of-the-art prediction methods, TooT-BERT-T achieves the highest accuracy of 93.89% and MCC of 0.86.

Keywords Transmembrane transport proteins · Machine learning · BERT · Language model · Transformers · Neural network

1 Introduction

Around one-third of the proteins in a cell are found in its membrane, and approximately one-third of these proteins are involved in molecule transport [21]. *Transmembrane transport proteins*, also known as *transporters*, are required for cell metabolism, ion homeostasis, signal transduction, binding with small molecules in the extracellular space, immune recognition, energy transduction, and physiological and developmental processes [21].

Protein research has advanced our knowledge of human health and disease treatment. The decreasing cost of sequencing technology has enabled the generation of

H. Ghazikhani (✉) · G. Butler

Department of Computer Science and Software Engineering, Concordia University,
Montreal, Canada
e-mail: hamed.ghazikhani@concordia.ca

G. Butler

e-mail: gregory.butler@concordia.ca

G. Butler

Centre for Structural and Functional Genomics, Concordia University, Montreal, Canada

massive datasets of naturally occurring proteins with enough information to build sophisticated machine learning models of protein sequences [23].

Since proteins, like human languages, are denoted by string concatenation, we can apply natural language processing (NLP) approaches [18]. Transformer neural networks (*Transformers*) have contributed significantly to the field of natural language processing [22]. Autoencoders, for example, BERT (Bidirectional Encoder Representations from Transformers) [9], are stacking models that are trained by corrupting input tokens and attempting to recover the original sentence [11]. While they can generate text as well, they are typically used to create vector representations for future tasks such as classification [11].

A massive collection of protein sequences from UniProt Archive (UniParc) [14] and the Big Fantastic Database (BFD) [11, 13] comprising over 390 billion amino acids resulted in *ProtTrans* [10], an amazing adaption to the protein domain of six available Transformer topologies which are Transformer-XL, BERT, Albert, XLnet, T5, and Electra.

TooT-BERT-T proposes a method for discriminating transport proteins from non-transport proteins using representations from ProtBERT-BFD and Logistic Regression. Our investigation can be summarised as follows: 1) Using ProtBERT-BFD to discriminate between transport and non-transport proteins for the first time. 2) Evaluation of frozen/fine-tuned ProtBERT-BFD representations. 3) Evaluation of frozen/fine-tuned *MembraneBERT* representations. 4) The fine-tuned *TransporterBERT* is a publicly accessible model pre-trained on the BFD database and fine-tuned using the transport proteins dataset (<https://huggingface.co/ghazikhanihamed/TransporterBERT>). 5) Proposing TooT-BERT-T as a method for classifying transport proteins that outperforms all other approaches.

The following is the outline for the paper: Sect. 2 describes the related work. Section 3 contains information about the dataset and experimental design used in this study. Section 4 compares and analyses the outcomes of TooT-BERT-T and Sect. 5 brings the paper to a close.

2 Related Work

Aplop and Butler [4, 5] provide a comprehensive overview of transport protein prediction methods. Earlier efforts used experimentally characterized databases to conduct homology searches for novel transporters. For example, TransATH [5] automates the Saier’s protocol via sequence similarity. TransATH improves transmembrane segment computations by including subcellular localization and claims an overall accuracy of 71.0%.

TrSSP (Transporter Substrate Specificity Prediction Service) [16] was developed to predict the substrate category of membrane transport proteins in an attempt to overcome the limitations of homology methods. The TrSSP tool predicts top-level transporters with an accuracy of 78.99 and 80.00%, respectively, and an MCC of 0.58 and 0.57 on the cross-validation and independent test sets.

SCMMTP [15] makes use of a novel scoring card method (SCM) to ascertain the dipeptide composition of potential membrane transport proteins. SCMMTP begins with a 400-dipeptide starting matrix and scores dipeptides based on the difference between positive and negative compositions. Following that, the matrix is optimized using a genetic algorithm. SCMMTP achieved an overall accuracy of 81.12% and 76.11% and an MCC of 0.62 and 0.47, respectively, on the training and independent datasets.

Nguyen et al. [17] characterize transporter protein sequences using a word-embedding technique. The protein sequence is defined by the word embedding and the protein's biological terms frequency. They achieved accurate results in terms of transporter substrate specificity but not in terms of transporter detection. When cross-validation was used, the prediction accuracy for transporters was only 83.94 and 85.00% using the independent dataset.

In 2020, Alballa and Butler developed TooT-T [2], an ensemble technique that combines the results of two distinct approaches: homology annotation transfer and machine learning. BLAST searches the Transporter Classification Database (TCDB) [20] for homology to a query protein. If a query meets three thresholds, it is projected as a transporter. It also computes three composition features for training their respective SVM models. Finally, the meta-model assigns a protein the transport protein classification. They claim accuracy of 90.07% and 92.22%, respectively, and MCC values of 0.80 and 0.82 for the cross-validation and independent test sets, respectively. While incorporating multiple feature sets and classifiers improves the classification of transport proteins in TooT-T, it also increases the task's complexity.

3 Materials and Methods

3.1 Dataset

This work utilizes the dataset from the TrSSP project [16] which can be accessed at the following URL: <https://www.zhaolab.org/TrSSP/>. The dataset was created using the UniProt database [14], in which 10,780 transporter, carrier, and channel proteins were initially well characterized at the protein level with different substrate specificity annotation. Mishra et al. [16] eliminated from this benchmarking dataset fragmented sequences, sequences with more than two substrate specificities, and biological function annotations based only on sequence similarity. As presented in Table 1 the final dataset contains 1,560 protein sequences for the training and test sets. This dataset is referred to as DS-T, which stands for a dataset for transporter proteins.

Table 1 DS-T: transport proteins dataset

Class	Training	Test	Total
Transporter	780	120	900
Nontransporter	600	60	660
Total	1,380	180	1,560

3.2 Protein Sequence Representation

As multiple studies demonstrate, representation learning, a branch of machine learning in which the representation is estimated concurrently with the statistical model, is gaining traction in biology. Works [3, 6, 19] highlight how representations can assist in extracting crucial biological information from the millions of observations collected by modern sequencing technologies [8].

BERT (Bidirectional Encoder Representations from Transformers) [9] is a language model used in natural language processing that employs a multi-layer bi-directional Transformer encoder that employs an attention mechanism in each encoder layer to condition both left and right context and process all words in the sentence in parallel. Each encoder layer comprises two sub-layers: multi-head self-attention and feed-forward neural networks. While encoding a specific word, the multi-head self-attention sublayer assists the encoder in looking at other words in the input sentence. The following formula is used to compute the scaled dot-product attention sublayer [22]:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_n)W^o \quad (1)$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

where Q (Query), K (Key) and V (Value) are various linear transformations of the input features in order to obtain information representations for various subspaces. The dimension of K is d_k and W_i^Q , W_i^K , W_i^V and W_i^O are weight matrices.

BERT is a two-step framework: pre-training and fine-tuning. Pre-training is training the model on a large amount of unlabeled data in an unsupervised manner. In contrast, fine-tuning is the process of initializing the model with the pre-trained parameters and fine-tuning all parameters using labeled data from downstream tasks via an additional classifier [9].

There are two methods for extracting representations from pre-trained BERT models: (i) frozen and (ii) fine-tuned. The former extracts features from a pre-trained BERT model without updating the model's weights, whereas the latter extracts

features after training the pre-trained BERT model on a smaller dataset and fine-tuning the model’s weights [9].

ProtBERT-BFD [10] is the BERT model which has been pre-trained on a large corpus of protein sequences from the BFD database (<https://bfd.mmseqs.com>) which contains 2.5 billion protein sequences. MembraneBERT is ProtBERT-BFD fine-tuned using the TooT-M membrane proteins dataset [1]. MembraneBERT can be found at (<https://huggingface.co/ghazikhanihamed/MembraneBERT>).

The representations from the final hidden layer of ProtBERT-BFD and MembraneBERT models are used in conjunction with a mean-pooling strategy, which is concluded to be the optimal method in ProtTrans [10].

3.3 Fine-Tuning a BERT Model

We add a classification layer and train the entire BERT model on the transporters training set to fine-tune a BERT model. We randomly chose 10% of the training samples as the validation set in this study. The downstream task dataset will update all initialized weights from pre-training during the fine-tuning phase. We fine-tuned the BERT models using the *Trainer API* from HuggingFace [24]. This is a preliminary investigation of BERT’s role in transport protein analysis, so we used the same hyperparameter settings as ProtTrans [10], except for the empirically determined number of training epochs of 13 for ProtBERT-BFD and 10 for MembraneBERT. We discovered these numbers when we have the maximum performance throughout the validation set results. Additional hyperparameters for fine-tuning are listed in Table 2 which are recommended and used in ProtTrans project.

Table 2 Fine-tuning ProtBERT-BFD and MembraneBERT hyperparameters

Hyperparameter	Value
Training batch size	1
Evaluation batch size	32
Warmup steps	1000
Weight decay	0.01
Gradient accumulation steps	64

Except for the training epochs, ProtBERT-BFD and MembraneBERT use the same fine-tuning hyperparameter settings as ProtTrans [10].

3.4 Logistic Regression

Logistic Regression is a widely used classification technique in medical/biological research [12]. The Logistic Regression algorithm used was the scikit-learn Python module (<https://scikit-learn.org>) and the study used the default hyperparameters.

3.5 Evaluation

A 10-fold cross-validation (CV) technique was used in this analysis to evaluate the model’s performance by partitioning the dataset into ten sections. For the purpose of fine-tuning the BERT, 10% of the training set was used as the validation set, while the remaining 90% was used for training. The independent test set is utilised for the sole purpose of evaluating the method.

3.6 Evaluation Metrics

Four key evaluation criteria are considered in this project: Sensitivity (Sen), Specificity (Sp), Accuracy (Acc), and MCC.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

MCC is an acronym for Matthew’s Correlation Coefficient. For imbalanced data, MCC is a more stable assessment metric [7].

4 Results and Discussion

4.1 Fine-Tuning ProtBERT-BFD and MembraneBERT

We compared both representations of ProtBERT-BFD and MembraneBERT, without (frozen) and with (fine-tuned) fine-tuning using the DS-T dataset. Figure 1 visualises the effect of fine-tuning ProtBERT-BFD and MembraneBERT for each epoch.

As demonstrated, the ProtBERT-BFD model improved representations in each epoch, increasing from zero MCC and 56% accuracy to 0.77 MCC and 87% accuracy on the validation set. The ProtBERT-BFD model outperforms the MembraneBERT model, indicating that a BERT model trained on a more extensive set of protein sequences has superior representation and performance in the downstream task fine-tuning. Additionally, the ProtBERT-BFD performs better in both frozen and

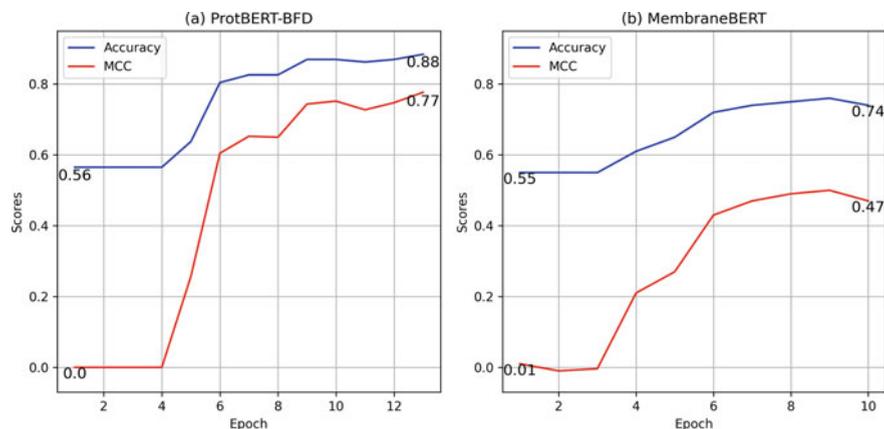


Fig. 1 The effect of fine-tuning (This figure depicts the results of fine-tuning the ProtBERT-BFD (left) and MembraneBERT (right) with accuracy and MCC metrics at each epoch on the validation set. The y-axis and x-axis display the scores and epochs, respectively)

Table 3 Logistic Regression performance with ProtBERT-BFD and MembraneBERT

Model	Sen (%)		Spc (%)		Acc (%)		MCC	
	Ind.	CV	Ind.	CV	Ind.	CV	Ind.	CV
ProtBERT-BFD frozen	76.67	80.00	90.83	82.69	86.11	81.52	0.6840	0.6262
ProtBERT-BFD fine-tuned	95.83	96.79	90.00	97.17	93.89	96.96	0.8620	0.9387
MembraneBERT frozen	88.33	80.51	68.33	77.50	81.67	79.20	0.5799	0.5797
MembraneBERT fine-tuned	86.67	98.08	85.00	97.00	86.11	97.61	0.6989	0.9512

This table summarizes the 10-fold CV and independent test set performance of frozen/fine-tuned representations from the ProtBERT-BFD and MembraneBERT models in terms of sensitivity, specificity, accuracy, and MCC. The maximum value for each column is displayed in boldface.

fine-tuned representations than MembraneBERT, with the exception of the frozen representation of sensitivity. Despite the high cost of fine-tuning the 420 million-parameter ProtBERT-BFD model, our results (Table 3) demonstrate that fine-tuning ProtBERT-BFD for transport protein prediction results in the best representation.

4.2 Logistic Regression with Fine-Tuned ProtBERT-BFD

We selected Logistic Regression as a preliminary good binary classifier because it is simple to implement and interpret, has been tested in the ProtTrans project, and produces competitive results [10]. Table 3 demonstrates that Logistic Regression with

both fine-tuned ProtBERT-BFD and MembraneBERT representations performs well, with fine-tuned ProtBERT-BFD outperforming MembraneBERT on all independent test set results, while MembraneBERT outperforms sensitivity, accuracy, and MCC on CV results.

4.3 Comparison of TooT-BERT-T with State-of-the-Art Models

Table 4 and Fig. 2 are used to compare TooT-BERT-T to other published methods that use only the protein sequence on the same dataset. As demonstrated, TooT-BERT-T outperforms other published works in all evaluation metrics except sensitivity, where Nguyen et al. [17] achieves 100% sensitivity.

Table 4 Comparative performance of TooT-BERT-T with state-of-the-art

Method	Sen (%)		Spc (%)		Acc (%)		MCC	
	Ind.	CV	Ind.	CV	Ind.	CV	Ind.	CV
SCMMTP [15]	80.00	83.76	68.33	77.68	76.11	81.12	0.47	0.62
TrSSP [16]	76.67	76.67	81.67	78.46	80.00	78.99	0.57	0.58
Nguyen et al. [17]	100.00	83.14	77.50	84.48	85.00	83.94	0.73	0.68
TooT-T [2]	94.17	90.15	88.33	89.97	92.22	90.07	0.82	0.80
TooT-BERT-T	95.83	96.79	90.00	97.17	93.89	96.96	0.86	0.94

This table compares the outcomes of various techniques using sensitivity, specificity, accuracy, and MCC metrics on the CV and independent test set. Results taken from [2]. The maximum value for each column is displayed in boldface.

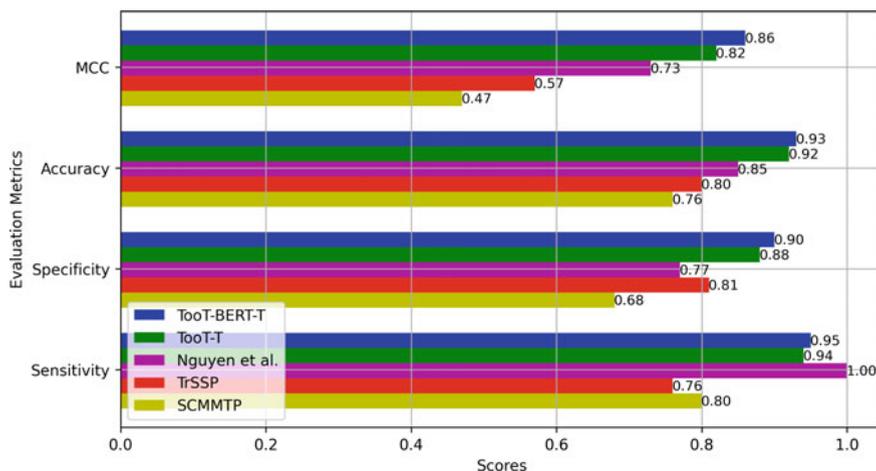


Fig. 2 Comparison of methodologies

Fig. 3 TooT-BERT-T confusion matrix (This figure summarises the performance of TooT-BERT-T, where *T* represents transport protein and *non-T* represents non-transport protein)

		Predicted values	
		T	non-T
Actual values	T	115	5
	non-T	6	54
		T	non-T

TooT-BERT-T has a greater specificity (rate of true negatives) than the approach of Nguyen et al. [17], indicating that it makes fewer false positive predictions (Fig. 3). This is essential for achieving a high true negative rate of 90% when describing non-transport proteins.

The proposed method, TooT-BERT-T, which employs fine-tuned ProtBERT-BFD representation and a Logistic Regression classifier using the dataset explained in Sect. 3.1, outperforms previous methods with an accuracy of 93.89% and an MCC of 0.86 on the independent test set.

The ProtBERT-BFD representation is effective because it understands the context of each amino acid in different protein sequences, whereas other methods rely on static protein-encoding techniques.

Figure 3 shows a confusion matrix of TooT-BERT-T for separating transport proteins from non-transport proteins. As depicted in the figure, despite the fact that the number of errors is quite low, the model makes more mistakes when identifying non-transporters as transporters (False positive = 6) than when predicting transporters as non-transporters (False negative = 5). This suggests that the proposed strategy is somewhat skewed towards predicting the positive class (transport proteins). This issue may occur when the dataset is imbalanced, with more positive class samples than negative class samples.

5 Conclusion

TooT-BERT-T distinguishes transport proteins from non-transport proteins using the fine-tuned ProtBERT-BFD representation. The representations of two BERT models, ProtBERT-BFD and MembraneBERT, were compared using frozen and fine-tuned representations. The ProtBERT-BFD fine-tuned representation outperforms the MembraneBERT representation on the independent test set. The proposed method, TooT-BERT-T, which utilizes fine-tuned ProtBERT-BFD and Logistic Regression, achieves an accuracy of 93.89% and an MCC of 0.86 on the independent test set and outperforms other methods. Given that this study was a preliminary examination of the BERT representation's performance in transport protein analysis, other classifiers such as SVM and CNN can be evaluated in the future.

References

1. Alballa M, Butler G (2020) Integrative approach for detecting membrane proteins. *BMC Bioinform* 21(19):575
2. Alballa M, Butler G (2020) TooT-T: discrimination of transport proteins from non-transport proteins. *BMC Bioinform* 21(3):25
3. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM (2019) Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* 16(12):1315–1322
4. Aplop F, Butler G (2015) On predicting transport proteins and their substrates for the reconstruction of metabolic networks. In: 2015 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB), pp 1–9
5. Aplop F, Butler G (2017) TransATH: transporter prediction via annotation transfer by homology. *ARPN J Eng Appl Sci* 12(2):8
6. Beppler T, Berger B (2019) Learning protein sequence embeddings using information from structure. [arXiv:1902.08661](https://arxiv.org/abs/1902.08661) [cs, q-bio, stat]
7. Chicco D, Jurman G (2020) The advantages of the Matthews Correlation Coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom* 21(1):6
8. Detlefsen NS, Hauberg S, Boomsma W (2022) Learning meaningful representations of protein sequences. *Nat Commun* 13(1):1914
9. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) [cs]
10. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, Bhowmik D, Rost B (2021) ProtTrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *IEEE Trans Pattern Anal Mach Intell* 1
11. Ferruz N, Höcker B (2022) Towards controllable protein design with conditional transformers. [arXiv:2201.07338](https://arxiv.org/abs/2201.07338) [q-bio]
12. Hess AS, Hess JR (2019) Logistic regression. *Transfusion* 59(7):2197–2198
13. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873):583–589
14. Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, Apweiler R (2004) UniProt archive. *Bioinformatics* 20(17):3236–3237
15. Liou YF, Vasylenko T, Yeh CL, Lin WC, Chiu SH, Charoenkwan P, Shu LS, Ho SY, Huang HL (2015) SCMMTP: identifying and characterizing membrane transport proteins using propensity scores of dipeptides. *BMC Genom* 16(12):S6
16. Mishra NK, Chang J, Zhao PX (2014) Prediction of membrane transport proteins and their substrate specificities using primary sequence information. *PLoS ONE* 9(6):e100278
17. Nguyen TTD, Le NQK, Ho QT, Phan DV, Ou YY (2019) Using word embedding technique to efficiently represent protein sequences for identifying substrate specificities of transporters. *Anal Biochem* 577:73–81
18. Ofer D, Brandes N, Linal M (2021) The language of proteins: NLP, machine learning & protein sequences. *Comput Struct Biotechnol J* 19:1750–1758
19. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen P, Canny J, Abbeel P, Song Y (2019) Evaluating protein transfer learning with TAPE. In: Wallach H, Larochelle H, Beygelzimer A, Alché-Buc Fd, Fox E, Garnett R (eds) *Advances in neural information processing systems*, vol 32. Curran Associates, Inc
20. Saier Jr MH, Tran CV, Barabote RD (2006) TCDB: the transporter classification database for membrane transport protein analyses and information. *Nucleic Acids Res* 34(suppl_1):D181–D186

21. Saier Jr MH (2002) Families of transporters and their classification. In: Transmembrane transporters. Wiley, pp 1–17
22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. arXiv
23. Vig J, Madani A, Varshney LR, Xiong C, Socher R, Rajani NF (2021) BERTology meets biology: interpreting attention in protein language models. [arXiv:2006.15222](https://arxiv.org/abs/2006.15222) [cs, q-bio]
24. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Scao TL, Gugger S, Drame M, Lhoest Q, Rush AM (2020) HuggingFace’s transformers: state-of-the-art natural language processing. arXiv

Machine Learning and Deep Learning Techniques for Epileptic Seizures Prediction: A Brief Review



Marco Hernández , Ángel Canal-Alonso , Fernando de la Prieta ,
Sara Rodríguez , Javier Prieto , and Juan Manuel Corchado 

Abstract The third most common neurological disorder, only behind stroke and migraines, is Epilepsy. The main criteria for its diagnosis are the occurrence of unprovoked seizures and the possibility of new seizures appearing. Usually, the professional in charge of detecting these seizures is a neurologist who interprets the patients' electroencephalography. However, more accurate, precise, and sensitive methods are needed. Machine learning has increased as a viable alternative, reducing costs and ensuring rapid diagnostic time. This work reviews the state of the art in machine learning applied to epileptic seizure detection and prediction as a prospective study before developing a novel seizure prediction algorithm.

Keywords Seizure prediction · Machine learning · Epilepsy · Electroencephalogram

This work was supported by the HERMES project, funded by the European Union under the Horizon 2020 FET-proactive program, Grant Agreement n. 824164., as well as the “XAI - XAI - Sistemas Inteligentes Auto Explicativos creados con Módulos de Mezcla de Expertos” project, ID SA082P20, financed by Junta Castilla y León, Consejería de Educación, and FEDER funds.

M. Hernández (✉) · Á. Canal-Alonso · F. de la Prieta · S. Rodríguez · J. Prieto · J. M. Corchado
BISITE Research Group, University of Salamanca, 37007 Salamanca, Spain
e-mail: marcohpez@usal.es

Á. Canal-Alonso
e-mail: acanal@usal.es

F. de la Prieta
e-mail: fer@usal.es

S. Rodríguez
e-mail: srg@usal.es

J. Prieto
e-mail: javierp@usal.es

J. M. Corchado
e-mail: corchado@usal.es

1 Introduction

Epilepsy is a neurological disorder affecting more than 39 million people in the United States [1], being the third most common only before stroke and migraines [31]. Due to its chronic nature, it is one of the most invalidating conditions, making it a suitable target for new therapies and biomedical research [11, 14, 34].

The International League Against Epilepsy defines an epileptic seizure as a “transient occurrence of signs and/or symptoms due to abnormal excessive or synchronous neuronal activity in the brain” [59].

The most widely used diagnostic method for epilepsy is electroencephalography (EEG). EEG records the electrical activity of the brain by the voltage changes provoked by the ion currents of the brain neurons [44].

These voltage changes are recorded with electrodes placed either along the scalp (sEEG) or over the cortex, inside the cranium (iEEG). Those electrodes do not have enough spatial resolution to detect each neuron’s action potential, but the simultaneous activity of millions of neurons creates a voltage wave that stimulates the electrodes.

The events observed in the EEG can be divided into three categories [16]: Ictal events, which are the events occurring during the seizure, preictal events, the ones that preceded the seizure, and interictal events, every other event that is not part of the ictal or preictal phase. The duration of the preictal phase varies between studies, and can range from a few minutes to hours.

In animal and *in vitro* models, microelectrodes are used instead of EEG to perform electrophysiology recordings, as the latter carries a considerable amount of drawbacks in such cases (Fig. 1).

2 Signal Processing and Feature Extraction

EEG signals usually come with several artifacts that may obscure the signal’s epilepsy-related information. These artifacts depend on the type of model of study or the specifics of the EEG recording.

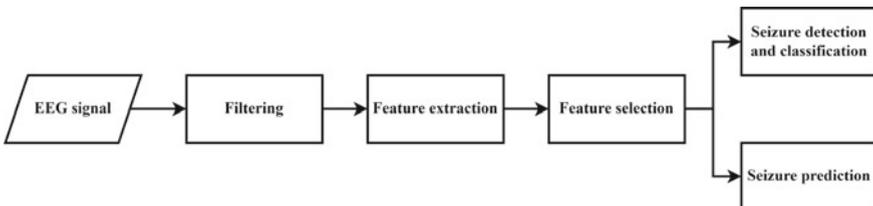


Fig. 1 Usual pipeline in epileptic seizure classification or prediction

The most typical processing for most EEG signals includes the removal of background noise, done by filtering the 50–60 Hz powerline. A vast number of features can then be extracted from the processed signal.

2.1 Feature Extraction

Statistical Features. Distribution and amplitude changes in EEG signals can be tracked using statistical parameters like kurtosis, mean, skewness, and variance [12, 38, 53]. Phase correlation [39, 46, 49, 57] can be used to analyze the patterns in ictal and preictal events. Common Spatial Pattern [56] extracts features by decomposing EEG signals.

Nonlinear Features. Correlation Dimension allows to measure the complexity of each event [2, 27] and the Largest Lyapunov Exponent calculates the chaos in EEG signals [60, 61]. Applying Fractality Dimension [17, 18, 41] enables the comparison of the rhythms between the EEG events and exposes the self-similarity in the data. The Repeatability of the events is also measurable using Lempel-Zic Complexity [2, 10, 71] and Approximate Entropy [58, 70]. Entropy can also be used as a measure of randomness with Spectral Entropy and level of disorganization using Wavelet Entropy.

Activity (variance of the signal of a time function), mobility (proportion of the standard deviation), and complexity (change in frequency) parameters can be used as descriptors in the Hjorth parameters analysis [19, 32]. To unveil the uniformity of the different frequency bands in EEG data Wavelet Energy can be applied [6, 24, 26].

Frequency Domain Features. Fourier transforms like Short Time Fourier Transform [63] and Fractional Fourier Transform [48] are used to obtain phase and magnitude components, while Spectral Power Analysis [7, 21] allows studying the different frequency bands in EEG.

Time-Frequency Domain Features. One of the most widely used time-frequency domain feature in EEG analysis is the wavelet transform, either Discrete Wavelet Transform [68], or Continuous Wavelet Transform [40]. Each wavelet transform offers a different decomposition; the CWT generates a scalogram from the dilation and translation, while the DWT filters the signal and breaks down the signal in different levels.

Some time-frequency domain features can be combined with nonlinear features to monitor hidden data properties. Higher-Order Spectra [4, 45] and Variational Model Decomposition [20, 35] are the most widely used features in these combinations.

2.2 Feature Selection

For most Machine Learning techniques, selecting an optimum number of features is essential for the algorithms to reach their full potential. Having features that carry similar information about the target variable or no information about this variable whatsoever can make the model too complex and impoverish its performance.

A roster of techniques is used to select the most suitable features for each study. Statistical approaches like Principal Components Analysis [23, 50, 66] or Partial Least Squares [30] are the easiest way of arriving at a conclusion. More complex techniques such as Minimum Redundancy Maximum Relevance [8, 51] or Gaussian Mixture Models [22, 52, 69] are used in more sensitive situations like human seizure prediction.

However, Deep Learning algorithms can use part of their architecture to automatically extract the most relevant information out of the initial input variables [65, 67]. In these cases, both feature extraction and selection can be skipped and still achieve competitive results.

3 Seizure Detection and Classification

Detection of seizures by EEG has traditionally been done manually by clinical professionals, evaluating the frequency, wavelength, voltage, amplitude, and waveforms. These features are suitable for being analyzed using automated learning algorithms [28]. Since the first computer analysis of EEG records in 2002 using wavelet transform [5], automated detection of ictal events has become an essential matter in epilepsy research [9, 25].

Many studies have been carried out in the last years to improve the performance of automated seizure detection. A wide range of Machine Learning and Deep Learning classifiers have been employed [15, 37] but, while detecting seizures may remain valuable for research purposes, patients and clinical professionals need tools that allow them to avoid the seizures instead of doing a post hoc analysis; here is where seizure prediction comes necessary.

4 Seizure Prediction

Having enough anticipation before a seizure is a crucial milestone for a clinical approach. The usual pipeline to seizure prediction is similar to seizure detection. However, instead of classifying interictal and ictal events, it focuses on separating interictal from preictal, leaving the actual seizures out of the analysis.

Establishing a correct stimulation protocol implies having enough anticipation before the seizure. Nevertheless, early prediction usually reduces sensitivity and specificity, ballasting the overall performance [29, 42].

4.1 Animal Models

The first experiments achieved times near 2.24 min and demonstrated the effectiveness of wavelet functions as predictors [47]. Best time results in animal models were obtained by [64] using canine EEG data and multiple machine learning algorithms, being able to detect the seizures 1 min ahead. That work established a proof of concept, so no diagnostic performance analyses were carried out. Nevertheless, some works have developed systems with performances over 90% of sensitivity but with a loose prediction time. Rajdev's team [55] developed a seizure prediction system for rat EEG recordings based on an adaptive wiener filter; his approach hits a 92% of sensitivity, being also the most sensitive of the works done on animal EEG recordings.

4.2 Human Subjects

The work of Iasemidis [33] established a ceiling of capacity in the time of prediction with 91 min and a precision of 91.3% (and sensitivity of 81.82%). Other authors aimed to maintain a prediction time horizon and keep the algorithm between those parameters. In such works [2, 70] 30 and 50-min horizons were fixed, and sensitivity between 79.9 and 90.2% were achieved.

Tsiouris [62] reached a prediction with 15 to 120 min ahead and a 99% of sensitivity, making use of Long Short-term Memory networks (LSTM), the first application of deep learning in the field.

Other deep learning approaches have reached similar results while automating feature extraction. Wei [65] uses an image of the EEG as input to an architecture based on Convolutional Neural Networks (CNN) for feature extraction and LSTM for sequence learning. This network achieves an average accuracy of 93.4% at an average warning time of 21 min. Transformers have been used in a similar manner [67] reaching prediction sensitivity and a False Positives Rate of 96.01% and 0.047/h, respectively with an average warning time between 3 and 30 min.

5 Conclusion

Although many advances have been made since the first works in automated seizure prediction, some gaps remain to be cleared.

To deploy accurate diagnostic and treatment tools, a correct balance between the time of prediction and the sensitivity and specificity must be reached. This balance can only be achieved with deep learning techniques such as CNN, LSTM, and Transformers. In this regard, some recent developments have been made, reaching promising results [3, 13, 36, 43, 54, 65, 67].

A combination of multiple techniques and features offers the best performance and results, but the computational requirements scale with each model implemented. Solving this issue is also a significant challenge in automated seizure prediction.

The solution to the automated seizure prediction problem will surely enhance the life quality of epilepsy patients and ameliorate the impact on the health services.

References

1. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the global burden of disease study. Technical report, GBD 2015 Disease and Injury Incidence and Prevalence Collaborators (2015)
2. Aarabi A, He B (2012) A rule-based seizure prediction method for focal neocortical epilepsy. *Clin Neurophysiol* 123:1111–1122
3. Acharya UR, Oh S, Hagiwara Y, Adeli H (2017) Deep convolutional neural network for the automated detection of seizure using EEG signals. *Comput Biol Med* 100:270–278
4. Acharya U, Yanti R, Zheng J, Mookiah M, Tan J, Martis R et al (2013) Automated diagnosis of epilepsy using CWT, HOS and texture parameters. *Int J Neural Syst* 23
5. Adeli H, Zhou Z, Dadmehr N (2003) Analysis of EEG records in an epileptic patient using wavelet transform. *J Neurosci Methods* 123:69–87
6. Alves N, Rodrigues R, Rocha M (2022) BioTMPy: a deep learning-based tool to classify biomedical literature. In: *Lecture notes in networks and systems*. LNNS, vol 325, pp 115–125
7. Bandarabadi M, Rasekhi J, Teixeira C, Karami M (2015) On the proper selection of preictal period for seizure prediction. *Epilepsy Behav* 45:158–166
8. Bandarabadi M, Teixeira C, Rasekhi J, Dourado A (2015) Epileptic seizure prediction using relative spectral power features. *Clin Neurophysiol* 126:237–248
9. Calvaresi D, Albanese G, Calbimonte J, Schumacher M (2020) Seamless: simulation and analysis for multi-agent system in time-constrained environments. In: *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. LNAI, vol 12092, pp 392–397
10. Casado-Vara R, González-Briones A, Prieto J, Corchado J (2019) Smart contract for monitoring and control of logistics activities: pharmaceutical utilities case study. *Adv Intell Syst Comput* 771:509–517
11. Casado-Vara R, Novais P, Gil A, Prieto J, Corchado J (2019) Distributed continuous-time fault estimation control for multiple devices in IoT networks. *IEEE Access* 7:11972–11984
12. Casado-Vara R, Prieto-Castrillo F, Corchado J (2018) A game theory approach for cooperative control to improve data quality and false data detection in WSN. *Int J Robust Nonlinear Control* 28(16):5087–5102
13. Chen H, Shen J, Wang L, Jin Y (2021) Towards a more effective bidirectional LSTM-based learning model for human-bacterium protein-protein interactions. In: *Advances in intelligent systems and computing*. AISC, vol 1240, pp 91–101
14. Costa Â, Novais P, Corchado J, Neves J (2012) Increased performance and better patient attendance in an hospital with the use of smart agendas. *Logic J IGPL* 20(4):689–698
15. Cristani M, Tomazzoli C, Olivieri F, Pasetto L (2020) An ontology of changes in normative systems from an agentive viewpoint. In: *Communications in computer and information science*. CCIS, vol 1233, pp 131–142

16. Cámpora NE, Mininni CJ, Kochen S, Lew SE (2019) Seizure localization using pre ictal phase-amplitude coupling in intracranial electroencephalography. *Sci Rep* 9:20022
17. D'Alessandro M, Esteller R, Vachtsevanos G, Hinson A, Echaz J, Litt B (2003) Epileptic seizure prediction using hybrid feature selection overmultiple intracranial EEG electrode contacts: a report of four patients. *IEEE Trans Inf Theory* 50:603–615
18. De Meo P, Falcone R, Sapienza A (2020) Fast and efficient partner selection in large agents' communities: when categories overcome direct experience. In: *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. LNAI, vol 12092, pp 106–117
19. Direito B, Teixeira CA, Sales F, Castelo-Branco M, Dourado A (2017) A realistic seizure prediction study based on multiclass SVM. *Int J Neural Syst* 27:1750006–1750021
20. Dragomiretskiy K, Zosso D (2014) Variational mode decomposition. *IEEE Trans Signal Process* 62:531–544
21. Dressler O, Schneider G, Stockmanns G, Kochs E (2004) Awareness and the EEG power spectrum: analysis of frequencies. *Br J Anaesth* 93
22. D'Auria M, Scott E, Lather R, Hilty J, Luke S (2020) Assisted parameter and behavior calibration in agent-based models with distributed optimization. In: *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. LNAI, vol 12092, pp 93–105
23. Fatima N (2020) Enhancing performance of a deep neural network: a comparative analysis of optimization algorithms. *ADCAIJ Adv Distrib Comput Artif Intell J* 9(2):79–90
24. Gadhouri K, Lina J, Gotman J (2013) Seizure prediction in patients with mesial temporal lobe epilepsy using EEG measures of state similarity. *Clin Neurophysiol* 124:1745–1754
25. García-Retuerta D, Canal-Alonso A, Casado-Vara R, Rey A, Panuccio G, Corchado J (2021) Bidirectional-pass algorithm for interictal event detection. In: *Advances in intelligent systems and computing*. AISC, vol 1240, pp 197–204
26. Garg G, Singh V, Gupta J, Mittal A (2011) Relative wavelet energy as a new feature extractor for sleep classification using EEG signals. *Int J Biomed Signal Process* 2:75–80
27. Grassberger P, Procaccia I (1983) Characterization of strange attractors. *Phys Rev Lett* 50:346–349
28. Gupta S, Meena J, Gupta O (2020) Neural network based epileptic EEG detection and classification. *ADCAIJ Adv Distrib Comput Artif Intell J* 9(2):23–32
29. Gupta S, Ranga V, Agrawal P (2022) EpilNet: a novel approach to IoT based epileptic seizure prediction and diagnosis system using artificial intelligence. *ADCAIJ Adv Distrib Comput Artif Intell J* 10(4):435–452
30. Haenlein A (2004) A beginner's guide to partial least squares analysis. *Underst Stat* 283–297
31. Hirtz D, Thurman DJ, Gwinn-Hardy K, Mohamed M, Chaudhuri AR, Zalutsky R (2007) How common are the "common" neurologic disorders? *Neurology* 68:326–337
32. Hjorth B, Elema-Schonander A (1970) EEG analysis based on time domain properties. *Electroencephalogr Clin Neurophysiol* 29:306–310
33. Iasemidis L, Shiau D, Pardalos P, Chaovalitwongse W, Narayanan K, Prasad A (2005) Long-termprospective on-line real-time seizure prediction. *Clin Neurophysiol* 116:532–544
34. Khan A, Zubair S, Khan S (2021) Comprehensive performance analysis of neurodegenerative disease incidence in the females of 60–96 year age group. *ADCAIJ Adv Distrib Comput Artif Intell J* 10(2):183–196
35. Kumar M, Rao Y (2018) Epileptic seizures classification in EEG signal based on semantic features and variational mode decomposition. *Clust Comput* 1–11
36. Lane N, Kahanda I (2021) DeepACPPred: a novel hybrid CNN-RNN architecture for predicting anti-cancer peptides. In: *Advances in intelligent systems and computing*. AISC, vol 1240, pp 60–69
37. Lee K, Jeong H, Kim S, Yang D, Kang HC, Choi E (2022) Real-time seizure detection using EEG: a comprehensive comparison of recent approaches under a realistic setting. [arXiv:2201.08780](https://arxiv.org/abs/2201.08780) [cs]

38. Li T, Fan H, García J, Corchado J (2018) Second-order statistics analysis and comparison between arithmetic and geometric average fusion: application to multi-sensor target tracking. *Inf Fusion* 51:233–243
39. Li T, Su J, Liu W, Corchado J (2017) Approximate gaussian conjugacy: parametric recursive filtering under nonlinearity, multimodality, uncertainty, and constraint, and beyond. *Front Inf Technol Electron Eng* 18(12):1913–1939
40. Mallat S, Hwang W (1992) Singularity detection and processing with wavelets. *IEEE Trans Inf Theory* 38:617–643
41. Mandelbrot B (1983) *Geometry of nature*. Freeman
42. Mena Mamani N (2020) Machine learning techniques and polygenic risk score application to prediction genetic diseases. *ADCAIJ Adv Distrib Comput Artif Intell J* 9(1):5–14
43. Muhamada AW, Mohammed AA (2022) Review on recent computer vision methods for human action recognition. *ADCAIJ Adv Distrib Comput Artif Intell J* 10(4):361–379
44. Niedermeyer E, da Silva F (2004) *Electroencephalography: basic principles, clinical applications, and related fields*. Williams & Wilkins
45. Nikias C, Petropulu A (1993) *Higher order spectra analysis: a nonlinear signal processing framework*. PTR Prentice Hall
46. Nugroho S, Weinmann A, Schindelbauer C, Christ A (2020) Averaging emulated time-series data using approximate histograms in peer to peer networks. In: *Communications in computer and information science*. CCIS, vol 1233, pp 339–346
47. Ouyang G, Li X, Guan X (2007) Application of wavelet-based similarity analysis to epileptic seizures prediction. *Comput Biol Med* 37:430–437
48. Ozaktas H, Zalevsky Z, Kutay M (2001) *The fractional Fourier transform*. Wiley
49. Parvez M, Paul M (2016) Epileptic seizure prediction by exploiting spatiotemporal relationship of EEG signals using phase correlation. *IEEE Trans Neural Syst Rehabil Eng* 24:158–168
50. Pearson K (1901) On lines and planes of closest fit to systems of points in space. *Philos Mag* 2:559–572
51. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 27:1226–1238
52. Press M (ed) (1999) *The infinite Gaussian mixture model*
53. Press W, Flannery B, Teukolsky S, Vetterling W (1992) *Numerical recipes in C: the art of scientific computing*. Cambridge University Press
54. Pérez-López R, Blanco G, Fdez-Riverola F, Lourenço A (2021) The activity of bioinformatics developers and users in stack overflow. In: *Advances in intelligent systems and computing*. AISC, vol 1240, pp 23–31
55. Rajev P, Ward M, Rickus J, Worth R, Irazoqui P (2010) Real-time seizure prediction from local field potentials using an adaptive wiener algorithm. *Comput Biol Med* 40:97–108
56. Ramoser H, Miller-Gerking J, Pfurtscheller G (2000) Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Trans Rehabil Eng* 8:441–446
57. Reddy B, Chatterji B (1996) An FFT-based technique for translation, rotation, and scale invariant image registration. *IEEE Trans Image Process* 5:1266–1271
58. Richman J, Moorman J (2000) Physiological time-series analysis using approximate entropy and sample entropy. *Am J Physiol Heart Circ Physiol* 278:2039–2049
59. Robert S, Fisher AC, Arzimanoglou A, Bogacz A, Cross JH, Elger Jr CE, Forsgren L, French JA, Glynn M, Hesdorffer DC, Lee B, Mathern GW, Moshé SL, Perucca E, Scheffer IE, Tomson T, Watanabe M, Wiebe S (2014) ILAE official report: a practical clinical definition of epilepsy. *Epilepsia* 55:475–482
60. Rosenstein M, Collins J, de Luca C (1993) A practical method for calculating largest Lyapunov exponents from small data sets. *Phys D* 65:117–134
61. Shafique A, Sayeed M, Tsakalis K (2018) *Nonlinear dynamical systems with chaos and big data: a case study of epileptic seizure prediction and control*. Guide to big data applications. Springer (2018)

62. Tiouris K, Pezoulas V, Zervakis M, Konitsiotis S, Koutsouris D, Fotiadis D (2018) A long short-term memory deep learning network for the prediction of epileptic seizures using EEG signals. *Comput Biol Med* 99:24–37
63. Truong N, Nguyen A, Kuhlmann L, Bonyadi M, Yang J, Ippolito S et al (2018) Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. *Neural Netw* 105:104–111
64. Varatharajah Y, Iyer R, Berry B, Worrell G, Brinkmann B (2017) Seizure forecasting and the preictal state in canine epilepsy. *Neural Syst* 27
65. Wei X, Zhou L, Zhang Z, Chen Z, Zhou Y (2019) Early prediction of epileptic seizures using a long-term recurrent convolutional network. *J Neurosci Methods* 327:108395
66. Williamson J, Bliss D, Browne D, Narayanan J (2012) Seizure prediction using EEG spatiotemporal correlation structure. *Epilepsy Behav* 25:230–238
67. Yan J, Li J, Xu H, Yu Y, Xu T (2022) Seizure prediction based on transformer using scalp electroencephalogram. *Appl Sci* 12:4158
68. Yves M (1992) *Wavelets and operators*. Cambridge University Press
69. Zandi A, Tafreshi R, Javidan M, Dumont G (2013) Predicting epileptic seizures in scalp EEG based on a variational Bayesian Gaussian mixture model of zero-crossing intervals. *IEEE Trans Biomed Eng* 60:1401–1413
70. Zhang Y, Zhou W, Yuan Q, Wu Q (2014) A low computation cost method for seizure prediction. *Epilepsy Res* 108:1357–1366
71. Ziv J, Lempel A (1977) A universal algorithm for sequential data compression. *IEEE Trans Inf Theory* 23:337–343

The Covid-19 Decision Support System (C19DSS) – A Mobile App



Pierpaolo Vittorini , Nicolò Casano , Gaia Sinatti ,
Silvano Junior Santini , and Clara Balsano 

Abstract The COVID-19 pandemic remains a concrete challenge, especially in communities and rural areas where health resources are scarce. We recently developed several classifiers, useful to predict safe discharge, disease severity, and mortality risk from COVID-19, fed by routine analyses collected in the Emergency Department. In this paper, we discuss a system, made up of an app and a server, that enables doctors to use these models during the management of COVID-19 patients. The app has been developed involving the doctors since the early phases of the app design, then revised in the light of two usability cycles. We report its main features and its ease of use. So far, it has been used during the fourth wave, producing accurate results with patients that did not complete the vaccination protocol (i.e., up to the second dose).

Keywords COVID-19 · App · Machine learning · User-centered design

1 Introduction

Health informatics can be defined as the application of computer science, engineering and telecommunication to healthcare [2]. It regards the use of methods, applications and devices in all aspects concerned with both individuals and public health [8, 11, 23, 25].

The pandemic caused by severe respiratory acute syndrome coronavirus 2 (SARS-CoV-2) was declared a global emergency by the World Health Organization (WHO) [24] on the 11th of March 2020. Although recent progress in the possible treatments has changed the face of the pandemic, some concerns still remain about threats related to SARS-CoV-2 [21]. Countries have not adopted a common global response to COVID-19 and vaccination inequities are manifest [17]. In this scenario, the pandemic risk remains a concrete challenge, especially in communities and rural

P. Vittorini (✉) · N. Casano · G. Sinatti · S. J. Santini · C. Balsano
University of L'Aquila, L'Aquila 67100, Italy
e-mail: pierpaolo.vittorini@univaq.it
URL: <https://vittorini.univaq.it/>

areas where health resources are scarce. To face these risks, we recently developed several classifiers, useful to predict safe discharge, disease severity, and mortality risk from COVID-19, fed by routine analyses collected in the Emergency Department (ED) [7].

In this paper, we focus on a system, called COVID-19 Decision Support System (C19DSS), whose aim is to enable doctors from EDs to take advantage of the aforementioned models during the management of COVID-19 patients. The C19DSS system has been developed following the User-Centered Design (UCD) methodology, i.e., involving the doctors since the early phases of the system design, and revising the development in the light of usability cycles [16]. So far, the system has been used during the fourth wave. It is currently producing accurate results with patients that did not complete the vaccination protocol (i.e., up to the second dose). Therefore, we consider the system suitable in these cases, as well as in all countries with low vaccination rates.

2 Background

To allow the paper to be self-contained, we briefly report the results concerning the models we developed, which we submitted in [7].

From a dataset containing the routine analyses of 779 patients collected in the ED, we devised several models for both the complete cases and through missing data imputation [6]. The following different models were tried from the available dataset: decision tree (DT) – as baseline [5], random forest (RF) [4] and gradient boosting machines (GBM) [9]. The models were developed to predict safe discharge (discharge/admit), disease severity (mild/severe), and mortality risk (no risk/risk). For all models and outcomes, we split the dataset into train and test (with 75% of data going for training, 25% for testing), used 10-fold cross-validation, tuned each classifier according to its specific hyper-parameters, calculated the confusion matrix and the ROC curve [10]. The results are summarized in Table 1. The table lists all details of each model, for each outcome, for the subset of the complete cases and for the complete dataset (with missing data imputation). The best AUC is reported in bold, together with the corresponding model.

Besides the limited size of the dataset and the constraint of using only routine clinical and laboratory data to devise the models, the performances of our models are in line with the best prediction models available in the scientific literature, that make use of similar data than our [3, 12, 26]. In particular: (i) concerning hospital admission, Jimenez-Solem et al. [12] developed a RF model that reached an AUC equal to 0.82, vs 0.89 and 0.94 of our models; (ii) for severity prediction, Yao et al. [26] devised a Support Vector Machines (SVM) model that reached an accuracy equal to 0.82, vs the 0.76 and 0.89 of accuracy of our models; (iii) for mortality prediction, Booth et al. [3] developed a SVM model that reached an AUC of 0.93, vs 0.84 and 0.87 of our models.

Table 1 Main statistics for all outcomes and classifiers. *Acc = Accuracy, Sens = Sensitivity, Spec = Specificity, AUC = Area Under the Curve*

	Complete cases				Missing data imp.				
	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC	
<i>Safe discharge</i>									
DT	0.870	0.941	0.546	0.937	0.824	0.842	0.770	0.858	DT
RF	0.886	0.960	0.546	0.938	0.829	0.869	0.780	0.894	RF
GBM	0.886	0.951	0.591	0.943	0.824	0.876	0.666	0.882	GBM
<i>Disease severity</i>									
DT	0.805	0.560	0.867	0.792	0.742	0.732	0.752	0.766	DT
RF	0.886	0.680	0.939	0.886	0.757	0.783	0.732	0.832	RF
GBM	0.846	0.600	0.908	0.893	0.762	0.804	0.721	0.827	GBM
<i>Mortality</i>									
DT	0.829	0.970	0.250	0.758	0.840	0.944	0.290	0.689	DT
RF	0.878	0.960	0.542	0.866	0.876	0.969	0.387	0.842	RF
GBM	0.854	0.939	0.500	0.857	0.860	0.944	0.419	0.844	GBM

With respect to similar tools available in the scientific literature, Liu et al. [15] propose a system to assist doctors in collecting data, assessing risk, triaging, managing, and following up on patients during the COVID-19 outbreak. The system uses logistic regression to predict risk, obtaining an AUC of 0.71. Furthermore, McRae et al. [18] developed an app that leverages models that use non-laboratory data to help determine whether hospitalization is necessary (AUC = 0.79) and that predicts the probability of mortality using bio-marker measurements (AUC = 0.95).

Our work continues in the same direction: it adopted state-of-the-art models based on routine data collected by the involved EDs, and developed a system supporting doctors in defining the need for hospitalization, disease severity and mortality risk, for patients accessing the ED.

3 C19DSS

In this section, we first describe the architecture of the overall system, and in particular of the C19CDSS app. Then, we present the usability results and the preliminary use within our Institution.

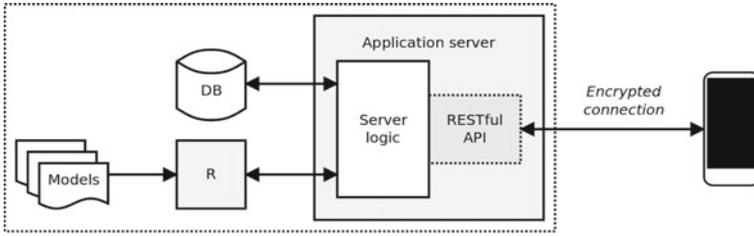


Fig. 1 System architecture

3.1 Architecture

The COVID-19 Decision Support System (C19DSS) is the system we developed to enable physicians to effectively use our models. The system is made up of a smartphone app used by clinicians, and a server that provides the “intelligence” to the app (Fig. 1).

The app is made up of four activities (see Fig. 2). The first activity is the dashboard, where a summary of the database and of the server connection status is reported. The second screen contains the patient list, how to filter patients according to different parameters, and the button to add a new patient. The third screen shows how to enter the laboratory/clinical data of a new patient, and the button to request the classification to the server. The fourth screen depicts how to edit the patient data, request the classification to the server, or delete the patient from the database.

In details, the first screen is the app dashboard (Fig. 2-a). The first card contains a summary of the data stored in the database. The second card shows the connection status with the server. By tapping on the “Go”, the user accesses the list of available patients (Fig. 2-b). If needed, the user can show a filtering panel through a menu item “Filters”. The filtering panel permits to include/exclude patients: (i) that have been classified and/or finalized (i.e., whose hospitalization has ended), or (ii) that have a name/surname containing a given string. The central part of the screen contains the list of patients (with information about name, surname, ID and status): by tapping on a patient, the user can edit the related data; by tapping on the floating action button, the user can add a new patient to the database. Figure 2-c shows the interface to introduce all data regarding a patient. On the lower part of the interface, two floating action buttons are available. The rightmost button saves the data, the leftmost requests the classification. Finally, Fig. 2-d shows how the user can edit the patient’s data. Three floating action buttons are available. From right to left, allow a user to update the data, request the classification, and delete the patient. Worth noting the two panels at the bottom of the interface, i.e., “Automated classification” and “Outcome”, that contain the results of the automated classification and the hospitalization outcome, respectively.

When the classification process is activated, the app opens an encrypted connection to the server, sends the laboratory data and the ID of the patient (so, no

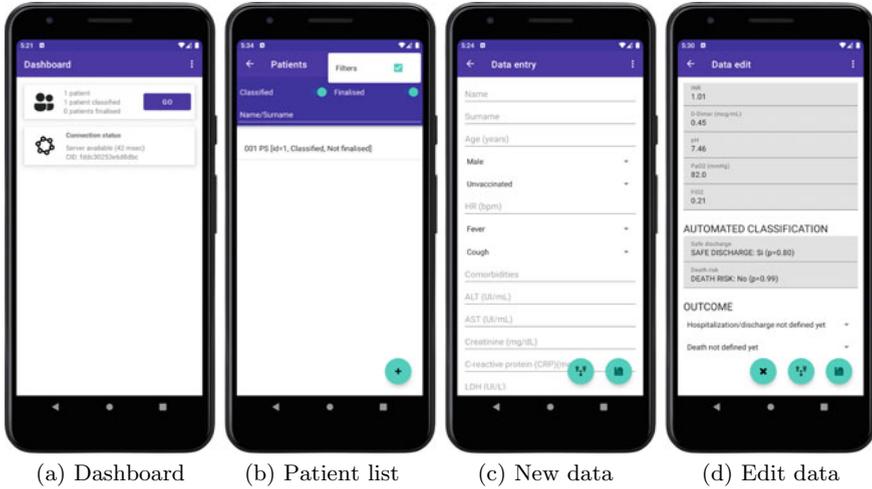


Fig. 2 C19DSS activities

personal data is ever communicated over the network) to the classification endpoint (the server follows the RESTful API paradigm) [20]. Then, the server uses R [19] to apply the correct model, depending on the request, on the received data. Hence, the server stores the received data for further analyses (i.e., to evaluate the quality of the predictions and potentially update the models), and finally returns the classification results to the app.

3.2 Usability Evaluation

To develop the app, we followed the UCD methodology, i.e., we involved the physicians from the very beginning phases of the design, and then we adapted and improved the design/implementation according to consecutive cycles of usability tests.

In the first phase, we discussed and defined with three physicians the navigational structure and the app user interface through mockups. After the system implementation, the first usability testing took place. The following three tasks were evaluated with seven physicians: (i) data entry, (ii) classification and (iii) data editing. We collected quantitative and qualitative measures based on the Single Ease Question (SEQ) and through unstructured interviews [22]. At the first iteration, we measured an average SEQ of 3.86/5, 3.71/5 and 4.00/5 for each task, and we collected a few issues and suggestions on how to improve the app. Among them, we added the automated calculation of the *P/F*, *NLR* and *PLR* values¹, we implemented a more clear

¹ *P/F* ($\text{PaO}_2/\text{FIO}_2$) = Oxygenation Index, *NLR* = Neutrophil-to-Lymphocyte Ratio, *PLR* = Platelet-to-Lymphocyte Ratio.

visualization of the classification, and we fixed a bug that blocked the classification. At the second iteration, the average SEQs increased to 4.71/5, 4.43/5 and 4.71/5.

In summary, we increased the overall average ease of completing all tasks from 3.86/5 to 4.62/5, from the first to the second implementation.

3.3 Preliminary Use

So far, the system is currently used in our Institution, as complementary support to physicians during the management of patients affected by COVID-19. The physicians that used the system reported that the application was easy and intuitive to use; the process of data entry and classification did not hamper the normal ED work routine. Conversely, it helped them to organize the workflow of COVID-19 patients.

Furthermore, with the current small cohort of patients managed through the C19DSS system, the mortality risk prediction model showed an accuracy of 0.92, whereas the model about safe discharge returned an accuracy of 0.57 (0.70 for the unvaccinated cohort). However, for safe discharge, the mistakes were conservative, i.e., the system never suggested discharging a patient that needed to be hospitalized, and took place mostly on vaccinated patients.

4 Discussion

The work presented in this paper starts from previous research finalized to devise state-of-the-art ML models, fed by routine clinical and laboratory analyses, to be used by physicians to manage safe discharge, severe disease (on the seventh day after medical presentation) and mortality during hospitalization.

Nevertheless, the models were devised from a cohort of unvaccinated patients, hence a cohort not previously immunized against SARS-CoV-2, and therefore the applicability of the models should be considered for unvaccinated patients.

At the time of writing, available data suggest long-term vaccine effectiveness in fully vaccinated healthy adults, but there are some uncertainties regarding vaccine waning in not fully vaccinated and in immunocompromised patients. Some evidence suggests that the risk of severe disease is higher in immunocompromised patients and in elderly ones [1, 13, 14]. On these bases, the app could be useful also in vulnerable patients where the immunizations seem to be less effective after a prolonged time.

In order to optimize the app performance also in fully vaccinated patients, during the data entry, for any new patient, we also save the vaccination status. So far, this information is not used by our models. However, when enough data will be collected, we could devise new models that will also consider the vaccination status. Moreover, given the client/server architecture and given that the predictions are provided by the server, the new models could be used by physicians without any change in the app, but only with a server upgrade, without affecting the user experience.

With specific regard to the C19DSS system, the adoption of the UCD methodology to design and develop the app, enabled us to gradually improve the user experience and collect useful suggestions on how to improve the overall system. Finally, the physicians that used the system reported that the application was easy and intuitive to use; the process of data entry and classification did not hamper the normal ED work routine; conversely, it helped them to organize the workflow of COVID-19 patients.

5 Conclusions

Presumably, in the next future, the SARS-CoV-2 pandemic will no longer be a global emergency, but in absence of an efficient global vaccination campaign, SARS-CoV-2 outbreaks could still be a threat to the communities where healthcare resources are limited and the immunization rate has not reached a protective stage.

Our study highlighted how AI-powered tools could be a valid support for emergency care. We do not suppose that mobile apps could replace the physician's bedside decision process, but we conceive that the interaction between emergency physicians and AI tools could improve healthcare assistance and have a significant impact on SARS-CoV-2 management.

References

1. Andrews N et al (2022) Duration of protection against mild and severe disease by COVID-19 vaccines. *N Engl J Med* 386(4):340–350. <https://doi.org/10.1056/NEJMOA2115481>
2. Bath PA (2008) Health informatics: current issues and challenges. *J Inf Sci* 34(4):501–518. <https://doi.org/10.1177/0165551508092267>
3. Booth AL, Abels E, McCaffrey P (2021) Development of a prognostic model for mortality in COVID-19 infection using machine learning. *Mod Pathol* 34(3):522–531. an official journal of the United States and Canadian Academy of Pathology, Inc. <https://doi.org/10.1038/S41379-020-00700-X>
4. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
5. Breiman L, Friedman JH, Olshen RA, Stone CJ (2017) Classification and Regression Trees. CRC Press, Boca Raton. <https://doi.org/10.1201/9781315139470>
6. Van Buuren S, Groothuis-Oudshoorn K (2011) MICE: multivariate imputation by chained equations in R. *J Stat Softw* 45(3):1–67. <https://doi.org/10.18637/JSS.V045.I03>
7. Casano N et al (2022) Application of machine learning approach in Emergency Department to support clinical decision making for SARS-CoV-2 infected patients. Submitted manuscript, under review
8. Chamoso P, De La Prieta F, Eibenstein A, Santos-Santos D, Tizio A, Vittorini P (2017) A device supporting the self management of tinnitus. In: Rojas I, Ortuño F (eds) IWBBIO 2017, vol 10209. LNCS. Springer, Cham, pp 399–410. https://doi.org/10.1007/978-3-319-56154-7_36
9. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 785–794. ACM, New York, NY, USA. <https://doi.org/10.1145/2939672>

10. Efron B (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. *J Am Stat Assoc* 78(382):316–331. <https://doi.org/10.1080/01621459.1983.10477973>
11. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K (2019) The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 25(1):30–36. <https://doi.org/10.1038/s41591-018-0307-0>
12. Jimenez-Solem E et al (2021) Developing and validating COVID-19 adverse outcome risk prediction models from a bi-national European cohort of 5594 patients. *Sci Rep* 11(1):1–12. 2021 11:1. <https://doi.org/10.1038/s41598-021-81844-x>
13. Kearns P et al (2021) Examining the immunological effects of COVID-19 vaccination in patients with conditions potentially leading to diminished immune response capacity—the OCTAVE trial. *SSRN Electron J*. <https://doi.org/10.2139/SSRN.3910058>
14. Lipsitch M, Krammer F, Regev-Yochay G, Lustig Y., Balicer RD (2021) SARS-CoV-2 breakthrough infections in vaccinated individuals: measurement, causes and impact. *Nat Rev Immunol* 22(1):57–65. 2021 22:1. <https://doi.org/10.1038/s41577-021-00662-4>
15. Liu Y, et al (2020) A COVID-19 risk assessment decision support system for general practitioners: design and development study. *J Med Internet Res* 22(6). <https://doi.org/10.2196/19786>
16. Mao JY, Vredenburg K, Smith PW, Carey T (2005) The state of user-centered design practice. *Commun ACM* 48(3):105–109. <https://doi.org/10.1145/1047671.1047677>
17. Mathieu E et al (2021) A global database of COVID-19 vaccinations. *Nat Hum Behav* 5(7):947–953. 2021 5:7. <https://doi.org/10.1038/s41562-021-01122-8>
18. McRae MP et al (2020) Managing COVID-19 with a clinical decision support tool in a community health network: algorithm development and validation. *J Med Internet Res* 22(8):e22033. <https://doi.org/10.2196/22033>
19. R Core Team: R (2018) A Language and Environment for Statistical Computing. <https://www.R-project.org/>
20. Richardson L, Ruby S (2007) RESTful Web Services. O’Reilly, Springfield
21. Skegg D et al (2021) Future scenarios for the COVID-19 pandemic. *Lancet* 397(10276):777–778. [https://doi.org/10.1016/S0140-6736\(21\)00424-4](https://doi.org/10.1016/S0140-6736(21)00424-4)
22. Tullis T, Albert W (2013) Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. Elsevier, Amsterdam
23. Vittorini P, Tarquinio A, di Orio F (2009) XML technologies for the Omaha system: a data model, a java tool and several case studies supporting home healthcare. *Comput Methods Program Biomed* 93(3). <https://doi.org/10.1016/j.cmpb.2008.10.009>
24. World Health Organization: WHO Coronavirus (COVID-19) Dashboard. <https://covid19.who.int/>
25. Yamin M (2018) IT applications in healthcare management: a survey. *Int J Inf Technol* 10(4):503–509. <https://doi.org/10.1007/s41870-018-0203-3>
26. Yao H et al (2020) Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests. *Front Cell Dev Biol* 8:683. <https://doi.org/10.3389/fcell.2020.00683>

Towards a Flexible and Portable Workflow for Analyzing miRNA-Seq Neuropsychiatric Data: An Initial Replicability Assessment



Daniel Pérez-Rodríguez, Mateo Pérez-Rodríguez, Roberto C. Agís-Balboa, and Hugo López-Fernández

Abstract In the last decade, miRNAs have attracted noticeable interest as potential biomarkers of neuropsychiatric conditions. However, a standard methodology for miRNA-Seq analysis does not yet exist, raising concerns about the reproducibility of the in-silico results and limiting their usefulness. This situation motivated us to design a miRNA-Seq pipeline specialized in the analysis of neuropsychiatric data, aiming to integrate the results of several bioinformatics tools in a highly reproducible workflow. In this study, we performed an initial test of the usefulness of our new pipeline, named myBrain-Seq, by reanalyzing four recent miRNA-Seq studies of neuropsychiatric conditions. We then compared the myBrain-Seq results with the original results and with an additional reanalysis done with another pipeline in order to make an estimation of the overall replicability. We found one of the three myBrain-Seq methodologies to be the one with best replicability, although the heterogeneity of the results and the absence of an experimental validation limits our conclusions.

D. Pérez-Rodríguez (✉) · R. C. Agís-Balboa

Translational Neuroscience Group-CIBERSAM, Galicia Sur Health Research Institute (IIS Galicia Sur), Área Sanitaria de Vigo-Hospital Álvaro Cunqueiro, SERGAS-UVIGO, 36213 Vigo, Spain
e-mail: daniel.perez.rodriguez@uvigo.es

R. C. Agís-Balboa

e-mail: roberto.carlos.agis.balboa@sergas.es

NeuroEpigenetics Lab., University Hospital Complex of Vigo, SERGAS-UVIGO, 36213 Vigo, Spain

M. Pérez-Rodríguez

Facultade de Matemáticas, Universidade de Santiago de Compostela (USC), Rúa de Lope Gómez de Marzoa, s/n, 15705 Santiago de Compostela, A Coruña, Spain
e-mail: mateo.perez.rodriguez@rai.usc.es

H. López-Fernández

CINBIO, Department of Computer Science, ESEI-Escuela Superior de Ingeniería Informática, Universidade de Vigo, 32004 Ourense, Spain
e-mail: hlfernandez@uvigo.es

SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, 36213 Vigo, Spain

Further work is required to assess myBrain-Seq' performance using a bigger dataset of studies with experimental validation data available.

Keywords miRNA-Seq · Pipeline · Neuropsychiatry · Docker

1 Introduction

MiRNAs are short non-coding RNA molecules that participate in the regulation of gene expression. They are almost ubiquitous in the regulation of the biological processes in eukaryotes, and their expression levels are known to be affected by lifestyle and environmental events such as age, diet, stress or medications [1–4]. In the last decade, these molecules have been studied as biomarkers of numerous conditions, gaining a particular importance in the field of neuropsychiatry.

Unlike other diseases, the complex etiology of mental illnesses limits their diagnosis to the identification of symptoms. Patients diagnosed with the same psychiatric disorder can present diverse clinical manifestations, often resulting in poor treatment efficacy and management. The necessity of biological biomarkers has focused the interest on miRNAs plasticity, increasing interest in the study of these molecules as biomarkers for neuropsychiatric conditions. Lots of studies have focused on finding differentially expressed miRNAs between groups of patients and healthy controls (DE miRNAs), and hundreds of these DE miRNAs have been proposed as potential candidates without being experimentally validated.

However, there is no standard methodology to perform a miRNA-Seq analysis, and this is usually performed using custom software and statistical thresholds [3, 4], and sometimes outdated reference genomes or annotations [5]. This lack of standardization raises a concern about reproducibility of the findings and questions the inference of the results to new data. To address this problem, pipelines as miARma-Seq [6] have emerged as an alternative to in-house solutions, integrating existing tools into a standardized process and offering a higher level of replicability and maintainability.

In our previous study [7], we used miARma-Seq to re-analyze the results of five miRNA-Seq studies in neuropsychiatric diseases and evaluate the reproducibility of the differential expression analysis (DEA) results. We found that only 28% of the original results were replicated with miARma-Seq on average, and tested the usefulness of the pipelines for comparing the replicability between studies. Furthermore, we hypothesized that the higher replicability of Mavrikaki et al. [8] might be related to the high quality of their raw data, perhaps as a result of using an animal model.

To the best of our knowledge, there are no pipelines for the analysis of miRNA-Seq data that offer specific results and features for the study of neuropsychiatric diseases. Since we believe that this field could greatly benefit from this type of tool, we are using Compi [9] to design an integrated solution named myBrain-Seq,¹ specifically adapted to work with miRNAs of neuropsychiatric patients. Our motivation to develop a new pipeline comes from three main aspects: (i) to tune the existing bioinformatics tools

¹ <https://www.sing-group.org/compithub/explore/625e719acc1507001943ab7f>.

used to enhance their performance with miRNA data, providing rationale default values, (ii) to focus on the analysis of data of neuropsychiatric diseases, adding specific features for different experimental designs; (iii) to provide a self-contained pipeline based on Docker, compatible with all operating systems with a Docker installation, where software versions can be easily changed and new tools can be easily added to perform custom analyses. In this study, we aim to pilot test the alpha version of myBrain-Seq by re-analyzing the same data as in our previous study [7] and comparing it to both miARma-Seq results and the original results. By doing this, we will be able to identify ways of improving our new pipeline and start enhancing its usability.

2 Materials and Methods

The dataset reanalyzed with myBrain-Seq are the same studies analyzed with miARma-Seq in our previous study [7] (Table 1), with the exception of one that was later discarded (see next section). Throughout this paper we will use the term “original studies” to refer to these five studies and “previous study” to refer to our previous miARma-Seq analysis [7].

2.1 Data Acquisition

MiRNA-Seq Data Acquisition. The raw data in FASTQ from our previous study was archived to a hard drive and retrieved for this study. More details about data

Table 1 Summary of the original studies and data to be reanalyzed

Year	Study	Organism	Contrast	Cases	Controls	Bioproject
2020	Nie et al.	Human	AD-Control PD-Control	5 AD 7 PD	34	PRJNA587017
2019	Mavrikaki et al.	Rattus norvegicus	Females I-G Males I-G	5 Females I-G 5 Males I-G	5 Females 5 Males	PRJNA543123
2018	Wang et al.	Human	ADHD-Control	1 ADHD (pool of 5 samples)	1 (pool of 5 samples)	PRJNA450485
2017	Martin et al.	Human	PTSD-Control	15 PTSD	9	PRJNA347370
2016	Hicks et al.	Human	ASD-Control	24 ASD	21	PRJNA310758
2016	Hoss et al.	Human	PD-C PDD-PDN	29 PD	33	PRJNA295431

acquisition are given in our previous study [7]. A summary of the original studies can be seen below in Table 1. As myBrain-Seq needs replicas to perform the DEA, Wang et al. 2018 [10] study was discarded because they exclusively used pooled samples.

Reference Genomes and Annotation Files. The reference genome and annotations are needed to convert the FASTQ sequences to genome coordinates first, and then to miRNA IDs. These references are dynamic and are subjected to continuous updates and modifications as the knowledge of the genome and epigenome grows. Changes on the reference genome include placement and orientation of new sequences, region relocations and deletions and inclusion/correction of alternative loci. Annotation changes include the identification of new miRNAs and the relocation or deletion of existing ones. These modifications are grouped and published in versions called assemblies or builds. All original studies aligned with old genomic builds: four of them [11–14] aligned with human genome hg19 and one [8] aligned with *Rattus norvegicus* genome rn6. Both genomes were downloaded from the NCBI Datasets [15] in their latest build: hg38 and mRatBN7.2. Regarding the annotation files, we used miRBase [16] to download the human annotations and NCBI [15] for the rat genome. Both genome and annotations were in the same version as used in our previous study using miARma-Seq [7]. This could imply that regions mapping to a miRNA in the original studies could be missing in miARma-Seq and myBrain-Seq analysis and vice versa.

2.2 *myBrain-Seq Implementation*

MyBrain-Seq is a container-based pipeline developed with Compi [9] and publicly available at Compi Hub [17]. Based on the knowledge gathered from our previous bibliographic review [5], the initial version of the pipeline comprises nine software and 14 tasks (Fig. 1), here is a brief summary of them: (1) pull-docker-images, download of the docker images from the pegi3s repository; (2) initialization, building of the directory tree for the results; (3) fastqc-qc, quality control of the fastQ files with FastQC [18]; (4) cut-sequences (optional), adapter removal with Cutadapt [19]; (5) build-genome-index (optional), genome index creation for the Bowtie alignment; (6) bowtie-alignment, alignment to the reference genome with Bowtie [20]; (7) sam-to-bam, format conversion of the Bowtie output files to bam using sam-tools [21]; (8) bam-stats, quality control of the alignments with sam-tools; (9) feature-counts, quantification and annotation with featureCounts [22]; (10) deseq, DEA with DESeq2 [23]; (11) edger, DEA with EdgeR [24]; (12) dea-integration (optional), intersection of the DESeq2 and EdgeR results and averaging of their q-values and FC; (13) venn, creation of a Venn diagram with the integrated results using VennDiagram [25] and (14) volcano, creation of a volcano plot with the DEA results using EnhancedVolcano [26].

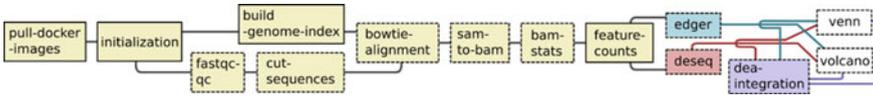


Fig. 1 Analysis workflow implemented in myBrain-Seq

Our pipeline uses independent Docker images from the `peg3s` Bioinformatics Docker Images project [27] to run the external software required in each task. This design improves the maintenance of the code, easing the version updates, and builds an isolated environment for each analysis at runtime to improve reproducibility. Also, the pipeline itself is distributed as a Docker image² and thus the only dependency required to execute it is Docker. The source code of myBrain-Seq is available at GitHub³ under a MIT License and the pipeline can be also seen at CompiHub.⁴

2.3 Experimental Setup

The analysis of the original data consisted on the following steps: (1) data arrangement into myBrain-Seq format requirement, (2) myBrain-Seq analysis, (3) quality control, (4) application of the same statistical thresholds as the original articles, and (5) comparison between the original results, miARma-Seq results and myBrain-Seq results.

Data Arrangement into myBrain-Seq Format Requirement. In addition to data, genome and annotations, the myBrain-Seq pipeline requires three additional files: (i) a parameters file with the paths of the data, references, output directory and paths to the other two additional files; (ii) a TSV file with sample names, conditions and labels; and (iii) a text file with the conditions to compare and a label for the contrast. Although these files are myBrain-Seq-specific, the information required to build the latter two files is similar to that used in our previous analysis. In order to diminish the probability of errors we adapted, whenever possible, the miARma-Seq files to myBrain-Seq files.

myBrain-Seq Analysis. MyBrain-Seq analysis was performed using all the tasks described in the Sect. 2.2 excepting for the adapter removal task. This task was only performed in the Hoss et al. 2016 samples as they explicitly state the adapter sequence.

Quality Control. To assess the performance of the myBrain-Seq analysis, MultiQC [28] reports were generated with the results of each study. We look for mapping and assignment rates in order to discard studies with low-quality data.

² <https://hub.docker.com/r/singgroup/my-brain-seq>.

³ <https://github.com/sing-group/my-brain-seq>.

⁴ <https://sing-group.org/compihub/explore/625e719acc1507001943ab7f#readme>.

Application of the Same Statistical Thresholds as the Original Articles. Results of DESeq2, EdgeR, and the integrated ones were filtered in a LibreOffice spreadsheet using the same criteria as in the original articles, namely: false discovery rate-corrected p-value (q-value) < 0.05 on the Mavrikaki [8], Hoss [14] and Martin [11] studies and q-value < 0.05 , fold change lower than -1 or greater than 1 on Nie [12] study. The resulting miRNAs were stored in an SQLite database, identified by study and contrast, to ease further queries and comparisons.

Comparison Between the Original Results, miARma-Seq Results and myBrain-Seq Results. Prior to the DEA result comparisons, the miRNAs annotations format was adapted to that used on the original studies: myBrain-Seq results were obtained in format “hsa-miR.../rnor-miR...” and were adapted to “miR-” nomenclature for the Martin [11] and Nie [12] studies by removing the “hsa-” prefix.

We aim to perform two comparisons: (1) myBrain-Seq results vs. miARma-Seq results and (2) myBrain-Seq results vs. original results. We chose the Jaccard index (J) as an estimator for the similarity between results (Fig. 2). To calculate J we considered n to be the number of contrasts in a study, therefore, for each contrast we could define pipeline results as three sets:

$$\begin{aligned} \text{MBS}_1^i &= \{ \text{DE miRNAs of myBrain-Seq using DESeq2} \}, \\ \text{MBS}_2^i &= \{ \text{DE miRNAs of myBrain-Seq using EdgeR} \}, \\ \text{MBS}_3^i &= \text{MBS}_1^i \cap \text{MBS}_2^i, \end{aligned}$$

where $i \in \{1, \dots, n\}$. Denote by $\text{ref}^i = \{ \text{DE miRNAs} \}$ the set of DE miRNAs found either in the original studies or miARma-Seq. Let us finally consider the number of coincidences in the i -th contrast using k software,

$$\text{coincidences}_k^i = | \text{MBS}_k^i \cap \text{ref}^i |, \quad i \in \{1, \dots, n\}, \quad k \in \{1, 2, 3\}.$$

Whenever $\text{MBS}_k^i \neq \emptyset$ and $\text{ref}^i \neq \emptyset$, we consider the Jaccard index of the i -th contrast using k software defined as

$$J_k^i = \frac{|\text{coincidences}_k^i|}{|\text{MBS}_k^i \cup \text{ref}^i|}, \quad i \in \{1, \dots, n\}, \quad k \in \{1, 2, 3\}.$$

Fig. 2 Graphical representation of the Jaccard index as Venn sets

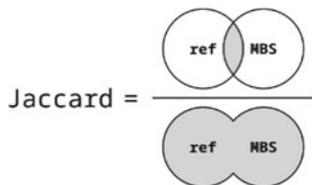


Table 2 Average number of sequences per study successfully aligned (% Alignment) and assigned (% Assignment)

Study	Contrast	% Alignment	% Assignment
Mavrikaki et al.	Females I-G, Males I-G	98.41	30.98
Hoss et al.	PD-C, PDD-PDN	94.20	39.77
Martin et al.	PTSD-Control	82.68	45.36
Nie et al.	AD-Control, PD-Control	64.03	10.24
Hicks et al.	ASD-Control	13.61	1.86

Note that the case $MBS = ref = \emptyset$ is interpreted as complete agreement, which means that $J = 1$.

3 Results and Discussion

3.1 Quality Control

MultiQC reports [28] were generated to summarize the Bowtie [20] and Feature-Counts [22] results. Alignment rates (Table 2) were above 80% for all samples except in Nie [12] (64.03%) and Hicks [13] (13.61%) studies. Similarly, assignment rates averaged 39% except for the Nie (10.24%) and Hicks (1.86%) samples. As done in our previous study, we discarded the Hicks et al. study for subsequent comparisons.

3.2 Application of the Same Statistical Thresholds as the Original Articles

Statistical criteria of the original articles were applied on myBrain-Seq results and names of the resulting miRNAs were stored in the SQLite database. Only in the Martin study [11] myBrain-Seq did not suggest any DE miRNAs.

3.3 Comparison Between the Original Results, MiARma-Seq Results and MyBrain-Seq Results

To get a first impression of the performance of myBrain-Seq, we re-analyzed the same data as in our previous study [7] and compared the results with those of miARma-Seq and the original studies. Table 3 contains all the comparisons mentioned in Sect. 2.3 along with the results of our previous study. In this section, we will refer

to myBrain-Seq DESeq2 results as MBS_D, to myBrain-Seq EdgeR results as MBSE and to myBrain-Seq integrated results as MBS_I (these are simply the intersection of MBS_D and MBSE). We will also use the term “methodologies” to refer to the original, miARma-Seq and myBrain-Seq analyses.

It can be observed from Table 3 that myBrain-Seq was the methodology with fewest DE miRNAs predictions, with MBS_D having more results on average than MBS_E and the latter more than MBS_I. The number of results per study and contrast was highly variable across methodologies, having Hoss et al. study [13] the contrast with more DE miRNA predictions, PD-control with 312 results in average, and fewer predictions, PDD-PDN with 0 in all the analysis including the original. This absence of results across all the analysis determines that PDD-PDN is the only contrast with J values of 1 for all the methodologies (see J definition in Sect. 2.3), but the relevance of this finding is limited, since it could be the result of comparing very unrelated groups.

As far as replicability is concerned, myBrain-Seq results were, on average, more similar to miARma-Seq results ($J = 0.251$) than to the original results ($J = 0.143$) (Table 4). Specifically, the comparisons between the results of MBS_D and miARma-Seq achieved the highest degree of replicability, with an average J of 0.37. Notably, in the case of the PD-control contrast in Hoss et al. [14], J is about 0.7 between miARma-Seq and myBrain-Seq (Table 3). This higher replicability could be the result of using the same reference genome and annotations in both pipelines, as well as the result of using similar software for alignment and quantification.

The first row of Table 4 shows the replicability of the original results taking in account the J values of miARma-Seq and myBrain-Seq: it could be interpreted as only 16% of the original results were replicated by these both pipelines. In the same line, miARma-Seq replicated 22.5% of the original results and myBrain-Seq 14.3% on average. The contrast best replicated was that of MBS_D vs. miARma-Seq ($J = 0.372$ on average), and the one with least replicability was MBS_I vs. original studies ($J = 0.117$ on average). MBS_E was the methodology with less predictions and greatest fluctuations in number of results, with almost the same number of predictions as the other methodologies in the Hoss et al. [14] contrast PD-control and Mavrikaki [8] contrast SM-GM, and up to 68 times less predictions in the Mavrikaki IF-GF contrast and 35 times less in Nie [12] AD-NC contrast. This result must be analysed further as miARma-Seq also uses edgeR but averages higher J values.

If we now turn to the results per study, Hoss et al. [14] was the study with highest replicability, with an overall average of $J = 0.713$, followed by Martin et al. study [11] with 0.286, Mavrikaki et al. study [8] with 0.214, and finally Nie et al. study [12], with 0.059. Poor replicability of the Nie et al. study [12] could be explained by the low alignment and assignment rates of their data, and appears to be consistent with the higher replicability of the Hoss and Martin studies [11, 14], as they both had the best alignment rates. In Mavrikaki [8] SM-GM contrast, myBrain-Seq methodologies made almost the same number of predictions, more than miARma-Seq and the original studies, but the J of such predictions was half of the J obtained

Table 4 Average Jaccard index (J) of each methodology for all the studies, with MBS_D being myBrain-Seq DESeq2 results, MBS_E myBrain-Seq EdgeR results, MBS_I myBrain-Seq integrated results, (M) miARma-Seq results and (O) the original results. In the “Summary” column, “All vs. O” is the average J of all the comparisons of MBS with O and M with O, whereas “M vs. O” is the average J result of all the comparisons of M with O

Summary	myBrain-Seq vs. Original		myBrain-Seq vs. miARma-Seq	
All vs. O	MBS_D vs. O	0.185	MBS_D vs. M	0.372
0.16	MBS_E vs. O	0.126	MBS_E vs. M	0.157
M vs. O	MBS_I vs. O	0.117	MBS_I vs. M	0.225
0.225	MBS vs. O	0.143	MBS vs. M	0.251

by miARma-Seq. Finally, in Nie AD-NC contrast, myBrain-Seq had better J values than miARma-Seq; both methodologies struggled to replicate the results of the Nie study [12].

4 Conclusion

MiRNAs are emerging as promising biomarkers for neuropsychiatric diseases, however there is still no standard methodology to perform a miRNA-Seq analysis. In this study, we aimed to test the replicability of a new pipeline that we are still developing, known as myBrain-Seq, which will be specially oriented to the analysis of miRNAs in neuropsychiatric data. To do that, we re-analyzed the same data as in our previous study [7] and compared it to both miARma-Seq results and the original results. We found that the results best replicated were those of myBrain-Seq when using DESeq2 software for differential expression analysis. However, replicability varied widely across studies, suggesting that it might be strongly related to the quality of the raw data. In addition, we found more similarities between miARma-Seq and myBrain-Seq results which could be attributed to the use of similar software for annotation and quantification. This would emphasize the influence of the methodology in drawing conclusions from the same data. On the other hand, the usefulness of the integrated methodology was limited due to the reduced number of EdgeR results, although it performed better than the EdgeR results alone. Additionally, as there is no experimental validation data in the original studies, our conclusions are limited by the predictions made by each methodology. Terms such as “good” and “bad” performance are related to how many results were replicated *in silico*, but are not a representation of true biological positives. In a future study, we would like to test and tune myBrain-Seq performance using a bigger dataset of studies with experimental validation data available. We will also start using our pipeline to analyze miRNA-Seq data from ongoing projects in our laboratory.

Acknowledgements This study was partially supported by: (i) Instituto de Salud Carlos III through the project PI18/01311 (co-funded by European Regional Development Fund, “A way to make

Europe”) to R.C. Agís-Balboa, and (ii) Consellería de Educación, Universidades e Formación Profesional (Xunta de Galicia) under the scope of the strategic funding ED431C2018/55-GRC Competitive Reference Group. H. López-Fernández is supported by a “María Zambrano” post-doctoral contract from Ministerio de Universidades (Gobierno de España). D. Pérez-Rodríguez is supported by an “Investigo program” predoctoral contract from Xunta de Galicia.

References

1. Esteller M (2011) Non-coding RNAs in human disease. *Nat Rev Genet* 12:861–874. <https://doi.org/10.1038/nrg3074>
2. Winkle M, El-Daly SM, Fabbri M, Calin GA (2021) Noncoding RNA therapeutics - challenges and potential solutions. *Nat Rev Drug Discov* 20:629–651. <https://doi.org/10.1038/s41573-021-00219-z>
3. Gebert LFR, MacRae IJ (2019) Regulation of microRNA function in animals. *Nat Rev Mol Cell Biol* 20:21–37. <https://doi.org/10.1038/s41580-018-0045-7>
4. Zovoilis A, Agbemenyah HY, Agis-Balboa RC, Stilling RM, Edbauer D, Rao P, Farinelli L, Delalle I, Schmitt A, Falkai P, Bahari-Javan S, Burkhardt S, Sananbenesi F, Fischer A (2011) microRNA-34c is a novel target to treat dementias. *EMBO J* 30:4299–4308. <https://doi.org/10.1038/emboj.2011.327>
5. Pérez-Rodríguez D, López-Fernández H, Agís-Balboa RC (2021) Application of miRNA-seq in neuropsychiatry: a methodological perspective. *Comput Biol Med* 135:104603 (2021). <https://doi.org/10.1016/j.combiomed.2021.104603>
6. Andrés-León E, Núñez-Torres R, Rojas AM (2016) miARma-Seq: a comprehensive tool for miRNA, mRNA and circRNA analysis. *Sci Rep* 6:25749. <https://doi.org/10.1038/srep25749>
7. Pérez-Rodríguez D, López-Fernández H, Agís-Balboa RC (2022) On the reproducibility of MiRNA-Seq differential expression analyses in neuropsychiatric diseases. In: Rocha M, Fdez-Riverola F, Mohamad MS, Casado-Vara R (eds) Practical applications of computational biology & bioinformatics, 15th international conference (PACBB 2021). Springer, Cham, pp 41–51. https://doi.org/10.1007/978-3-030-86258-9_5.
8. Mavrikaki M, Pantano L, Potter D, Rogers-Grazado MA, Anastasiadou E, Slack FJ, Amr SS, Ressler KJ, Daskalakis NP, Chartoff E (2019) Sex-dependent changes in miRNA expression in the bed nucleus of the stria terminalis following stress. *Front Mol Neurosci* 12. <https://doi.org/10.3389/fnmol.2019.00236>
9. López-Fernández H, Graña-Castro O, Nogueira-Rodríguez A, Reboiro-Jato M, Glez-Peña D (2021) Compi: a framework for portable and reproducible pipelines. *PeerJ Comput Sci* 7:e593. <https://doi.org/10.7717/peerj-cs.593>
10. Wang LJ, Li SC, Lee MJ, Chou MC, Chou WJ, Lee SY, Hsu CW, Huang LH, Kuo HC (2018) Blood-Bourne microRNA biomarker evaluation in attention-deficit/hyperactivity disorder of Han Chinese individuals: an exploratory study. *Front Psychiat* 9. <https://doi.org/10.3389/fpsy.2018.00227>
11. Martin CG, Kim H, Yun S, Livingston W, Fetta J, Mysliwiec V, Baxter T, Gill JM (2017) Circulating miRNA associated with posttraumatic stress disorder in a cohort of military combat veterans. *Psychiatry Res* 251:261–265. <https://doi.org/10.1016/j.psychres.2017.01.081>
12. Nie C, Sun Y, Zhen H, Guo M, Ye J, Liu Z, Yang Y, Zhang X (2020) Differential expression of plasma Exo-miRNA in neurodegenerative diseases by next-generation sequencing. *Front Neurosci* 14. <https://doi.org/10.3389/fnins.2020.00438>
13. Hicks SD, Ignacio C, Gentile K, Middleton FA (2016) Salivary miRNA profiles identify children with autism spectrum disorder, correlate with adaptive behavior, and implicate ASD candidate genes involved in neurodevelopment. *BMC Pediatrics* 16. <https://doi.org/10.1186/s12887-016-0586-x>

14. Hoss AG, Labadorf A, Beach TG, Latourelle JC, Myers RH (2016) microRNA profiles in Parkinson's disease prefrontal cortex. *Front Aging Neurosci* 8. <https://doi.org/10.3389/fnagi.2016.00036>
15. NCBI Datasets. <https://www.ncbi.nlm.nih.gov/datasets/>. Accessed 11 May 2021
16. Kozomara A, Birgaoanu M, Griffiths-Jones S (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res* 47:D155–D162. <https://doi.org/10.1093/nar/gky1141>
17. Nogueira-Rodríguez A, López-Fernández H, Graña-Castro O, Reboiro-Jato M, Glez-Peña D (2021) Compi Hub: a public repository for sharing and discovering Compi pipelines. In: Panuccio G, Rocha M, Fdez-Riverola F, Mohamad MS, Casado-Vara R (eds) *Practical applications of computational biology & bioinformatics*, 14th international conference (PACBB 2020), pp 51–59. Springer, Cham. https://doi.org/10.1007/978-3-030-54568-0_6
18. Andrews S (2010) FASTQC. A quality control tool for high throughput sequence data
19. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal* 17:10–12. <https://doi.org/10.14806/ej.17.1.200>
20. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25. <https://doi.org/10.1186/gb-2009-10-3-r25>
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) 1000 genome project data processing subgroup: the sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
22. Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30:923–930. <https://doi.org/10.1093/bioinformatics/btt656>
23. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15:550. <https://doi.org/10.1186/s13059-014-0550-8>
24. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140. <https://doi.org/10.1093/bioinformatics/btp616>
25. Chen H, Boutros PC (2011) VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinform* 12:35. <https://doi.org/10.1186/1471-2105-12-35>
26. Blighe K (2022) EnhancedVolcano: publication-ready volcano plots with enhanced colouring and labeling
27. López-Fernández H, Ferreira P, Reboiro-Jato M, Vieira CP, Vieira J (2021) The pegi3s bioinformatics docker images project. In: Rocha M, Fdez-Riverola F, Mohamad MS, Casado-Vara R (eds) *Practical applications of computational biology & bioinformatics*, 15th international conference (PACBB 2021). Springer, pp 31–40
28. Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32:3047–3048. <https://doi.org/10.1093/bioinformatics/btw354>

The NAD Interactome, Identification of Putative New NAD-Binding Proteins



Sara Duarte-Pereira , Sérgio Matos, José Luís Oliveira, and Raquel M. Silva

Abstract Nicotinamide adenine dinucleotide (NAD) is an essential metabolite in normal cellular physiology and its deregulation may lead to several pathological conditions. NAD interacts with a vast number of proteins, acting as a coenzyme, as a substrate and regulating the interaction between proteins. The goals of this study were to characterize the proteins involved in NAD metabolism and to identify putative new NAD regulated proteins. Using an in silico approach, we first defined a NAD-binding dataset, that we characterized through pathway enrichment analysis and protein structural domains analysis. We then screened the full human proteome and further analyzed a selection of potential NAD-binding proteins. This global study of the NAD interactome resulted in the identification of new potentially NAD-binding proteins (NADPBs), including TRPC3 and a few isoforms of DGA kinases, which are involved in calcium signaling. NADPBs participate in several metabolic pathways and signaling processes in the cell, while proteins interacting with NADPBs are mostly involved in signaling pathways, including pathways related to disease, namely three major neurodegenerative diseases, Alzheimer's, Huntington's, and Parkinson's.

Keywords Nicotinamide Adenine Dinucleotide (NAD) · Protein-protein interactions (PPIs) · Protein domains

S. Duarte-Pereira (✉) · S. Matos · J. L. Oliveira
IEETA – Institute of Electronics and Informatics Engineering of Aveiro, 3810-193 Aveiro, Portugal
e-mail: sdp@ua.pt

S. Duarte-Pereira · R. M. Silva
Department of Medical Sciences and iBiMED – Institute of Biomedicine, University of Aveiro, 3810-193 Aveiro, Portugal

S. Matos · J. L. Oliveira
DETI – Department of Electronics, Telecommunications and Informatics, University of Aveiro, 3810-193 Aveiro, Portugal

R. M. Silva
Faculty of Dental Medicine, Center for Interdisciplinary Research in Health (CIIS), Universidade Católica Portuguesa, 3504-505 Viseu, Portugal

1 Introduction

Proteins are responsible for virtually all fundamental cellular processes and often their function depends on the physical interaction with other proteins and with small molecules. Thus, protein-protein or protein-metabolite interactions are highly specific and regulated. Over the years, the study of protein-protein interactions has shown that the specificity of the interaction between a protein and its target depends mostly on the structure of the interface of the two proteins and that the interaction patterns between similar proteins or domains are more conserved than their amino acid sequence [1]. On the one hand, structural analysis showed similarities between interfaces and, on the other, a preference of certain amino acids on protein interfaces has been observed, which differ from the amino acids at the periphery of the protein [2]. In addition, protein domains are usually associated to a specific protein function or interaction. Therefore, the analysis of both the sequence and the structural units of a protein may lead to the prediction of associated functions [3].

Nicotinamide adenine dinucleotide (NAD) is a small molecule essential for cellular functions such as energy metabolism, by acting as a coenzyme in redox reactions in several metabolic pathways. Additionally, NAD serves as a substrate for proteins involved in critical physiological processes, such as transcription regulation, DNA damage repair, calcium signaling, cell survival, among others. The major groups of NAD-dependent enzymes are sirtuins (SIRTs) [4], poly- and mono-(ADP-ribose) polymerases (PARPs and MARTs) [5], and cyclic ADP-ribose hydrolases, such as CD38 [6].

More recently, a third role for NAD has been suggested, where NAD would function as a direct regulator of protein-protein interactions (PPIs). In their report, Li and collaborators [7] have shown that NAD binds to the NUDIX homology domain (NHD) of the Deleted in Breast Cancer 1 (DBC1) protein, preventing its interaction with PARP1. PARP1 is one of the most important players in the DNA damage repair process and the DBC1-PARP1 interaction inhibits PARP1 normal function. On the other side, DBC1 regulates the activity of several proteins with key roles in the cellular physiology, such as the transcription factor p53 [8].

Given the vast range of NAD functions in the cell and the importance of NAD metabolism in the normal physiology of the cell and pathological conditions, in this study we first aimed to characterize the proteins involved in NAD metabolism. Due to NAD specific role of regulating PPIs, we focused on NAD-binding proteins and their interactions, aiming to identify putative new NAD regulated proteins. To achieve these goals, we chose an *in-silico* approach where we first defined a NAD-binding dataset, that we characterized through pathway enrichment analysis and protein structural domains analysis. We then screened the full human proteome and further analyzed a selection of potential NAD-binding proteins.

2 Methods

2.1 Data Collection and Definition of Protein Datasets

NAD-Binding Proteins (NADBPs) Dataset. We defined a first dataset composed by proteins known to bind the NAD molecule. We searched for NAD-protein interactions in several chemical databases, based on experimental data. Namely, we downloaded data from the Human Metabolome Database¹ [9] (accession HMDB0000902), from the STITCH database v.4² [10] (beta-NAD, all types of evidence, confidence level of 0.9), from Drugbank v.5.0³ [11] (accession DB00157), from ChEMBL database⁴ [12] release 23 (ChEMBL1234613), from PubChem⁵ [13], (ID 5893) and from Protein Data Bank⁶ (PDB) [14] (NAD as a ligand). For all proteins found, we mapped either the names or IDs obtained from each database to the corresponding UniProt ID⁷ [15], and removed duplicates. The resulting dataset was composed by all the proteins identified in the interactions from these six databases.

NAD-Protein-Protein Interactions (NAD-PPIs) Dataset. We then built a dataset composed by the proteins that interact with the NAD-binding proteins, i.e., the NAD-PPIs dataset. For that purpose, we searched for the interactions of the proteins from the NADBPs dataset, using three main sources: BIOGRID⁸ [16], STRING⁹ v.10 [17] and IMEX Consortium¹⁰ [18]. On STRING database, we selected only the interactions with highest confidence (>0.9) of the combined score provided. We merged the results from these three databases, by identifying unique interactions and by removing the duplicates, and mapped the proteins to the UniProt ID.

2.2 Datasets Analysis

Gene Ontology (GO) Analysis. To perform a GO analysis, we used PANTHER¹¹ [19] overrepresentation test (Fisher's exact, False Discovery Rate correction), using

¹ <https://hmdb.ca/>.

² <https://stitch4.embl.de/>.

³ <https://go.drugbank.com/>.

⁴ <https://www.ebi.ac.uk/chembl/>.

⁵ <https://pubchem.ncbi.nlm.nih.gov/>.

⁶ <https://www.rcsb.org/>.

⁷ <https://www.uniprot.org/>.

⁸ <https://thebiogrid.org/>.

⁹ <https://string-db.org/>.

¹⁰ <http://www.imexconsortium.org/>.

¹¹ <http://pantherdb.org/>.

the Pathways annotation dataset (version 13.0). We analyzed the NAD-binding and the NAD-PPIs datasets.

Protein Structural Domain Analysis. To identify the most frequent protein domains and protein families within the NADDBPs dataset, we used PFAM¹² database [20]. We performed a batch search using as input a fasta file containing the NADDBPs sequences, retrieved from UniProt. The analysis was executed through the HMMER software [21], that uses profile hidden Markov models.

We also obtained the domain analysis of the full human Uniprot proteome (NCBI tax. ID 9606) and selected the results with an E-value below 1. We considered all human reviewed proteins from the Uniprot database as a reference dataset, in a total of 20,303 proteins, and the unreviewed proteins as a test dataset, in a total of 50,588 proteins.

2.3 Identification of Putative New NAD-Binding Proteins

From the protein domains obtained from the NADDBPs dataset, we identified 15 domains that appeared in more than 10 proteins in the dataset. Then, we retrieved from both reference and test datasets the proteins that presented at least one of those 15 domains. All protein fragments were excluded and the mapping to a corresponding single gene identifier was performed. The genes/proteins that were found exclusively within the test dataset of unreviewed proteins were identified.

We analyzed the resultant proteins from the test dataset using the NADbinder¹³ [22] to predict the number of NAD interacting residues. The amino acid fasta format sequence of each protein was used as input, and a threshold of 0.3 was selected.

The STRING database v.11 [17] was used to obtain the interactions of each of those proteins. Only interactions based on experiments were retrieved, with a 0.4 confidence level.

3 Results

3.1 NAD-Binding Proteins (NADDBPs) Dataset

The combination of the data obtained from the different databases resulted in a NADDBPs dataset with a total of 439 proteins. Around 80% of these proteins were enzymes, most with catalytic activity, involved in metabolite interconversion. The major protein classes found were dehydrogenases (92 proteins), from which over 30

¹² <http://pfam.xfam.org/>.

¹³ <http://crdd.osdd.net/raghava/nadbinder/>.

were NADH dehydrogenase, and oxidoreductases (55 proteins), but several others were identified. In addition to enzymes that use NAD as cofactor in redox reactions, there were also enzymes that use NAD as a substrate, such as all SIRT's and all PARPs. Considering their molecular function, a small number of proteins were involved in regulation or transporter activities. Of note, over one hundred proteins corresponded to mitochondrial isoforms of enzymes, mostly involved in the chain of reactions responsible for ATP production.

3.2 NAD-Protein-Protein Interactions (NAD-PPIs) Dataset

After mapping every ID retrieved from each database to the UniProtKB ID, with reviewed annotation, we removed the duplicated entries that were mainly due to gene or protein alternative names. We then identified the proteins common to the three sources of PPIs, remaining with a final list of 10,020 proteins involved in PPIs with NADBPs. For further analysis, we considered the 1368 proteins that were common to all databases.

3.3 Pathway Enrichment Analysis

We performed a GO analysis on the the NAD-PPIs and the NADBPs datasets (Fig. 1).

Pathways specific of the NADBPs dataset were related to biosynthesis or metabolism of nucleic acids, carbohydrates, and amino acids, while NAD-PPIs dataset presented an enrichment in several signaling pathways. The pathways with the highest number of genes (over 50) were related to hormone receptors signaling, namely for gonadotropin and for the gastrointestinal peptide hormones cholecystokinin and gastrin, followed by the Wnt signaling and angiogenesis pathways. Several other pathways were related to hormone or growth factor signaling, and disease pathways also emerged, namely three major neurodegenerative diseases, Alzheimer's, Huntington's, and Parkinson's.

3.4 Characterization of Protein Domains of the NADBPs Dataset

From the analysis of the protein domain performed on the 439 NADBPs through PFAM database, 1101 identifications were made, which corresponded to a total of 412 different domains, that belonged to a total of 114 clans. More than half of the proteins (56%–247 proteins) belonged to the FAD/NAD(P)-binding Rossmann fold

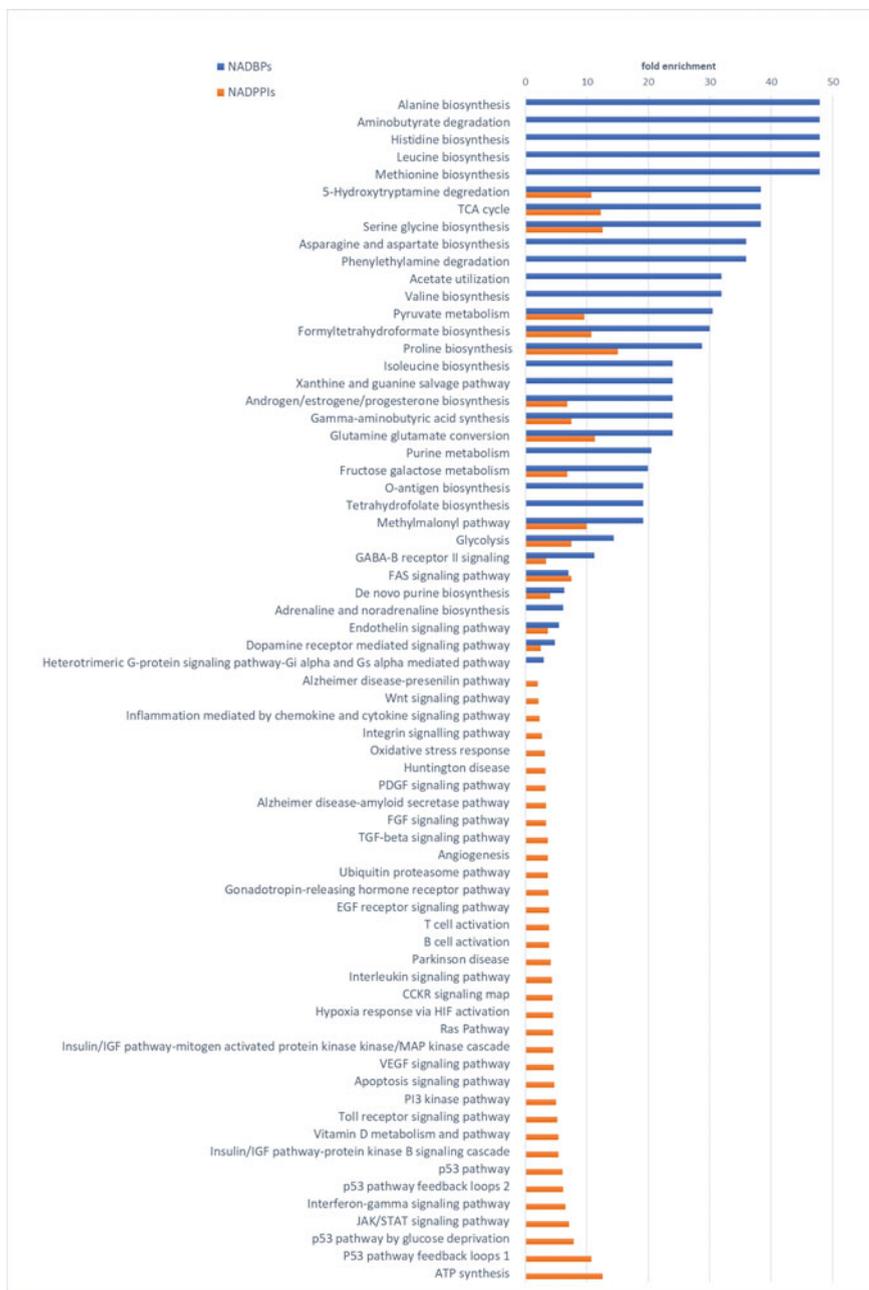


Fig. 1 Graphical representation of the Pathway enrichment analysis of the NADBPs (blue bars) and the NAD-PPIs (orange bars) datasets

superfamily (clan id:CL0063), and 27% belonged to the Ankyrin repeat superfamily (clan id: CL0465).

The 15 more common domains appeared in more than ten proteins. They included five different ankyrin repeats, the short chain dehydrogenase, the aldehyde dehydrogenase family, the cytochrome P450 and the poly(ADP-ribose) polymerase (PARP) catalytic domain.

The NUDIX domain was found only in two proteins from the NAD-binding dataset, namely NUDT12 and NUDT7.

3.5 Identification of Putative NAD-Binding Proteins

Top Domains from the NADBP Dataset. We searched for the 15 domains that were identified in ten or more proteins from the NADBP dataset within the dataset of the full human proteome unreviewed proteins (test dataset) and obtained 901 protein sequences. After removing all protein fragments and duplicates, we identified 255 proteins, which corresponded to 204 single genes. We performed a similar approach in the reference dataset and obtained 474 genes. Given our aim to identify uncharacterized proteins, from the 204 genes, we excluded 195 that were also identified in the reference dataset and 8 genes remained, corresponding to 13 protein sequences, found uniquely in the test dataset.

Among the 13 proteins, there were five isoforms of the Diacylglycerol (DAG) kinase, four encoded by the DGKI gene (UniProt IDs: A0A087WV00, E7EM72, E7EWQ4 and E9PFX6) and one encoded by DGKZ gene (E9PNL8). There were two other kinase isoforms, from the Leucine-rich repeat serine/threonine-protein kinase 1, encoded by the LRRK1 gene (E9PK39 and E9PMK9). There were also two proteins related to membrane transport, the Sodium/hydrogen exchanger 9B2 (SLC9B2 gene, UniProt ID D6R9P2) and two isoforms of a short transient receptor potential channel encoded by the TRPC3 gene (D6RC49 and J3QTB0). A smaller isoform of the POTE member of the ankyrin family was also found (Q495V5). Of note, POTE was the only protein that presented simultaneously two of the 15 domains (Ank_2 e Ank_5). Additionally, there were two proteins resultant from the readthrough of two genes, CYP3A7-CYP3A51P (UniProt ID A0A087WV96), which belong to a subfamily of the Cytochrome P450, and FPGT-TNNI3K (UniProt ID V9GXZ4), from the neighboring fucose-1-phosphate guanylyltransferase (FPGT) and TNNI3 interacting kinase (TNNI3K) genes.

NADbinder Analysis. We further analyzed the 13 identified proteins using the NADbinder software (Fig. 2). Here, instead of the protein structure, the protein sequence is considered. The highest number of NAD-interacting residues (33) was identified in the longest isoform of TRPC3, with 793 amino acids, followed by the longest isoform of DGKI with 1078 amino acids, where 31 residues were identified. We observed a positive correlation between the amino acid length and the number of NAD-interacting residues identified.

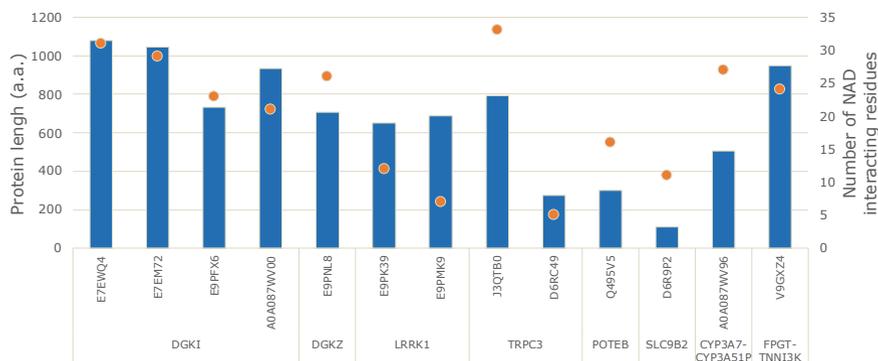


Fig. 2 NADbinder results for the potential NAD-binding proteins identified. For each protein, the number of NAD residues (orange dots) was identified from the protein sequence, which length is represented by the blue bars

Protein-Protein Interactions. To evaluate the possibility that NAD has an impact on the interactions between these proteins, we further searched for the interactions of each of the proteins (Fig. 3). DGKI and SLC9B2 had no reported interactions, as well as the proteins resultant from the two readthrough events. LRRK1 had the highest number of interactions, followed by TRPC3.

Among the 23 proteins that interact with the potential NAD-binding proteins, seven were already present in the NAD-PPIs dataset described previously, meaning that they were also found among the proteins that interact with known NAD-binding proteins. They were the following: the SHC-transforming protein 1 (SHC1), the phospholipase C gamma 1 (PLCG1), the serine/threonine Polo-like kinase 1 (PLK1), the leucine rich repeat kinase 2 (LRRK2), the FKBP (FK506-Binding Protein) Prolyl Isomerase 5 (FKBP5), the 14–3-3 protein sigma Stratifin (SFN), also known as the Epithelial cell marker protein 1, and the Retinoblastoma-associated protein (RB1). Among the interactors, we found that PLCG1 interacts both with TRPC3 and DGKZ.

4 Discussion

To accomplish the diversity of functions in which NAD participates within the cell, this small molecule binds to a high number of different proteins. In those reactions, NAD can play one of three roles: a) acts as an enzymatic cofactor in redox reactions, b) is consumed by NAD-dependent enzymes, and c) intervenes in protein-protein interactions, therefore regulating several cellular processes. In this exploratory study, our approach to identify potential NAD-binding proteins, led us to a global analysis of the NAD interactome.

The pathway enrichment analysis revealed a diversity of cellular pathways in which the NADBP are involved. Interestingly, the comparison with the NAD-PPIs

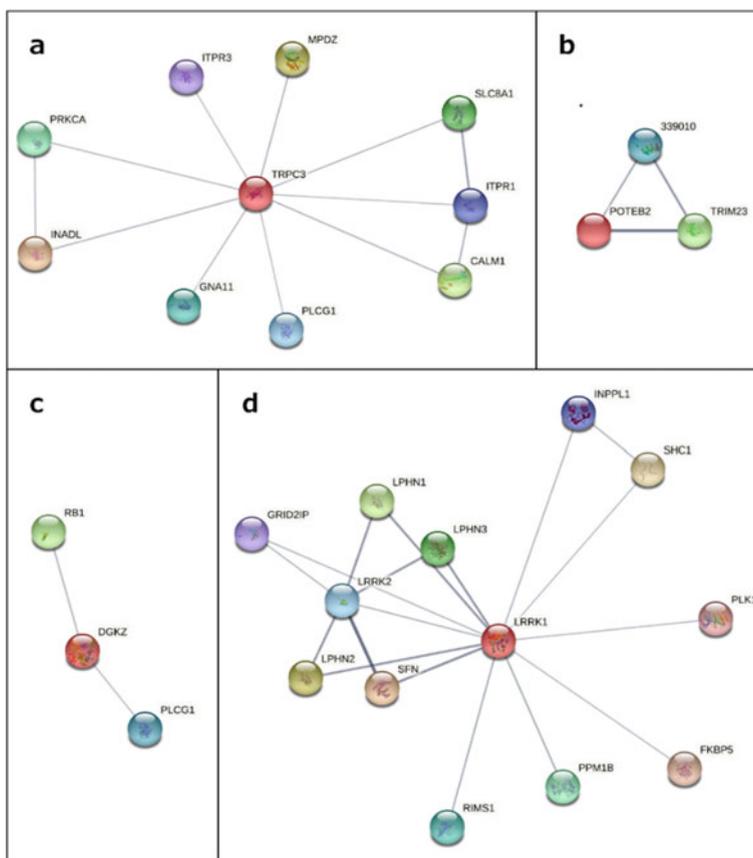


Fig. 3 Protein-protein interactions of potential NAD-binding proteins. **a** TRPC3, **b** POTEB, **c** DGKZ and **d** LRRK1. Queried proteins are represented by red nodes and the line thickness indicates the confidence level of the interaction. Only physical interactions are represented. The network was obtained through STRING (string-db.org)

dataset highlighted the key role of NADDBPs in basic metabolism and biosynthetic processes. Nevertheless, several pathways were common to both datasets, such as glycolysis and TCA cycle, which are essential metabolic pathways, or signaling pathways mediated by GABA or dopamine receptors, for example. On the other hand, the pathways found in NAD-PPIs dataset analysis, showed how vast is the action of this small molecule. Besides the large number of proteins that directly interact with NAD, the highest number of protein interactions in which it participates is related to critical signaling pathways, from development and apoptosis to general immune and hormone responses, and including many disease pathways.

Then, our pursue for previously unidentified proteins as NADDBPs was primarily based on the presence of the most frequent domains and then on the presence of NAD-interacting residues.

Besides being, by definition, a structural protein motif that binds nucleotides, particularly NAD, NADP and FAD, the Rossmann fold is also one of the most common folds found in all human proteome [23]. So, it was expected to find that the majority of the NADBPs dataset belonged to this superfamily of proteins. In what concerns the protein domains, the ankyrin repeats were the most frequent, and some proteins presented more than one ankyrin repeat in their structures. The ankyrin domain is also very frequent in all human proteome and it mediates PPIs [24].

In addition to protein structural domains, from which only the frequency was evaluated, we considered the number of NAD interacting residues. In fact, the direct binding of NAD at specific sites of a protein ultimately determines its action [22]. In the case of the report from Li and collaborators [7], the NAD binding to the NUDIX homology domain of DBC1 regulates its action on PARP1, by preventing the interaction between the two proteins. In their study, they identified no more than 10 residues within the NUDIX domain that are conserved across various species. So, even when considering the presence of a specific domain with a folding favorable to an interaction with a small molecule, only a small number of residues might be responsible for the actual interaction.

Among the smaller set of proteins identified that might potentially bind NAD, TRPC3 (UniProt ID J3QTB0) had the ankyrin repeat domain and had the highest number of NAD-interacting residues. The corresponding reviewed protein (UniProt ID: Q13507) of TRPC3 is longer than the two isoforms detected here, with 836 amino acids. Its known interactions were found to be mostly involved in signal transduction, response to stress, anatomical structure development, and transport processes, many of them related to calcium transport and signaling, such as the inositol trisphosphate (IP3) receptors ITPR1 and ITPR3, and the Sodium/calcium exchanger 1 SLC8A1.

TRPC3 is a member of the transient receptor potential (TRP) channels family, which regulates calcium concentration [25]. The canonical subfamily of the TRP channels is directly activated by lipids, specifically diacylglycerol (DAG). Together with IP3, DAG is a product of the hydrolysis of a phospholipid catalyzed by the phospholipase C (PLC) enzymes, that are key components of intracellular calcium signaling, in response to the activation of different receptors by neurotransmitters, hormones and growth factors. Some PLCG1 functions have been associated to a specific protein domain that directly interacts with TRPC3 and PLCG1, regulating calcium entry [26]. Very recently, the role of PLC gamma enzymes in disease development has been explored [27]. Of note, PLCG1 was also found in our dataset of NAD-PPIs, showing that it already binds other NADBPs.

5 Conclusion

With this study, we obtained the main pathways in which NADBPs and NAD-PPIs are mostly involved and identified putative NADBPs. Both NAD-dependent signaling and calcium-dependent signaling are essential in the cell and therefore their dysregulation is often associated with disease. In particular, the role of NAD as a regulator of

calcium channels has been recently reviewed, due to its impact on cancer treatment research [28], where calcium channels emerge as potential targets for anticancer therapy. In addition to cancer, the TRP channels, namely the TRPC3 group, regulate functions in neurons and are involved in various neurological and psychiatric disorders [29]. Generally, the proteins highlighted throughout this study were involved in several critical cellular pathways and processes that, when disrupted, may lead to pathological conditions.

References

1. Res I, Lichtarge O (2005) Character and evolution of protein-protein interfaces. *Phys Biol* 2(2):S36–43. <https://doi.org/10.1088/1478-3975/2/2/S04> (in English)
2. Chakrabarti P, Janin J (2002) Dissecting protein-protein recognition sites. *Proteins* 47(3):334–343. <https://doi.org/10.1002/prot.10085> (in English)
3. Wang Y, Zhang H, Zhong H, Xue Z (2021) Protein domain identification methods and online resources. *Comput Struct Biotechnol J* 19:1145–1153. <https://doi.org/10.1016/j.csbj.2021.01.041> (in English)
4. Haigis MC, Sinclair DA (2010) Mammalian sirtuins: biological insights and disease relevance. *Annu Rev Pathol* 5:253–295. <https://doi.org/10.1146/annurev.pathol.4.110807.092250> (in English)
5. Gupte R, Liu Z, Kraus WL (2017) PARPs and ADP-ribosylation: recent advances linking molecular functions to biological outcomes. *Genes Dev* 31(2):101–126. <https://doi.org/10.1101/gad.291518.116> (in English)
6. Chini EN (2009) CD38 as a regulator of cellular NAD: a novel potential pharmacological target for metabolic conditions. *Curr Pharm Des* 15(1):57–63. <https://doi.org/10.2174/138161209787185788> (in English)
7. Li J et al (2017) A conserved NAD⁺ binding pocket that regulates protein-protein interactions during aging. *Science* 355(6331):1312–1317. <https://doi.org/10.1126/science.aad8242> (in English)
8. Santos L et al (2019) A novel form of Deleted in breast cancer 1 (DBC1) lacking the N-terminal domain does not bind SIRT1 and is dynamically regulated in vivo. *Sci Rep* 9(1):14381. <https://doi.org/10.1038/s41598-019-50789-7> (in English)
9. Wishart DS et al (2018) HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* 46(D1):D608–D617. <https://doi.org/10.1093/nar/gkx1089> (in English)
10. Kuhn M et al (2014) STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Res* 42(Database issue):D401–7. <https://doi.org/10.1093/nar/gkt1207> (in English)
11. Wishart DS et al. (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 46(D1):D1074–D1082. <https://doi.org/10.1093/nar/gkx1037> (in English)
12. Gaulton A et al (2017) The ChEMBL database in 2017. *Nucleic Acids Res* 45(D1):D945–D954. <https://doi.org/10.1093/nar/gkw1074> (in English)
13. Kim S et al (2016) PubChem substance and compound databases. *Nucleic Acids Res* 44(D1):D1202–D1213. <https://doi.org/10.1093/nar/gkv951> (in English)
14. Berman HM et al (2000) The protein data bank. *Nucleic Acids Res* 28(1):235–242. <https://doi.org/10.1093/nar/28.1.235> (in English)
15. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 45(D1):D158–D169. <https://doi.org/10.1093/nar/gkw1099> (in English)
16. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 34(Database issue):D535–9. <https://doi.org/10.1093/nar/gkj109> (in English)

17. Szklarczyk D et al (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 47(D1):D607–D613. <https://doi.org/10.1093/nar/gky1131> (in English)
18. Orchard S et al (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods* 9(4):345–350. <https://doi.org/10.1038/nmeth.1931> (in English)
19. Mi H et al (2019) Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat Protoc* 14(3):703–721. <https://doi.org/10.1038/s41596-019-0128-8> (in English)
20. Finn RD et al (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 44(D1):D279–D285. <https://doi.org/10.1093/nar/gkv1344> (in English)
21. Finn RD et al (2015) HMMER web server: 2015 update. *Nucleic Acids Res* 43(W1):W30–W38. <https://doi.org/10.1093/nar/gkv397> (in English)
22. Ansari HR, Raghava GP (2010) Identification of NAD interacting residues in proteins. *BMC Bioinform* 11:160. <https://doi.org/10.1186/1471-2105-11-160> (in English)
23. Medvedev KE, Kinch LN, Schaeffer RD, Grishin NV (2019) Functional analysis of Rossmann-like domains reveals convergent evolution of topology and reaction pathways. *PLoS Comput Biol* 15(12):e1007569. <https://doi.org/10.1371/journal.pcbi.1007569> (in English)
24. Li J, Mahajan A, Tsai MD (2006) Ankyrin repeat: a unique motif mediating protein-protein interactions. *Biochemistry* 45(51):15168–15178. <https://doi.org/10.1021/bi062188q> (in English)
25. Samanta A, Hughes TET, Moiseenkova-Bell VY (2018) Transient receptor potential (TRP) channels. *Subcell Biochem* 87:141–165. https://doi.org/10.1007/978-981-10-7757-9_6 (in English)
26. Wen W, Yan J, Zhang M (2006) Structural characterization of the split pleckstrin homology domain in phospholipase C-gamma1 and its interaction with TRPC3. *J Biol Chem* 281(17):12060–12068. <https://doi.org/10.1074/jbc.M600336200> (in English)
27. Liu Y et al (2020) Structural insights and activating mutations in diverse pathologies define mechanisms of deregulation for phospholipase C gamma enzymes. *EBioMedicine* 51:102607. <https://doi.org/10.1016/j.ebiom.2019.102607> (in English)
28. Yu P, Cai X, Liang Y, Wang M, Yang W (2020) Roles of NAD. *Molecules* 25(20). <https://doi.org/10.3390/molecules25204826> (in English)
29. Huang Q, Wang X, Lin X, Zhang J, You X, Shao A (2020) The role of transient receptor potential channels in blood-brain barrier dysfunction after ischemic stroke. *Biomed Pharmacother* 131:110647. <https://doi.org/10.1016/j.biopha.2020.110647> (in English)

Multiple Instance Learning Based on Mol2vec Molecular Substructure Embeddings for Discovery of NDM-1 Inhibitors



Thomas Papastergiou , Jérôme Azé , Sandra Bringay ,
Maxime Louet , Pascal Poncelet , and Laurent Gavara 

Abstract In this paper, we first present a new dataset of NDM-1 biological activities that is compiled by a cleaned version of the NMDI database. A literature review enriched the former database by 741 new compounds, comprising activities against NDM-1 classified in three classes (inactive, weakly and strongly active compounds) by specifying a unifying procedure for the labeling, which covers a range of different activity properties. Second, we restate the classification problem in the Multiple Instance Learning (MIL) setting by representing the compounds as a collection of Mol2vec vectors, each of them corresponding to a specific substructure (either atom or atom including their first neighbors). We observe an amelioration up to 45.7% and 38.47% in respect to balanced accuracy and F1-score, respectively, for the strongly active class in the MIL approach when compared to the classical Machine Learning paradigm. Finally, we present a classification and ranking framework based on classifiers learned by a k-fold CV procedure, which possess different hyper-parameters per fold, learnt by a Bayes optimization procedure. We observe that the top-3 and

T. Papastergiou (✉) · J. Azé · S. Bringay · P. Poncelet
LIRMM, University of Montpellier, CNRS, Montpellier, France
e-mail: thomas.papastergiou@lirmm.fr

J. Azé
e-mail: jerome.aze@lirmm.fr

S. Bringay
e-mail: sandra.bringay@lirmm.fr

P. Poncelet
e-mail: pascal.poncelet@lirmm.fr

S. Bringay
AMIS, Paul Valéry University, 34000 Montpellier, France

T. Papastergiou · M. Louet · L. Gavara
IBMM, CNRS, University of Montpellier, ENSCM, Montpellier, France
e-mail: maxime.louet@umontpellier.fr

L. Gavara
e-mail: laurent.gavara@umontpellier.fr

top-5 ranked accuracies of the strongly active classified compounds yield 100% for the MIL setting.

Keywords Machine learning · Multiple instance learning · Drug discovery · NDM-1 inhibitors

1 Introduction

New Delhi Metallo- β -lactamase (NDM-1) is a recent bacterial enzyme highly involved in bacterial resistance phenomenon by its capacity to inactivate the main available class of antibiotics: the β -lactam agents [1]. The common way to fight this kind of resistance is the adjuvant strategy, which consists in a combination of a β -lactam agent and a β -lactamase inhibitor [2]. Some combinations are already on the market but sadly are not effective on NDM-1 producers. Due to the specific mode of action of NDM-1, involving zinc atoms into the active site, the design of efficient inhibitors remains an unmet therapeutic need [3]. This major threat on human health has to be addressed to avoid return to the pre-antibiotic era.

The drug discovery process is a very time-consuming (approximately 10–14 years) and costly (1 billion USD magnitude) procedure, characterized by high attrition rates, to reach marketing authorization [4]. Thus, *in silico* strategies, (e.g. Virtual Screening (VS) techniques) are often used as starting point for medicinal chemistry, for speeding-up the drug discovery process by identifying compounds of high potential against specific targets. VS can be categorized in three main areas: (1) structure-based (requiring knowledge of the 3D structure of the target), (2) ligand-based (requiring knowledge of active ligands) and (3) hybrid approaches [5]. As the number of ligands in openly available databases is constantly increasing (e.g. ZINC 15 [6], ChEMBL [7] etc.) Machine Learning (ML) techniques are used for constructing efficient models used in VS for hit identification (i.e. discovery of small molecules as a starting point for medicinal chemistry programs), drug repurposing, activity scoring [8] or activity prediction [9]. In order to tackle the latter problem in an efficient manner specialized, annotated data are needed, since ligand-activity data that refer to different targets or to general target categories (e.g. antibacterial, anti-cancer, anti-inflammatory etc.) will produce ML models with low efficiency on specified tasks (e.g. discovery of effective NDM-1 inhibitors).

Multiple Instance Learning (MIL) is a paradigm of weakly supervised learning where the samples to be classified (i.e. bags) are represented by multiple vectors (i.e. instances) and labels are only available for the bags. MIL was first introduced by Dietterich et al. in 1997 [10] tackling a musk odor prediction task. In this structure-activity prediction problem, each molecule was represented by their different conformations captured by various feature vectors representing the shape of the molecule in each conformation. The standard MIL assumption was then applied stating that a bag is positive if it contains at least one positive instance (i.e. an active molecule conformation) and negative otherwise. The MIL paradigm has been used in different

application areas including medical imaging classification, frailty prediction using physiological signals [11], natural images classification [12, 13], drug discovery [14] etc.

As the numerical representation of molecules is crucial in order to construct ML models, different approaches have been proposed including Extended-Connectivity Fingerprints (a.k.a. Morgan Fingerprints (MF)), Molecular Graphs or computer learned representations [15] like the Mol2vec representation [16], a NLP-inspired technique that considers compound substructures, extracted by MF, as words and compounds as sentences. In this frame, a compound is represented by a collection of vectors, each of which corresponds to a substructure of the molecule, and a vector representation is obtained by adding-up these substructure vectors.

ML have been extensively used in the drug design process for various purposes: prediction on drug-protein interactions, discovering of drug efficacy or ensuring the safety biomarkers, with applications ranging from prediction of protein folding or target identification to hit discovery [8]. More specifically, Shi et al. [17] compiled a NDM-1 activities database, comprising strongly and weakly active compounds of known NDM-1 activities and provided a list of “hypothetical” inactive compounds, based on their physicochemical properties. They have applied classical ML and deep learning models for activity prediction based on physicochemical features extracted by the commercial software MOE2018.¹

In this paper we present a framework to tackle the problem of discovering potential strongly active NDM-1 inhibitors by the use of ML models. For this purpose, (1) we compiled a database of 868 compounds of known activity against NDM-1, by collecting compounds from the recent literature and by considering only compounds referring to the NDM-1 enzyme, coming from the NDMI database, proposed by Shi et al. [17]; (2) we established a unifying set of rules for labelling compounds as inactive, weakly active or strongly active, by considering different experimental properties; (3) we restated the activity classification problem as a MIL problem by representing molecules by a collection of Mol2vec vectors representing molecular substructures; (4) we proposed an ensemble classification framework, which is able to rank the classification outputs per predicted class.

The contributions of this paper can be resumed as follows:

1. The compilation of a dataset of known activities against NDM-1 annotated by a set of unifying rules for incorporating different experimental properties;
2. The restatement of the activity classification problem in the MIL paradigm, by representing compounds by Mol2vec representations of their substructures, that shows experimentally better performance than state-of-the-art Mol2vec classical ML models;
3. The introduction of an homogeneous ensemble classifier framework that classifies and ranks the classification results per class, and shows very promising classification and ranking results for the strongly active class in terms of top-5 to top-15 accuracy, when evaluated on an independent test set showing good generalization capabilities for the MIL ensemble models.

¹ <https://www.chemcomp.com/Products.htm>.

2 Materials and Methods

2.1 Dataset Collection

In [17], Shi et al. introduced a database of active and “hypothetical” inactive compounds, found in the literature, comprising 511 and 6,358 compounds respectively. The “hypothetical” inactive compounds were specified by considering physicochemical properties of 51,280 compounds of the ZINC database, lacking of activity data against NDM-1. To compile a database comprising only compounds with known activities against NDM-1, we considered only the 511 compounds of NDMI. For each of these compounds, we tried to verify the existence of the publications by performing database searches on the PubMed² database using the provided Digital Identification Number (DOI) of each publication. In a subsequent step, the relevance to NDM-1 inhibitors activities of the publications were checked, and irrelevant entries were discarded. Subsequently, the corresponding Canonical SMILES representation was produced, using the RDKit³ library, and duplicate entries were discarded. This procedure yielded 127 compounds with known activity scores. Furthermore, a thorough search in the existing literature for compounds with known activities on NDM-1 returned 741 new unique compounds. In total the new NDM-1 activity database comprises 868 unique compounds.

2.2 Labeling the Database

The activity against NDM-1 is measured by experimental properties based on enzymatic inhibition: (K_i , IC_{50} , pIC_{50} , enzyme inhibition at a set concentration, or K_d) [18] or in vitro bacterial growth inhibition (MIC) [19]. Our goal is to identify potential strongly active compounds against NDM-1. We classified the compounds in the new database in three classes: inactive, weakly active and strongly active compounds, inspired by the classification in [17] but with different, stricter, cut-off values for the strongly active compounds, since the aim is to deliver a classifier that can predict strongly active molecules with high enzymatic inhibition potency. We adopt a unifying strategy that comprises all activity properties. We include, in contrast to [17], only compounds with known activities against NDM-1 and classify them according to the cut-off values shown in Table 1.

As the compounds found in the literature often possess activity measurements for multiple properties and as different papers report different values for the same compound which sometimes leads to different labeling of the same compound, we need a unifying approach to cure these inconsistencies. We adopted a ranking order for the properties and classified each compound according to the property with the

² <https://pubmed.ncbi.nlm.nih.gov/>.

³ <https://github.com/rdkit/rdkit>.

Table 1 Labeling cut-off scores for activity properties

Rank		Inactive	Weakly active	Strongly active
1	K_i (μM)	>10	[0.5, 10)	≤ 0.5
2	IC_{50} (μM)	>20	(1, 20]	≤ 1
3	p IC_{50}	<4.7	[4.7, 6]	≥ 6
4	%100 μM	<60%	>60%	–
5	K_d (μM)	>10	[0.5, 10)	≤ 0.5
6	MIC ($\mu\text{g/ml}$)	>8	(0.5, 8]	≤ 0.5

highest rank. The ranking of the properties is shown in Table 1. Furthermore, if the classification of a compound according to two different publications is ambiguous, respecting the ranking of the properties, the more active label is assigned to the compound, since there is evidence in at least one experiment of the highest activity. We need to note here that when we applied the above procedure to the 127 compounds retained from the NMDI database [17], labels of 51 compounds (40.16%) changed.

The rationale behind the ranking of the activity properties is following the main objective of this work, which is to deliver a classification model for the discovery of active NDM-1 inhibitors. In this sense, properties which refer to enzymatic inhibition (e.g. K_i , IC_{50}) are placed in higher ranks than activity properties that refer to the NDM-1 agent inhibition (e.g. MIC). In this sense, for compounds that both enzymatic and bacterial inhibition activity are provided, we rely on the enzymatic activity property for their classification. On the other hand, when only the agent’s inhibition property is provided, we rely on properties like MIC, although that in vivo experiments tend to possess a higher degree of complexity, than enzymatic assays, and MIC values are indirect observations. Indeed, it’s the concentration of β -lactam agents to have antibacterial effect protected by a fixed concentration of a NDM-1 inhibitor. In this sense, the adopted ranking procedure resolves these ambiguities, in the aforementioned direction, and has a mild effect on the labeling of the dataset, since if the ranking of the activity properties is e.g. reversed only about 3% of the compounds would change labels.

2.3 Calculating Mol2vec Embeddings

ML Embeddings

For calculating the embeddings, we used the Mol2vec pre-trained model of [16]. The model was trained on 19.9 M compounds of ZINC and ChEMBL databases, as a skip-gram word2vec model, with window size of 10 using radius 1 for the MF (for a more elaborate description of the extraction of the MF refer to [20]). For training the Mol2vec model all MF identifiers of radii 0 and 1 were generated, and considered as words, while each molecule was considered as sentence. The rare identifiers (i.e. identifiers that occurred less than 3 times in the training database) were marked as

“unknown” and were attributed to a special identifier called ‘UNK’. After training the word2vec model, with such a specification, the individual vectors of each molecule, that corresponded to the MF substructures, were added up to produce a single vector for each molecule, and thus a molecule is represented by a vector of 300 real values.

MIL Embeddings

In order to restate the classification problem in the MIL setting, each molecule (i.e. bag) has to be represented by a collection of the individual’s MF substructures vectors (i.e. instances). The labels for each bag are known as the activity of each corresponding molecule is known, but the individual labels for each instance are unknown, since there is no activity information concerning each substructure. Thus, for the molecule to be bound to the target, one or multiple substructures of the molecule must be involved (i.e. active) in the binding affinity.

As the Mol2vec model calculates the embedding vectors of all the substructures of each molecule, up to a specified radius r , , after removing all duplicate vectors corresponding to the same substructure, we introduce two different types of MIL representations: (1) each molecule can be represented as a collection of all the substructure vectors of all radii (used in this work), or alternatively (2) each molecule can be represented as a collection of substructure vectors corresponding to a specific radius k , with $0 \leq k \leq r$. . In contrast to the Mol2vec model, where all the substructure vectors (i.e. vectors corresponding to MF of different radii) are added up to construct a vector representation for each compound, in the MIL representation, each unique substructure vector is explicitly included in the compound’s representation. As we will show experimentally, this contributes positively to the performance of the models. Indeed, according to the MIL assumption, the inactivity of a molecule suggests that all his substructures must be inactive (i.e. not contributing to the binding affinity). Weak or strong activity suggests that a portion of its substructures is involved to the binding affinity.

2.4 Classification and Ranking Frame Work

In this section, we introduce a homogeneous classification and ranking framework, which is based on different models acquired by a k -fold Cross Validation (CV) procedure. Let $f_i^{h_i} : \mathbb{R}^m \rightarrow \{cl_1, \dots, cl_n\}$, and $d_i^{h_i} : \mathbb{R}^m \rightarrow \mathbb{R}$, $i = 1, \dots, k$, k classification functions and their corresponding decision functions obtained by a k -fold CV procedure, where n and m are the number of classes and features respectively and $h_i \in \mathbb{R}^l$ are the corresponding hyper-parameters specified by a hyper-parameter optimization procedure for each individual fold. In this sense, we are equipped with k homogeneous classifiers trained and evaluated in different training-validation sets having different hyper-parameters. The decision of the ensemble classifier is then given by a voting procedure $g(f_0^{h_0}, \dots, f_k^{h_k}) = c$ and the per class rank of the ensemble’s classification output for each sample can be given by

$r_c(x) = \text{mean}_i \left\{ d_i^{h_i}(x), \text{iff}_i^{h_i}(x) == c \right\}$, $c = cl_1, \dots, cl_n$. Thus, by calculating for each sample the mean decision value of these classifiers, which have predicted the decision of the ensemble classifier, we obtain the rank per class of each sample.

3 Results and Discussion

For the evaluation of the proposed methods, we used the NDM-1 activities database described in Sect. 2.1. The 868 known activities compounds’ database, included 345 (39.75%) inactive, 254 (29.26%) weakly active and 269 (30.99%) strongly active molecules, making it a relative balanced dataset.

For representing numerically the compounds of the database for the classical ML paradigm, we generated Mol2vec vectors, employing the 300 dimensional pre-trained model of [16] resulting to 863 unique identifiers and 21 “unknown” structures. After generating the numerical representation for the MIL algorithms, we were equipped by 19,082 instances of radii 0 and 1, from which 1 radius 0 and 55 radius 1 structures were “unknown”. Furthermore, we obtained 7,264 and 11,818 instances of radii 0 and 1 respectively. The “unknown” structures were removed from the training and test sets, since they do not contribute to the representation of a bag, because they represent potential different substructures. The removal of the “unknown” structures does not result to bags (i.e. molecules) without representation, since each compound was represented by at least one known substructure.

The performance evaluation of the ranking and ensemble classification framework was performed by an independent Test Set (TS), acquired by a stratified (90% Training (TrS)-10% (TS)) split of the database, while the evaluation of the classifiers was performed by tenfold CV on the TrS split. Support Vector Machines (SVM) with Radial Basis Kernel (RBF), Linear Discriminant Analysis (LDA) and Random Forest (RF) [8] have been used as representatives of classical ML algorithms and TensMIL [11] and TensMIL2 [12] as MIL state-of-the-art algorithms, from which we decoupled the feature extraction by tensor decomposition phase, since our data are of 2D nature, and used only the classification procedure.

TensMIL and TensMIL2 consist of two inference phases: in the first phase, a score for each instance (i.e. substructure) is calculated and the bags’ scores distributions are estimated. These distributions are then fed to a bag classifier that yields the classification result. The difference of TensMIL2 is that, in the first phase, incorporates an instance selection procedure for choosing the most informative instances (i.e. substructures) per bag.

For tuning the hyper-parameters for each algorithm, a Bayes optimization approach was adopted like in [11], using as objective function the mean twofold CV balanced accuracy (Bacc) on a validation set. The hyper-parameters were tuned separately for each one of the 10-folds, resulting thus to 10 different classifiers with different sets of hyper-parameters. The hyper-parameters tuned for each classifier were: C and γ for SVM, $nrOfForestTrees$ for RF, ϑ_H and ϑ_p for TensMIL and q and

p for TensMIL2, where ϑ_H corresponds to the number of the histogram bins for the distribution estimation, ϑ_p and p to the variance retained of the PCA applied to the instances' feature matrix and q to the quantile defining the threshold for the instance selection procedure of TensMIL2. For the required ϑ_H parameters of TensMIL2 we used for each experiment the $\lceil \text{mean } \theta_H^i \rceil$, $i = 1, \dots, 10$, where ϑ_H^i is the parameter acquired by TensMIL on the i -th fold of the corresponding experiment. For the LDA algorithm none hyper-parameter was tuned. For discussion on the hyper-parameters, the interested reader may refer to the corresponding publications.

For the ensemble classifier, we used a majority voting approach in the sense that the class predicted by the majority of the classifiers is attributed to the corresponding sample.

The metrics used for evaluating the ML, MIL and ensemble classifiers were the mean of tenfold CV accuracy, balanced accuracy, precision-, recall- and F1-score-per class. For the evaluation of the ranking procedure, we used the per class top- k accuracy: $TopAcc_c^{(k)} = \frac{\#top-k \text{ ranked True Positives}}{k}$, with c being the corresponding class.

3.1 Results

Classification and Generalization Evaluation

Since we are interested in discovering strongly active NDM-1 inhibitors, special attention on the presentation of the results will be given to the strongly active class. In Table 2 we compare the classification performance of the ML and MIL paradigms. The reported Precision (Prec.) and Recall metrics refer to the corresponding metrics of the strongly active class.

As presented in Table 2, the MIL approach resulted in an amelioration from 38.43% up to 45.7% in terms of balanced accuracy (Bacc.) in respect to the ML approach. Precision (Prec.) and Recall for the strong activity class was augmented up to 40.07% and 29.30% respectively in the case of the MIL setting in comparison to the ML paradigm. The improvement of the classification performance could

Table 2 Comparison between classical ML and MIL for NDM-1 activity classification

Tenfold CV	Ensemble classifier			
	Acc. (%)	Bacc. (%)	Prec. ^a (%)	Recall ^a (%)
SVM	52.25	50.83	64.38	66.40
LDA	51.09	50.24	58.95	68.00
RF	52.76	51.15	62.41	62.28
TensMIL	72.08	70.81	80.65	80.53
TensMIL2	74.40	73.20	82.57	80.52

^a Refers to the strong activity class metric

Table 3 Per class CV and ensemble classifier F1-scores

	Tenfold CV F1-score			Ensemble classifier F1-score		
	Inactive class	Weakly active class	Strong active class	Inactive class	Weakly active class	Strong active class
SVM	0.5946	0.2021	0.6107	0.5546	0.0714	Inf
LDA	0.5794	0.2243	0.6082	0.1053	0.1875	0.5
RF	0.6216	0.2142	0.5861	0.5667	0.0741	Inf
TensMIL	0.7767	0.5357	0.8023	0.8462	0.5263	0.7931
TensMIL2	0.7942	0.5866	0.8116	0.8354	0.5116	0.7692

be attributed to the compounds' MIL representation. Instead of representing each compound by the sum of the vectors corresponding to each substructure, as is the case of the ML paradigm, each molecule is represented by the set of vectors of their substructures. As, in the frame of MIL, the individual activity labels of each instance (i.e. substructure) are unknown, and as the binding of a ligand to a target is a subject of specific substructures of a compound (i.e. the binding site of the ligand may concern a part of the compound having special structures and binding properties) the MIL representation has been proven beneficial to the activity classification performance.

Furthermore, for evaluating the generalization ability of the models as well as for assessing the ranking performance of the ensemble classification framework introduced in Sect. 2.4, we evaluated their performance in an independent test set that was not subject of the training, hyper-parameter tuning and CV evaluation of the models. As presented in Table 2, the ensemble classifier, in the frame of classical ML models, performs worse than the individual classifiers, suggesting that the generalization ability of these classifiers are poor. In contrast, in the MIL setting we see that the ensemble classification framework, in the case of TensMIL, performs better than the individual classifiers, in terms of Acc., Bacc. and Recall for the strongly active class, and in the case of TensMIL2 it performs slightly worse than the initial classifiers, suggesting the generalization ability of the initial classifiers. The ensemble classifiers in the MIL setting performed, in terms of balanced accuracy, from 86.29% to 123.32% better than in the classical ML setting (Fig. 1). In the case of the LDA model, the ensemble classifier displays a 96% recall, but only 34% precision for the strong activity class, suggesting that, in this case, a significant amount of compounds are predicted as strongly active and thus the False Positive predictions are relatively high.

For further assessing the performance of the classifiers and their generalization ability, the tenfold CV and on the independent TS F1-scores of the classifier and the ensemble classification framework are presented in Table 3. Overall, the MIL classifiers performed better in comparison to the classical ML classifiers. More specifically, in the MIL setting, we had from 24.95% to 37.07%, from 138.84% to 190.24% and from 31.36% to 38.47% better F1-scores for the inactive, weakly active and strongly active classes respectively, for the tenfold CV evaluation. For the MIL algorithms, the F1-score performance is better or slightly worse for the ensemble classifier on

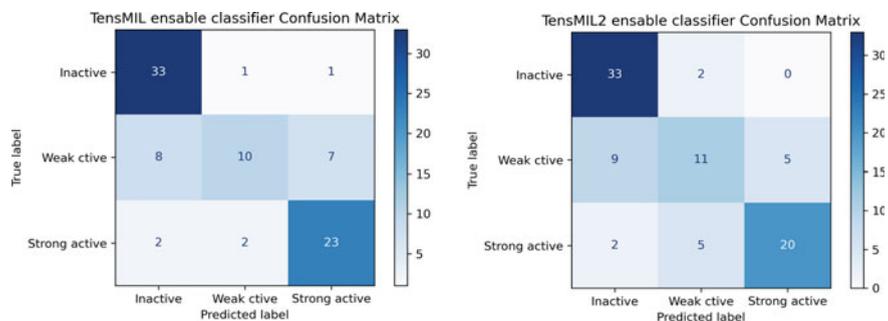


Fig. 1 Confusion matrices of TensMIL and TensMIL2, for the ensemble classifier on the independent test set

the independent TS, in contrast to the ML algorithms. Furthermore, we observed that the ensemble classifier based on the SVM and RF algorithms was not able to predict samples of the strong active class. In general, we observed lower performances in respect to the F1-score for the weakly activity class, than for the inactive and strong active classes.

Finally, comparing the results in [17], where handcrafted features and “hypothetical” inactive compounds were used, to our experiments, we conclude that in general, the classification performance, with respect to the F1 score in [17], is better for the inactive and weakly active class. In contrast, TensMIL2 performs from 15.36% to 41.66% better than the models in [17] for the strongly active class. Although, the two experiments are not fully comparable, we can conclude that the use of Mol2vec representations in the MIL setting and the stricter labeling for the strongly active class had a positive effect in the performance of the classification of the strongly active class.

Ranking Evaluation

The results of the ranking procedure are displayed in Table 4 where the top-3, 5, 10 and 15 ranked compounds accuracy per class are presented.

The improvement of the MIL algorithms in comparison to the classical ML algorithms in terms of the top-k ranking accuracy for the inactive class is from 1.33x to 7x (top-15 accuracy), for the weakly active compounds up to 10x and for the active

Table 4 Ranking performance (top-k accuracy) of the ensemble classifiers per class

	Inactive class				Weak active class				Strong active class			
	Top-3	Top-5	Top-10	Top-15	Top-3	Top-5	Top-10	Top-15	Top-3	Top-5	Top-10	Top-15
SVM	0.333	0.6	0.6	0.6	0.333	0.2	0.1	0.0667	0	0	0	0
LDA	0.667	0.4	0.2	0.133	0.667	0.6	0.3	0.2	0.333	0.6	0.7	0.6
RF	0.333	0.4	0.5	0.467	0.333	0.2	0.1	0.067	0	0	0	0
TensMIL	1	0.8	0.9	0.867	0.667	0.6	0.8	0.667	1	1	0.9	0.933
TensMIL2	1	0.8	0.9	0.933	0.667	0.8	0.6	0.6	1	1	0.9	0.933

class up to 3x. TensMIL and TensMIL2 are displaying 100% top-3 and top-5 accuracy, meaning that the top-5 ranked compounds are strongly active. In contrast, the ranking based on the ensembles of RF and SVM algorithms did not yield strongly active compounds in the top-15 ranks. In the evaluation of the classification performance of the ensemble classifier, MIL algorithms display better ranking accuracy than ML algorithms, furthermore, their performances on inactive and strongly active class are better than on the weakly active class.

4 Conclusion

To conclude, the compilation of a new database comprising compounds with known activities against NDM-1 (excluding “hypothetical” inactive compounds), as well as the unifying labeling procedure, that comprises a stricter, in comparison to former approaches, rules for the strongly active compounds, can be beneficial for discovering strongly active compounds against NDM-1. Furthermore, the restatement of the classification problem in the MIL framework, by representing a compound as a bag of vectors corresponding to their substructures, showed promising results, in terms of the efficiency in the three-class classification problem. Indeed, a part of the molecule corresponding to certain substructures, is responsible for the binding of the ligand to the target. The introduction of the homogeneous ensemble classifier and the ranking procedure, especially if MIL algorithms are used, showed promising results, as in the case of TensMIL and TensMIL2 classifiers ensembles, where the top-3 and top-5 ranked strongly active predicted compounds belong to the strongly active class, as predicted on an independent test set. This fact suggests that a screening for active compounds could reveal strongly active compounds among the top ranked results of the ensemble classifier. Finally, the classification evaluation of the ensemble classifier on an independent test set showed a great generalization ability for the MIL classifiers.

Acknowledgements This project was publicly funded through ANR (the French National Research Agency) under the “Investissements d’avenir” programme with the reference ANR-16-IDEX-0006.

References

1. Mojica MF, Bonomo RA, Fast W (2016) B1-Metallo- β -Lactamases: where do we stand? *Curr Drug Targets* 17(9):1029–1050
2. González-Bello C (2017) Antibiotic adjuvants—a strategy to unlock bacterial resistance to antibiotics. *Bioorg Med Chem Lett* 27(18):4221–4228
3. Linciano P et al (2019) Ten Years with New Delhi Metallo- β -lactamase-1 (NDM-1): from structural insights to inhibitor design. *ACS Infect Dis* 5(1):9–34
4. DiMasi JA, Grabowski HG, Hansen RW (2016) Innovation in the pharmaceutical industry: new estimates of R&D costs. *J Health Econ* 47:20–33

5. Khan AU (2015) Virtual screening strategies: a state of art to combat with multiple drug resistance strains. *MOJ Proteomics Bioinform* 2(2):61–66
6. Sterling T, Irwin JJ (2015) ZINC 15–ligand discovery for everyone. *J Chem Inf Model* 55(11):2324–2337
7. Gaulton A et al (2017) The ChEMBL database in 2017. *Nucleic Acids Res* 45(D1):D945–D954
8. Dara S et al (2022) Machine learning in drug discovery: a review. *Artif Intell Rev* 55(3):1947–1999
9. Chan HCS et al (2019) Advancing drug discovery via artificial intelligence. *Trends Pharmacol Sci* 40(8):592–604
10. Dieterich TG, Lathrop RH, Lozano-Pérez T (1997) Solving the multiple instance problem with axis-parallel rectangles. *Artif Intell* 89(1):31–71
11. Papastergiou T, Zacharaki EI, Megalooikonomou V (2018) Tensor decomposition for multiple-instance classification of high-order medical data. *Complexity* 2018:1–13
12. Papastergiou T, Zacharaki EI, Megalooikonomou V (2019) TensMIL2: improved multiple instance classification through tensor decomposition and instance selection. In: 2019 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, September 2019, pp 1–5
13. Branikas E et al (2019) Instance selection techniques for multiple instance classification. In: 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), PATRAS, Greece, July 2019, pp 1–7
14. Carbonneau M-A et al (2018) Multiple instance learning: a survey of problem characteristics and applications. *Pattern Recognit* 77:329–353
15. Wigh DS, Goodman JM, Lapkin AA (2022) A review of molecular representation in the age of machine learning. *WIREs Comput Mol Sci* e1603
16. Jaeger S, Fulle S, Turk S (2018) Mol2vec: unsupervised machine learning approach with chemical intuition. *J Chem Inf Model* 58(1):27–35
17. Shi C et al (2020) Applications of machine-learning methods for the discovery of NDM-1 inhibitors. *Chem Biol Drug Des* 96(5):1232–1243
18. Burlingham BT, Widlanski TS (2003) An intuitive look at the relationship of K_i and IC_{50} : a more general use for the Dixon plot. *J Chem Educ* 80(2):214
19. Andrews JM (2001) Determination of minimum inhibitory concentrations. *J Antimicrob Chemother* 48(suppl_1):5–16
20. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50(5):742–754

Towards Improving Bio-Image Segmentation Quality Through Ensemble Post-processing of Deep Learning and Classical 3D Segmentation Pipelines



Anuradha Kar 

Abstract In biological image analysis, 3D instance segmentation is a crucial step towards extracting information on objects of interest from microscopy datasets. Existing instance segmentation pipelines are frequently affected by errors such as missing boundary layer cells or poorly segmented regions. In this study, we propose several ensembles as post-processing methods for improving the quality of outputs obtained from deep learning and classical 3D segmentation pipelines. These methods take as input the results from two independent 3D segmentation pipelines and combine them using different fusion algorithms. The first algorithm uses label set intersection, the second one involves adjacency graph composition and the third one works through segmented object boundary fusion followed by 3D watershed. These 3 algorithms are tested on a dataset of 3D confocal microscopy images of floral tissues. The third fusion algorithm is found to perform best and has better global and local accuracies compared to its input segmentations. The specialty of the proposed ensemble methods is that these are model agnostic, i.e., they can be used to combine segmentation results from deep learning as well as non-deep learning or classical pipelines. These methods could be highly beneficial in correcting segmentation errors arising from missing cells in the boundary layer or under segmentation in the inner tissue layers and ultimately provide us robust segmentation results in presence of variable image qualities in biological datasets.

Keywords Segmentation · Deep learning · Bio-imaging · Microscopy

1 Introduction

Accurate segmentation of 3D microscopy images is an essential first step in many biological analysis procedures like estimating cell lineages, studying cell morphology, growth and gene expression patterns [1–3]. In the past few years, a number of 3D segmentation algorithms have been developed which use watershed

A. Kar (✉)

Institut du Cerveau – Paris Brain Institute, 47 Bd de l’Hôpital, 75013 Paris, France
e-mail: anuradha.kar@icm-institute.org

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
F. Fdez-Riverola et al. (eds.), *Practical Applications of Computational Biology and Bioinformatics, 16th International Conference (PACBB 2022)*, Lecture Notes in Networks and Systems 553, https://doi.org/10.1007/978-3-031-17024-9_7

[4, 5], graph partitioning [6] or active contour models [7]. Recently a number of deep learning based 3D segmentation pipelines such as [1, 8, 9] have been developed which achieve highly accurate instance segmentation of 3D microscopy images to extract 3D objects from them on cellular levels. Both deep learning and non-deep learning pipelines for 3D segmentation are affected by errors arising from image quality variations or inherent pipeline characteristics. For example, watershed-based methods need extensive parameter tuning when the image intensity levels drop in the inner tissue layers [10]. For deep learning-based pipelines, segmentation quality may degrade when subjected to images that have artifacts or are different from images in their training datasets. In general, segmentation pipelines are frequently affected by errors in which objects of interest are either missed or segmented regions have incorrectly identified boundaries.

In order to mitigate these common segmentation errors, several ensemble methods are proposed and tested in this study. These methods operate by taking segmentation results from any two independent segmentation pipelines (we call them component segmentations here) as inputs and applying fusion strategies to produce a resultant segmented output. Three segmentation fusion algorithms are explored. The first one is based on label set based intersection of two 3D segmentations. The second algorithm creates region adjacency graphs [11] from the component segmentations and merges them using graph composition. The third algorithm extracts and adds object boundaries from the component segmentations. A 3D watershed is applied to the fused boundary images to produce the final 3D instance segmentations. All the three algorithms are tested on a dataset of confocal microscopy images of floral meristem where the two component segmentations for each 3D image are from a deep learning [1] and a classical segmentation pipeline [12]. The outputs from the fusion algorithms are evaluated using a 3D Jaccard index metric by comparing them with ground truth segmented images. From the results, it is observed that the third ensemble method is successful in mitigating errors present in the component segmentations such as missed boundary cells and under-segmented regions. Using this method, an overall improvement in segmentation quality in terms of volumetric accuracy is observed for a test dataset of 20 confocal microscopy stacks.

2 Previous Research on Ensemble Methods for Segmentation

Ensemble methods have been reported in several works for image segmentation such as [13–15]. In [13] a multi-model ensemble framework is presented which integrates multiple state of the art deep learning architectures into a more powerful ensemble model which outperforms the single models in terms of accuracy. In [14] an ensemble segmentation algorithm is generated using AdaBoost for liver lesion extraction problems, where the component algorithms are built to work on lesions of different sizes. In [15] an ensemble system for combining multiple deep learning

architectures in a layer wise manner is proposed. The prediction by the first layer is used as the augmented data of the training image for the next layer of the ensemble and the predictions of the second layer is then combined by using a weights-based scheme. Ensemble methods combining two deep learning models Mask R-CNN and DeeplabV3+ are presented in [16] which achieve higher Sensitivity and specificity in segmenting skin cancer lesion boundaries and outperformed several standalone models like U-Net, and SegNet. An ensemble of convolutional neural networks are presented in [17] for semantic segmentation and tested on polyp and skin segmentation datasets. DeepLabV3+ and variants of the Resnet models and diversity in the ensemble is achieved through the use of different loss functions. In [18] first a deep network is used for generating feature maps from input images that are easy to segment. The second network in the ensemble is used for segmenting the generated images. The segmented images generated by the first and second networks with weight averaging. For segmenting an Invasive coronary angiography (ICA) dataset, [19] presents methods for combining results from multiple deep learning models that are trained using different loss functions. The results were evaluated and weighted based on the segmentation accuracy of the models. Other ensemble algorithms for segmentation are discussed in [20–22].

Key differences between above works and the ensemble methods presented in the study are that: 1) The approaches in the current study are model agnostic, i.e., independent of the segmentation mechanism used in the individual pipelines, which allows combining of results from deep learning as well as non-deep learning pipelines irrespective of their underlying concepts. 2) The methods presented in this study are post processing algorithms, therefore they do not require modification, retraining or retuning of the original segmentation model architectures, making the methods computationally simple and easily implementable. 3) Finally, most ensemble methods are designed for either classification or semantic segmentation tasks, while the methods in this study are designed for 3D instance segmentation problem for microscopy datasets, which is uncommon in contemporary literature.

3 Methodology

3.1 3D Segmentation Pipelines

In this study, two 3D segmentation pipelines are used for applying the ensemble methods. The first pipeline is deep learning based and the second one is a classical 3D segmentation method. For the deep learning pipeline, training images are 3D confocal stacks acquired from growing multicellular tissues from the plant *Arabidopsis* and are provided in the open dataset of [23]. The first pipeline (called P1 in this work) is adapted from [1]. It is a deep learning based pipeline consisting of a residual 3D UNet model. The architecture and workflow of this pipeline is shown in Fig. 1.

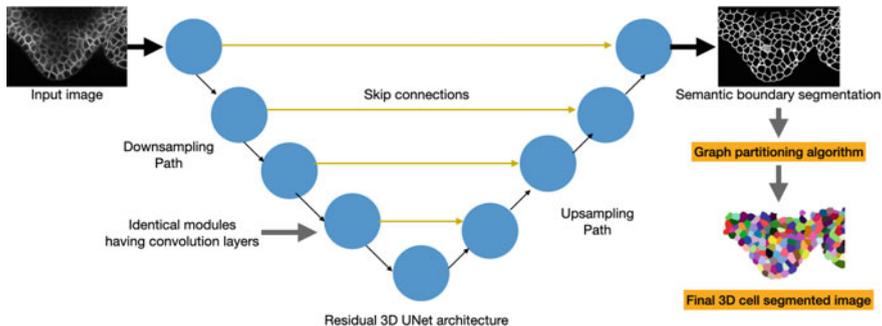


Fig. 1 Workflow of pipeline P1. The deep learning 3D UNet model is trained using confocal images to produce 3D cell boundary images which are post processed using graph partitioning techniques to obtain the final 3D instance segmented outputs

The second pipeline P2 is a classical 3D watershed technique [12]. In this, seed regions in a confocal image are defined using a h-minima parameter value and are then grown using morphological watershed transformation to identify segmented regions.

3.2 Dataset

For implementing the ensemble methods, a test dataset of twenty 3D stacks of plant floral multicellular tissues from *Arabidopsis thaliana* are used which are segmented independently by the two pipelines P1 and P2 described above. Each image has their expert 3D ground truth segmentations (3D stacks where voxels in each segmented cell has a unique label). The training data for P1 consists of 80 stacks of raw confocal images and corresponding 3D segmentations as ground truth (See Sect. 6 for training and test datasets used in this work).

3.3 Fusion Algorithm 1: Label Set Intersection

In this technique, fusing two component segmentations from pipelines P1 and P2 is based on a set intersection method (or simple sum of the label arrays). The objects in the component segmentations are treated as labeled voxels. Thus, if S_1 and S_2 are two voxel label arrays corresponding to the two component segmentations from pipelines P1 and P2 respectively, the output set of voxel labels is a label array j given by:

$$j = (\max(S_2) + 1) * S_1 + S_2 \quad (1)$$

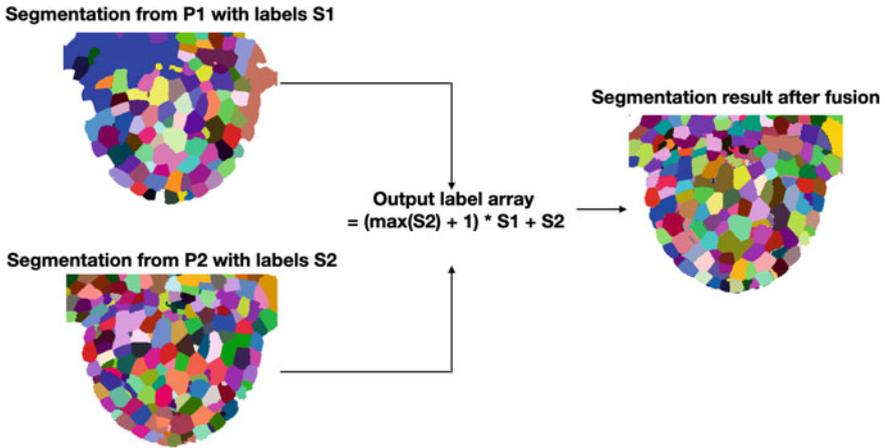


Fig. 2 Combining segmented images based on intersection of voxel label sets from each component segmentation S_1 and S_2

where j is the array of labels in the output segmentation. This forms our baseline or simplest method for combining the component segmentations.

The logic of this algorithm can be stated as: a pair of voxels from segmentations S_1 and S_2 are determined to fall in the same segment of the output segmentation if and only if they are in the same segment in both S_1 and S_2 (Fig. 2). Thus, A voxel is said to belong to a region in the output segmentation only if it belongs to the corresponding regions in both of the input segmentations.

3.4 Fusion Algorithm 2: Region Adjacency Graphs

The Region Adjacency Graphs (RAG) are used to model regions within an image as nodes of a graph which represent the neighboring relationships between pixels. A Region is defined as a collection of connected pixels sharing common properties, e.g., color of the pixels. Thus, a segmentation of an image can be associated with a RAG. For a graph with vertices V and edges E such that graph $G' = \{V, E\}$ and where a node represents a pixel (for 2D), a partition into R connected regions is done such that:

$$V_1 \cup V_2 \cdots \cup V_R = V \quad (2)$$

$$V_1 \cap V_2 \cdots \cap V_R = \emptyset \quad (3)$$

These regions may be identified with a new graph defined as $G' = \{V', E'\}$ where each partition of V is identified with a node of V' . The new edge set E' is defined by the criteria that edge weight between each node in V' is equal to the sum of edge weights e connecting each original node in the set, so that for $e_{ij} \in E'$:

$$w_{ij} = \sum_{e_{ks}, v_k \in V_i, v_s \in V_j} w_{ks} \quad (4)$$

In this work, RAGs R1 and R2 (undirected graphs) are created from two component segmentations (S1 and S2 respectively) obtained from the two pipelines P1 and P2 as shown in Fig. 3. The voxel labels in S1 and S2 are converted to nodes of RAGs R1 and R2. The criteria for creating RAGs from the segmentations is the color (or voxel label) of each segmented region. Then R1 and R2 are combined using graph composition which is the union of node and edge sets from the two component RAGs R1 and R2. The union of two graphs $R = R_1 \cup R_2$ with set of vertices V_1 and V_2 and set of edges E_1, E_2 has a set of vertices and edges given by:

$$V_{res} = V_1 \cup V_2 \quad (5)$$

$$E_{res} = E_1 \cup E_2 \quad (6)$$

Finally, the nodes of this composed resultant RAG RAG_res or R are converted to voxel labels to get back the resultant 3D fused segmentation (Fig. 3).

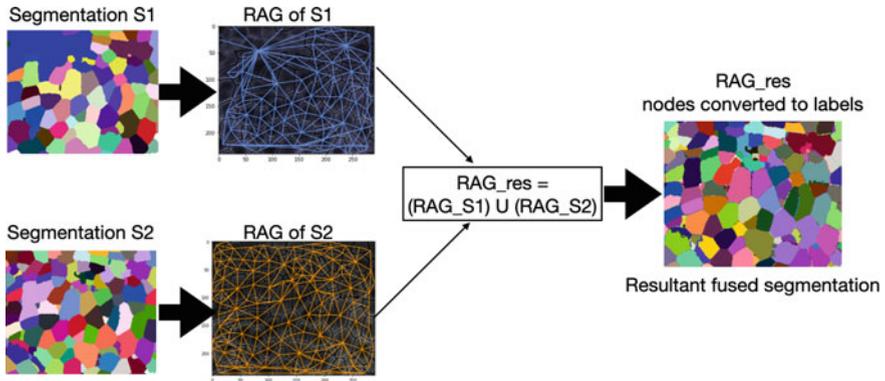


Fig. 3 Concept of converting segmented images to RAGs and composing component RAGs to get fused segmentation. Example shown in 2D above works similarly for 3D data

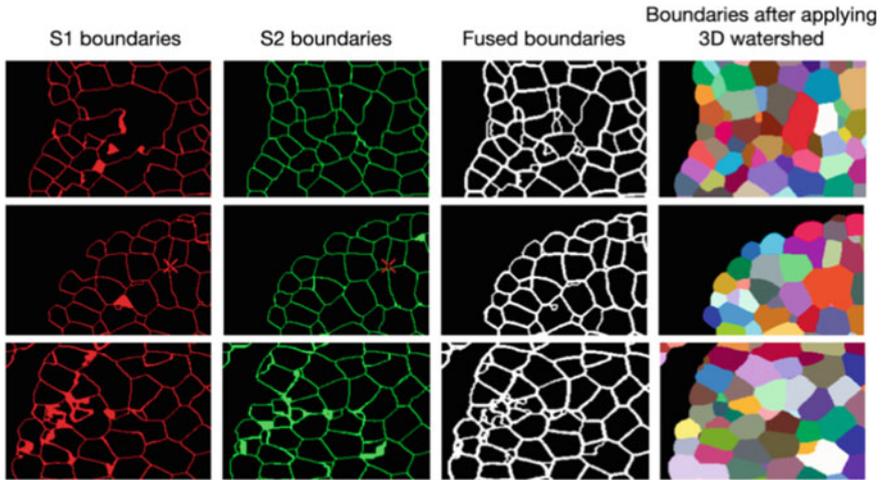


Fig. 4 Boundary fusion strategy followed by applying 3D watershed

3.5 Fusion Algorithm 3: Boundary Fusion with 3D Watershed

This ensemble method is a two-step approach. First, from each component segmentation, the object boundaries are extracted followed by pixel-by-pixel addition of the boundaries from each component segmentation. Object boundaries are obtained as a 3D binary image where the boundary of each object has a value of 1 and the cell interiors and background have a value of 0 (Fig. 4). The binary 3D boundary images from the two component segmentations are added to produce a fused boundary image. To this image, 3D watershed with the h-minima parameter = 0 (since cell interiors are zero) is applied to produce the final 3D instance segmentation.

This is a new ensemble method that can help to counter significant segmentation errors occurring in both deep learning and non-deep learning based segmentation pipelines. For example, from Fig. 4 it is seen that this method helps to mitigate two key problems in the segmentation results from P1 and P2- which are under-segmentation (first row) and missing boundary cells (2nd and 3rd row).

3.6 Segmentation Evaluation Metric

For investigating the quality of segmentations obtained from each ensemble method, an averaged Jaccard Index metric is used. In this, the volumetric Jaccard Index (JI) between a ground truth object (from the ground truth segmented image for the

corresponding image) and a predicted object is first estimated. The JI is then weighted with the volume of the ground truth object and summed for all ground truth object labels (k) and finally averaged using the sum of volume of all ground truth objects.

$$\text{Average JI} = \frac{\sum_i (G_i * \text{Jaccard Index}_i)}{\sum_i G_i} \quad (7)$$

where $i = 1, 2 \dots k$

This metric takes on values ranging between 0 and 1. Better segmentation accuracy is indicated by a metric result closer to 1. With this, the Jaccard Index score for individual segmented objects can be estimated without losing object location information. Implementation of this metric is provided in the open repository with this paper (Sect. 6).

4 Results

4.1 Results from Original Pipelines

The original segmentation accuracy results from the two pipelines P1 and P2 are shown in Fig. 5A. For this, the averaged JI is computed for the segmentation results of P1 and P2 on the 20 test image stacks, using respective ground truth stacks. From the violin plots, it is seen that pipeline P1 (deep learning based) achieves a mean average Jaccard index of 0.7 while the 3D watershed-based pipeline P2 gets 0.8 on the test dataset.

4.2 Results from Fusion Algorithm 1

The average JI values are computed for the results of fusion algorithm 1 and shown in Fig. 5B. The x-axis shows the names of the 20 test stacks and the y-axis presents the segmentation accuracy values for P1, P2 and their fusion result for those stacks. For most of the test stacks, the average JI of fusion results are close to the best results from the two component pipelines. Fusion algorithm 1 helps to recover many missing cells in the segmentation but it does not perform very well to remove large under-segmented regions in the middle of the tissue. Also, the resultant cell boundaries in the fused image are not very smooth. This algorithm however works in real time and has a simple implementation.

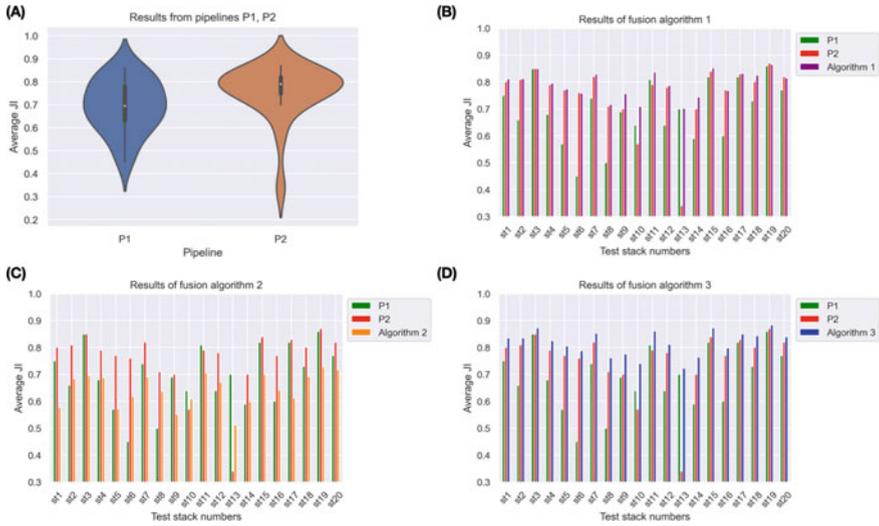


Fig. 5 Segmentation accuracy plots (y-axis view limited between 0.3 and 1) for the two segmentation pipelines and the three different ensemble methods (using fusion algorithms 1, 2, 3)

4.3 Results from Fusion Algorithm 2

The average JI values for fusion algorithm 2 are plotted in Fig. 5C. The JI values of the RAG based ensemble result fall between those from P1, P2 or in some cases lower than both of the P1 and P2’s results. The outputs from this algorithm mitigates the missing cells problem but not under-segmentations. Also, the cell boundaries remain irregular. Another issue is the high processing time for RAG creation especially for large stacks with more than 100 objects. Overall, this algorithm achieves the goal of combining two segmentations but does not help to mitigate all types of segmentation errors. It also produces global segmentation accuracy levels that are often lower than those of the two component segmentations and processing times are in the order of 20–60 min for segmented stacks having more than 100 objects.

4.4 Results from Fusion Algorithm 3

The segmentation accuracies of the results from fusion algorithm 3 are plotted along with those of P1, P2 in Fig. 5D. It is seen that for all the test image stacks, this ensemble method achieves a higher segmentation accuracy compared to those obtained from its component segmentations on the same stacks. It works in real time and is independent of the number of objects in the component segmentations. In Fig. 6, the results from

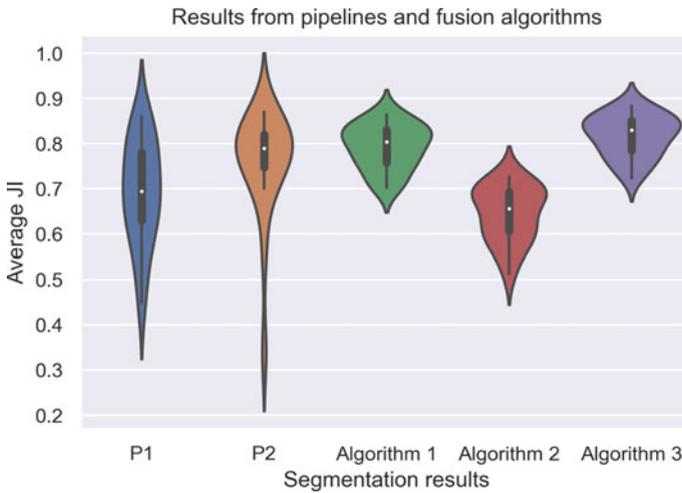


Fig. 6 Statistical plots of segmentation accuracies of individual pipelines P1, P2 along with accuracy results from fusion algorithms 1, 2 and 3

the original pipelines are included and denoted as P1 (deep learning based) and P2 (3D Watershed) for results from Pipelines 1 and 2 respectively.

With algorithm 3, the accuracy is better than those from P1 and P2. Algorithm 1 achieves a mean accuracy close to the better performing pipeline out of P1, P2. Algorithm 2 achieves a mean accuracy that is closer to the pipeline with lower accuracy out of the two. The fusion algorithms have different impacts but the third is most efficient.

5 Discussions

Quality of segmentation of 3D microscopy images is critical for biological analysis procedures that rely on the segmented image contents. Poor segmentation quality with missing cells or erroneous cell boundaries may lead to errors in successive biological analysis steps. Although several 3D instance segmentation pipelines for microscopy images exist, they face limitations in performance when image quality is degraded, which often leads to rendering large volumes of bio-image datasets unusable. In this study, several post-processing techniques are developed for segmentation quality improvement based on ensembles of segmentation results. These post-processing techniques are model agnostic, that is they can be used irrespective of the types of the component 3D segmentation pipelines. The benefits of these methods are that these are simple to implement and do not require retraining of the deep learning pipelines or re-tuning of the classical ones but still achieves improved segmentation by combining results from each pipeline. The ensemble post-processing concepts

presented here could be highly useful to mitigate segmentation errors like missing cells and under-segmentation by utilizing the efficiencies of individual pipelines. This could also make the segmentation results robust to image quality and segmentation pipeline performance variations. The future work in this direction is to test fusion strategies with more than two segmentation pipelines and extend the analysis on other biological datasets.

6 Data and Code Availability

All data and codes used in this work are available as open resources. The training data for the 3D UNet model may be found at: <https://www.repository.cam.ac.uk/handle/1810/262530>. The test dataset containing 3D confocal stacks used in this work are obtained from: <https://www.repository.cam.ac.uk/handle/1810/318119>.

Codes used for different parts of this work may be found in the GitHub repository at: <https://github.com/anuradhakar49/SegFusion>.

References

1. Wolny A, et al (2020) Accurate and versatile 3D segmentation of plant tissues at cellular resolution. *eLife* **9**. <https://doi.org/10.7554/eLife.57613>
2. Yang L, Zhang Y, Guldner IH, Zhang S, Chen DZ (2016) 3d segmentation of glial cells using fully convolutional networks and k-terminal cut. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W (eds) MICCAI 2016. Springer, Cham pp 658–666 (2016). https://doi.org/10.1007/978-3-319-46723-8_76
3. Jiang J, Kao P-Y, Belteton SA, Szymanski DB, Manjunath BS (2019) Accurate 3D cell segmentation using deep features and CRF refinement. In: 2019 IEEE ICIP, pp 1555–1559
4. Kornilov A, Safonov I (2018) An overview of watershed algorithm implementations in open source libraries. *J Imaging* **4**:123. <https://doi.org/10.3390/jimaging4100123>
5. Lou S, Pagani L, Zeng W, Jiang X, Scott PJ (2020) Watershed segmentation of topographical features on freeform surfaces and its application to additively manufactured surfaces. *Precis Eng* **63**:177–186. <https://doi.org/10.1016/j.precisioneng.2020.02.005>
6. Kappes JH, Speth M, Andres B, Reinelt G, Schn C (2011) Globally optimal image partitioning by multicuts. In: International workshop on energy minimization methods in computer vision and pattern recognition. Springer, Heidelberg, pp 31–44
7. Zheng Q, Dong E, Cao Z, Sun W, Li Z (2014) Active contour model driven by linear speed function for local segmentation with robust initialization and applications in MR brain images. *Signal Process* **97**:117–133. <https://doi.org/10.1016/j.sigpro.2013.10.008>
8. Eschweiler D, Spina TV, Choudhury RC, Meyerowitz E, Cunha A, Stegmaier J (2019) CNN-based preprocessing to optimize watershed-based cell segmentation in 3D confocal microscopy images. In 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019). IEEE, pp 223–227
9. Stringer C, Wang T, Michaelos M, Pachitariu M (2021) Cellpose: a generalist algorithm for cellular segmentation. *Nat Methods* **18**:100–106
10. Kiss A, Moreau T, Mirabet V, Calugaru CI, Boudaoud A, Das P (2017) Segmentation of 3D images of plant tissues at multiple scales using the level set method. *Plant Methods* **13**:114. <https://doi.org/10.1186/s13007-017-0264-5>

11. Tremeau A, Colantoni P (2000) Regions adjacency graph applied to color image segmentation. *IEEE Trans Image Process* 9:735–744. <https://doi.org/10.1109/83.841950>
12. Fernandez R, Das P, Mirabet V, Moscardi E, Traas J, Verdeil J-L, Malandain G, Godin C (2010) Imaging plant growth in 4D: robust tissue reconstruction and lineaging at cell resolution. *Nat Methods* 7:547–553. <https://doi.org/10.1038/nmeth.1472>
13. Chen Y-H, Kuo P-H, Fang Y-Z, Wang W-L (2021) More birds in the hand-medical image segmentation using a multi-model ensemble framework. *Nat Mach Intell* 1:23–25
14. Shimizu A, Narihira T, Furukawa D, Kobatake H, Nawano S, Shinozaki K (2008) Ensemble segmentation using AdaBoost with application to liver lesion extraction from a CT volume. *MIDAS J*. <https://doi.org/10.54294/wrtw01>
15. Dang T, Nguyen TT, Moreno-Garcia CF, Elyan E, McCall J (2021) Weighted ensemble of deep learning models based on comprehensive learning particle swarm optimization for medical image segmentation. In: 2021 IEEE CEC. IEEE, pp 744–751
16. Goyal M, Oakley A, Bansal P, Dancey D, Yap MH (2020) Skin lesion segmentation in dermoscopic images with ensemble deep learning methods. *IEEE Access*. 8:4171–4181
17. Nanni L, Cuza D, Lumini A, Loreggia A, Brahnam S (2021) Deep ensembles in bioimage segmentation. *CoRR*. abs/2112.12955
18. Kato S, Hotta K (2021) Automatic preprocessing and ensemble learning for low quality cell image segmentation. *ArXiv*. abs/2108.13118
19. Park J, Kweon J, Bark H, Kim YI, Back I, Chae J, Roh J-H, Kang D-Y, Lee PH, Ahn J-M, Kang S-J, Park D-W, Lee S-W, Lee CW, Park S-W, Park S-J, Kim Y-H (2021) Selective ensemble methods for deep learning segmentation of major vessels in invasive coronary angiography. *medRxiv*
20. Kim H, Yoon H, Thakur N, Hwang G, Lee EJ, Kim C, Chong Y (2021) Deep learning-based histopathological segmentation for whole slide images of colorectal cancer in a compressed domain. *Sci Rep* 11:22520. <https://doi.org/10.1038/s41598-021-01905-z>
21. Zheng H, et al (2019) A new ensemble learning methods for 3D biomedical image segmentation. In: *AAAI*, vol 33, pp 5909–5916
22. Bousselham W, Thibault G, Pagano L, Machireddy A, Gray J, Chang YH, Song X (2021) Efficient self-ensemble for semantic segmentation. *arXiv*
23. Willis L et al (2016) Cell size and growth regulation in the *Arabidopsis thaliana* apical stem cell niche. *Proc Natl Acad Sci USA* 113:E8238–E8246

Exploring *Xylella fastidiosa*'s Metabolic Traits Using a GSM Model of the Phytopathogenic Bacterium



Alexandre Oliveira , Emanuel Cunha , Miguel Silva ,
Cristiana Faria , and Oscar Dias 

Abstract *Xylella fastidiosa* is a gram-negative phytopathogenic bacterium able to infect over 500 plant species, with devastating consequences for agricultural and forest-based economies. In the last decade, genome-scale metabolic (GSM) models have become important systems biology tools for studying the metabolic behaviour of different organisms. In this work, a GSM model of *X. fastidiosa* subsp. *pauca* De Donno is presented, comprising 1164 reactions, 1379 metabolites, and 508 genes. The model was validated by comparing *in silico* simulations with available experimental data. The GSM model allowed identifying potential drug targets using a pipeline based on a gene essentiality analysis of the model.

Keywords *Xylella fastidiosa* · GSM modelling · Essentially analysis · *Merlin*

1 Introduction

Xylella fastidiosa, first described in 1987 [36], is a gram-negative phytopathogenic bacterium transmitted among plants by xylem-fluid feeding insects. *X. fastidiosa* is able to infect over 500 plant species, as demonstrated in the *Xylella* spp. host plant database [14]. Also, many host plants, despite the infection, may remain symptomless and function as a reservoir for the phytopathogen [3]. In Europe, it led to a severe outbreak of olive quick decline syndrome in Italy, specifically in the Apulia region. This resulted in the death of thousands of trees, causing economic and environmental consequences. Hence, *X. fastidiosa* was declared as a quarantine pest by the European Commission [35].

A. Oliveira · E. Cunha · M. Silva · C. Faria · O. Dias (✉)
Centre of Biological Engineering, University of Minho, 4710-057 Braga, Portugal
e-mail: odias@ceb.uminho.pt

A. Oliveira · E. Cunha · M. Silva · O. Dias
LABBELS –Associate Laboratory, Braga, Guimarães, Portugal

C. Faria
SilicoLife Lda., Braga, Portugal

The field of systems biology aims at studying whole-cell mechanisms, using genomic and metabolic data [30]. As a result of developments in high-throughput sequencing techniques, which leads to an upsurge of the available genomic data, systems biology approaches have gained increased interest as means of obtaining metabolic insights. Genome-Scale Metabolic (GSM) Models comprise both genetic and metabolic data on a given organism and can be used extensively to predict the phenotype of the organism in different environmental conditions [10]. Additionally, GSM models have been applied in the discovery of potential drug targets for pathogenic organisms [6, 31].

Recently, a GSM model of *X. fastidiosa* subsp. multiplex CFBP 8418 was released to explore metabolic properties associated with the fastidious growth of the phytopathogenic organism [16]. This work aimed at reconstructing a GSM model of *X. fastidiosa pauca* De Donno, which is the causal agent of the olive quick decline syndrome outbreak in Italy [32]. Due to the current spread of *X. fastidiosa* and the lack of a cure that effectively stops the phytopathogen, the developed model was also used to study intrinsic metabolic traits of the organism and find new ways to fight its spread by investigating potential drug targets.

2 Materials and Methods

2.1 Software

The major steps of the GSM model reconstruction were performed within the *merlin* [11] framework. This software provides a friendly interface for manual curation and contains multiple tools that accelerate the reconstruction process. Furthermore, all simulations related to the final stage of reconstruction (model validation) were performed in COBRApy [13]. Lastly, the drug targeting pipeline performed in this work required both COBRApy and Biopython [8] packages. *X. fastidiosa* subsp. *pauca* De Donno RefSeq assembly genome files, accessible via the NCBI [33] assembly accession number ASM211787v1 [17], were automatically retrieved with *merlin*.

2.2 Metabolic Model Reconstruction

Genome Annotation. The genome annotation of the genome of the bacterium was based on similarity searches with the basic local alignment search tool (BLAST) [1]. First, the similarity search was run against UniProtKB/SwissProt [9], using an expected value (e-Value) of $1e-30$ as threshold. Genes without hits were submitted to an additional BLAST against UniProtKB/TrEMBL. The genome functional annotation was performed using the *merlin*'s "annotation workflow" tool, which annotates

all candidate genes (a gene that has at least one of its homologous is associated with a metabolic function) based on a set of defined organism.

Metabolic Network Assembly. An initial draft metabolic network for *X. fastidiosa* was assembled by coupling the information obtained in the genome annotation step with metabolic data retrieved from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [18]. The subcellular compartments of each protein was predicted using PSORTb 3.0 [37]. Transport reactions were generated using the Transport Systems Tracker (transyt.bio.di.uminho.pt).

Both the *Escherichia coli* GSM model (iOJ1366) [28] and *Xanthomonas campestris* large-scale metabolic model [34] were used as templates to infer the macromolecular composition of the biomass equation. The biomass reaction accounts for a total of seven macromolecular entities, namely Deoxyribonucleic acid, ribonucleic acid, Protein, Cofactor, Peptidoglycan, Lipopolysaccharide, and Lipid (Supplementary File S1). The detailed description of the steps involved in the reconstruction of the GSM model is available in Supplementary File S2.

Model Validation. The reconstructed GSM model's performance was evaluated by comparing the predicted results to available information from the literature. As the amount of information on *X. fastidiosa*'s metabolism is limited, *X. campestris*' data was used whenever necessary. In this work, all *in silico* simulations were considered one of the biomass reactions ('R_e-Biomass' or 'R_Coupled_Biomass') as the objective function, which was maximized through a Parsimonious Flux Balance Analysis (pFBA) [25], except otherwise indicated. Different simulations addressing the bacterium aerobic metabolism, amino acid auxotrophies, and glucose flux pattern were performed to validate the model (more details in Supplementary File S2 Sect. 1.4).

2.3 Identification of Potential Drug Targets

One of the most important applications of the *X. fastidiosa*'s reconstructed GSM model is the identification of potential drug targets that could impair the organism's survival. A medium based on the olive tree's xylem fluid composition [12] was replicated to reflect the natural environment of the phytopathogenic bacterium. Additionally, a chemically defined media, CHARD2 [22], which replicates the xylem environment was used in this step.

The model's essential genes were determined with adequate methods available in COBRApy, to pinpoint potential drug targets. The genes identified were filtered using a simple pipeline: a BLAST search (e-Value threshold of 0.0001) against the *Olea europaea* genome, followed by an evaluation of the results. However, before discarding a gene as a potential drug target, due to homologous genes present in the host's genome, a metabolic potential analysis of the host, using KEGG and MetaCyc pathways was performed to evaluate the possibility of alternative metabolic routes for the same objective. In other words, if a reaction encoded by an *in silico* essential gene was essential for *X. fastidiosa* but not for the host (*Olea europaea*),

the gene could be considered a potential drug target and further analysed. Finally, the DrugBank database was used to find potential inhibitors of the filtered genes. Hence, BLAST alignments and searches of EC number(s) associated with essential genes were performed on said database. Only compounds classified as inhibitors or antagonists were included. Using information retrieved from the DrugBank and BRENDA, the collected compounds were filtered to include only drugs with known activity in gram-negative bacteria.

Furthermore, to find other strategies to possibly kill the phytopathogenic organism, synthetic lethals, which are pairs of non-essential genes whose simultaneous knockout leads to model infeasibility towards biomass maximization, and triple gene knockout *in silico* simulations were performed.

3 Results and Discussion

3.1 Model Validation

The metabolic behavior of the GSM model was first assessed in aerobic conditions. As expected, an aerated environment supports the growth of the pathogen, while an oxygen restriction leads to infeasible solutions, as not even the energetic maintenance requirements can be fulfilled. Therefore, the *in vivo* aerobic metabolism can be properly simulated, and viable energy production is achieved. Furthermore, as expected, all oxidative phosphorylation enzymatic complexes included in the model, namely, NADH dehydrogenase, succinate dehydrogenase, ubiquinol oxidase and ATP synthase, display flux through their reactions. The following test relied on the validation of amino acid auxotrophies. To date, no auxotrophies have been reported for *X. fastidiosa*, as the phytopathogen seems to be able to grow using exclusively glutamine as a nitrogen source [23]. In fact, it is expected for *X. fastidiosa* to have biosynthetic capabilities for all amino acids as it thrives in poor and limited nutrient environments. Furthermore, single amino acid omissions, simulated with the model, comply with the reported data as no auxotrophies were detected. Results from *in silico* simulations are presented in Supplementary File S3.

Model Summary. The GSM model reconstructed in this work, comprises a total of 508 genes, 1379 metabolites and 1160 reactions. This model also includes 1014 gene-protein-reaction associations, 241 transport reactions, and 94 exchange reactions. The iMS508 model is available in BioModels [26] with the identifier MODEL2205020002. An overview of the available metabolic models for *X. fastidiosa*, *E. coli*, and *X. campestris* pv. *campestris* B100 is presented in Table 1.

The number of genes, reactions, and metabolites included in iMS508 are similar to the one presented by the *X. fastidiosa* multiplex model. However, these values are lower than the ones found for *E. coli*, which is related with the high amount of information available for this model organism. The *X. campestris* pv. *campestris* B100 model presents a smaller network, probably due to the lack of information for

Table 1 Comparison of the number of genes, metabolites and reactions between the GSM models of *X. fastidiosa pauca* De Donno (this work), *X. fastidiosa multiplex* CFBP 8414, *E. coli* K-12 substr. MG1655 (iJO1366), and *X. campestris* pv. *campestris* B100 (iSS352)

	<i>X. fastidiosa pauca</i> De Donno	<i>X. fastidiosa multiplex</i>	<i>E. coli</i> K-12	<i>X. campestris</i> pv. B100
Genes	508	572	1,367	352
Gene coverage (%)	24.6	26.6	31.6	7.8
Metabolites	1285	1107	1805	338
Reactions	1160	1158	2583	447

this species at the reconstruction moment (2013). The SBML file for the reconstructed model can be found in Supplementary File S4.

Glucose Flux Pattern. In several xanthomonads, glucose catabolism occurs mainly through the Entner-Doudoroff pathway (EDP) and to a lesser extent through the pentose phosphate pathway, ranging from 81–93% and 7–19%, respectively [38]. Assuming the maximization of biomass production as objective function, the model predicts 90% and 8% of the glucose flux towards the Entner-Doudoroff and pentose phosphate pathways, respectively.

X. fastidiosa seems to present a pyrophosphate-dependent phosphofructokinase (EC 2.7.1.90), similar to the one identified in *X. campestris* [15]. This enzyme could explain a flux bottleneck through the glycolytic pathway, which is a glucose degradation route energetically more efficient than the Entner-Doudoroff pathway, as the enzyme only shares a minimal portion of the total proteome of *X. campestris* [15]. Furthermore, phosphofructokinase deletion mutants of *X. campestris* are unaffected when compared with the wild type [34]. Previous studies indicated the impossibility of this organism to use gluconeogenesis, due to the lack of fructose-1-6-bisphosphatase (EC 3.1.3.11) [24]. However, the detected phosphofructokinase could close this gap found in the gluconeogenic pathway, as there is not a preferred direction of the reaction catalysed by the enzyme [15]. Moreover, according to *in silico* simulations, gluconeogenesis is mandatory when organic acids and amino acids, which are usually available in a xylem fluid environment [19], are used as carbon sources for *X. fastidiosa*'s growth.

Plant cells can produce reactive oxygen species (ROS) in response to environmental stress, including the invasion by pathogenic organisms [2]. ROS levels are substantially increased to act as defence mechanisms to fight off invasive pathogens [27]. Therefore, using the EDP as a catabolic pathway for carbohydrates could offer a metabolic advantage in comparison to glycolysis, as this pathway generates NADPH, which is especially meaningful in the detoxification of ROS. For instance, a study performed with *Pseudomonas putida*, known to catabolise carbohydrates through the EDP, shows that the organism is highly tolerant against oxidative stress, while a transgenic strain with activated glycolysis becomes sensitive to ROS [7]. Therefore,

Table 2 Potential drug targets evaluated for CHARD2 and Olive medium. The data was retrieved from the DrugBank database, considering only drugs with known inhibitory or antagonistic effects

Methods for integrating transcriptomics data into GEMs

Locus Tag	EC number(s)	Pathways	Drugs
B9J09_RS07720	2.5.1.7	Peptidoglycan	Fosfomycin
B9J09_RS10475	6.3.2.4	Peptidoglycan	Cycloserine
B9J09_RS10245	5.1.1.1	Amino acids	Cycloserine
B9J09_RS03350	1.5.1.3	Cofactors	Trimethoprim Methotrexate Trimetrexate Pemetrexed Pyrimethamine Pralatrexate Aminopterin
B9J09_RS04800	2.1.2.3; 3.5.4.10	Nucleotide	Methotrexate
B9J09_RS08730	1.17.4.1	Nucleotide	Gemcitabine
B9J09_RS08735			Hydroxyurea

X. fastidiosa may use the Entner-Doudoroff pathway as a source of redox potential as means to act against plant defence mechanisms.

3.2 Drug Targeting

The identification of drug targets is one of the main applications of GSM models. These models have been used to identify drug targets in gram-negative pathogens, such as *Klebsiella pneumoniae* [6], *Acinetobacter baumannii* [20], and *Pseudomonas aeruginosa* [4]. The first step to identify potential drug targets was to perform a gene essentiality analysis, using both the CHARD2 and the olive media (Supplemental File S5). To reduce the potential interactions between the drugs and the host, a BLAST search against the genome of *O. europaea* was performed for each essential gene. Moreover, a manual pathway analysis allowed to determine alternative routes in the olive tree's genome.

The DrugBank database was used to identify drugs affecting these genes. As expected, most identified drugs are antibiotics affecting the peptidoglycan and cell wall assembly (Supplemental File S6). Specifically, penicillin-binding proteins had a high number of identified potential drugs. Besides genes encoding such proteins, seven genes were also identified as potential drug targets (Table 2).

The genes B9J09_RS07720 and B9J09_RS10475 are also associated with the peptidoglycan biosynthesis, encoding however a UDP-N-acetylglucosamine 1-carboxyvinyltransferase (*murA*) and D-Alanine-D-alanine ligase, respectively. Fosfomycin and cycloserine were identified as inhibitors for these enzymes. Fos-

fomycin, a broad-spectrum antibiotic produced by *Streptomyces* species, has demonstrated inhibition against gram-negative bacteria, including *E. coli*, *A. baumannii*, and *P. aeruginosa*. This antibiotic acts as an irreversible inhibitor by binding covalently to a cysteine in the active site of the murA enzyme. Thus, the peptidoglycan precursor UDP-N-acetylmuramic acid is not produced, and the cell wall synthesis is disrupted. *Chlamydia trachomatis*, a gram-negative bacterium, has demonstrated resistance to Fosfomycin due to the exchange of cysteine by aspartate in the active site. The sequence of murA in the *X. fastidiosa* genome presents a cysteine in that position, indicating that Fosfomycin might have activity against this species.

Resistance against cycloserine was already reported in *X. fastidiosa* [5]. This antibiotic was also identified as a potential target for alanine racemase (encoded by B9J09_RS10245). Dihydrofolate reductase, encoded by B9J09_RS03350 in the *X. fastidiosa* genome, is an enzyme responsible for the synthesis of tetrahydrofolate. Seven drugs were identified as an inhibitor for this enzyme. From these, there are reports of resistance against trimethoprim [29]. Moreover, hydroxyurea has also been reported to inhibit ribonucleoside-diphosphate reductase in *E. coli* [21]. Penicillin-binding proteins are widely used as targets for antibiotics. These proteins play a major role in the cell wall assembly, essential for the virulence of pathogens. Using the searches by the EC numbers (3.4.16.4 and 2.4.1.129) and BLAST searches, 61 different antibiotics were identified as potential inhibitors of *X. fastidiosa* PBP.

The susceptibility of *X. fastidiosa* against some of them, such as cephaloridine, cephaloglycin, carbenicillin, and ampicillin was already screened [5]. Integration of data regarding drug resistance mechanisms into the model, such as antibiotics degrading enzymes and multidrug resistance efflux pumps, would allow improving the prediction capabilities of the model to specific drugs.

4 Conclusion

This work reports the reconstruction of iMS508, a GSM model for *X. fastidiosa subsp. pauca* De Donno. The reconstruction of the model was based on a semi-automatic genome annotation, and retrieval of information from biological databases and literature. The model can simulate aerobic growth of *X. fastidiosa* with accurate production rates of fastidian gum, and the lesA protein. The iMS508 was used for a brief drug targeting analysis, identifying 20 potential drug targets. Most of the drug targets identified are penicillin binding proteins, although genes related with the nucleotide, amino acids, and cofactors metabolism also presented results. Moreover, this approach identified drugs screened in previous studies that aimed at devising new compounds to fight this phytopathogenic bacterium.

5 Supplementary Materials

All Supplementary Material files mentioned in the manuscript are available at <https://nextcloud.bio.di.uminho.pt/s/NBEwEoLaCRY8DxF>.

Data Availability

The model presented in this work can be found in BioModels [26] with the identifier MODEL2205020002. To access the model:

- Visit <https://www.ebi.ac.uk/biomodels/login/auth>.
- Log in with the username reviewerForMODEL2205020002 and password DEMPB4.
- Access <https://www.ebi.ac.uk/biomodels/MODEL2205020002> to view the model.

Acknowledgements This study was supported by the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UIDB/04469/2020 unit. A. Oliveira (DFA/BD/10205/2020), E. Cunha (DFA/BD/8076/2020) hold a doctoral fellowship provided by the FCT. Oscar Dias acknowledge FCT for the Assistant Research contract obtained under CEEC Individual 2018.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Molec Biol* 215(3):403–410
2. Bailey-Serres J, Mittler R (2006) The roles of reactive oxygen species in plant cells. *Plant Physiol* 141:311
3. Baldi P, Porta NL (2017) *Xylella fastidiosa*: host range and advance in molecular identification techniques. *Front Plant Sci* 8:1–22
4. Bartell JA, Blazier AS, Yen P, Thøgersen JC, Jelsbak L, Goldberg JB, Papin JA (2017) Reconstruction of the metabolic network of *Pseudomonas aeruginosa* to interrogate virulence factor synthesis. *Nat Commun* 8(1):1–13
5. Blevé G, Gallo A, Altomare C, Vurro M, Maiorano G, Cardinali A, D’Antuono I, Marchi G, Mita G (2018) In vitro activity of antimicrobial compounds against *Xylella fastidiosa*, the causal agent of the olive quick decline syndrome in Apulia (Italy). *FEMS Microbiol Lett* 365(5):281
6. Cesur MF, Siraj B, Uddin R, Durmuş S, Çakır T (2020) Network-based metabolism-centered screening of potential drug targets in *Klebsiella pneumoniae* at genome scale. *Front Cell Infect Microbiol* 9:447
7. Chavarría M, Nikel PI, Pérez-pantoja D, Lorenzo VD, José S, Rica C (2013) The Entner - Doudoroff pathway empowers *Pseudomonas putida* KT2440 with a high tolerance to oxidative stress. *Environ Microbiol* 15(6):1772–1785
8. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11):1422–1423

9. Consortium TU (2020) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 49(D1):D480–D489
10. Dias O, Rocha I (2015) Systems biology in fungi. *Molec Biol Food Water Borne Mycotoxicogenic Mycotic Fungi*, 69–92 (2015)
11. Dias O, Rocha M, Ferreira EC, Rocha I (2015) Reconstructing genome-scale metabolic models with merlin. *Nucleic Acids Res* 43(8):3899–3910
12. Drossopoulos J, Niavis C (1988) Seasonal changes of the metabolites in the leaves, bark and xylem tissues of olive tree (*olea europaea*. L) i. nitrogenous compounds. *Ann Botany* 62(3):313–320
13. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR (2013) Cobrapy: constraints-based reconstruction and analysis for python. *BMC Syst Biol* 7(1):1–6
14. (EFSA), E.F.S.A (2018) Update of the *Xylella* spp . host plant database. Technical Report
15. Frese M, Schatschneider S, Voss J, Vorhölter FJ, Niehaus K (2014) Characterization of the pyrophosphate-dependent 6-phosphofructokinase from *Xanthomonas campestris* pv. *campestris*. *Arch Biochem Biophys* 546:53–63
16. Gerlin L, Cottret L, Cesbron S, Taghouti G, Jacques MA, Genin S, Baroukh C (2020) Genome-scale investigation of the metabolic determinants generating bacterial fastidious growth. *mSystems* 5(2):1–15
17. Giampetruzzi A, Saponari M, Almeida RPP, Essakhi S, Boscia D, Loconsole G, Saldarelli P (2017) Complete genome sequence of the olive-infecting strain *Xylella fastidiosa* subsp *pauca* De Donno. *Genome Announc* 5(27):5–6
18. Kanehisa M (2002) The KEGG database. In: Novartis foundation symposium, pp 91–100
19. Killiny N, Hijaz F (2015) Chemical composition of xylem sap of citrus sinensis *L. Osbeck* (sweet orange). *Proc Florida State Horticult Soc* 128(1):114–118
20. Kim HU, Kim TY, Lee SY (2010) Genome-scale metabolic network analysis and drug targeting of multi-drug resistant pathogen *Acinetobacter baumannii* AYE. *Molec BioSyst* 6(2):339–348 (2010)
21. Kjølner Larsen I, Sjöoberg BM, Thelander L (1982) Characterization of the active site of ribonucleotide reductase of *Escherichia coli*, bacteriophage T4 and mammalian cells by inhibition studies with *Hydroxyurea Analogues*. *Eur J Biochem* 125(1):75–81 (1982)
22. Leite B, Andersen PC, Ishida ML (2004) Colony aggregation and biofilm formation in xylem chemistry-based media for *Xylella fastidiosa*. *FEMS Microbiol Lett* 230(2):283–290
23. Lemos EGD, Alves LMC, Campanharo JC (2003) Genomics-based design of defined growth media for the plant pathogen *Xylella fastidiosa*. *FEMS Microbiol Lett* 219:39–45
24. Letisse F, Chevallereau P, Simon JL, Lindley ND (2001) Kinetic analysis of growth and xanthan gum production with *Xanthomonas campestris* on sucrose, using sequentially consumed nitrogen sources. *Appl Microbiol Biotechnol* 55:417–422
25. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, Adkins JN, Schramm G, Purvine SO, Lopez-Ferrer D, Weitz KK, Eils R, König R, Smith RD, Palsson BØ (2010) Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Molec Syst Biol* 6(1):390
26. Malik-Sheriff RS, Glont M, Nguyen TV, Tiwari K, Roberts MG, Xavier A, Vu MT, Men J, Maire M, Kananathan S et al (2020) Biomodels-15 years of sharing computational models in life science. *Nucleic Acids Res* 48(D1):D407–D415
27. O'Brien JA, Daudi A, Butt VS, Bolwell GP (2012) Reactive oxygen species and their role in plant defence and cell wall metabolism. *Planta* 236:765–779
28. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BØ (2011) A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism. *Molec Syst Biol* 7(535):1–9
29. Ribeiro MDP, Dellias MTF, Tsai SM, Bolmström A, Meinhardt LW, Bellato CM (2005) Utilization of the etest assay for comparative antibiotic susceptibility profiles of citrus variegated chlorosis and pierce's disease strains of *Xylella fastidiosa*. *Curr Microbiol* 51:262–266
30. Palsson BØ (2006) Systems biology - properties of reconstructed networks
31. Presta L, Bosi E, Mansouri L, Dijkshoorn L, Fani R, Fondi M (2017) Constraint-based modeling identifies new putative targets to fight colistin-resistant *A. baumannii* infections. *Sci Rep* 7(1):3706

32. Saponari M, Giampetruzzi A, Loconsole G, Boscia D, Saldarelli P (2018) *Xylella fastidiosa* in olive in Apulia: where we stand. In: Phytopathology, pp 1–43 (2018)
33. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrahi I, Ostell J, Panchenko A, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, John Wilbur W, Yaschenko E, Ye J (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 38(D1):5–16
34. Schatschneider S, Persicke M, Watt SA, Hublik G, Pühler A, Niehaus K, Vorhölter FJ (2013) Establishment, in silico analysis, and experimental verification of a large-scale metabolic network of the xanthan producing *Xanthomonas campestris* pv. *campestris* strain B100. *J Biotechnol* 167(2):123–134
35. Vos S, Camilleri M, Diakaki M, Lázaro E, Parnell S, Schrader G, Vicent A (2019) Pest survey card on *Xylella fastidiosa*. EFSA Supporting Publications
36. Wells JM, Raju BC, Hung HY, Weisburg WG, Mandelco-Paul L, Brenner DJ (1987) *Xylella fastidiosa* gen. nov. , sp. nov.: Gram-negative, xylem- limited, fastidious plant bacteria related to *Xanthomonas* spp. *Int J Syst Bacteriol* 37(2):136–143
37. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FSL (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26(13):1608–1615
38. Zagallo AC, Wang CH (1967) Comparative glucose catabolism of *Xanthomonas Species*. *J Bacteriol* 93(3):970–975

Genomic Regions with Atypical Concentration of Inverted Repeats



Carlos A. C. Bastos , Vera Afreixo , João M. O. S. Rodrigues ,
and Armando J. Pinho 

Abstract Hairpin/cruciform structures, as well as other non-B DNA structures, are important regulators for biological processes and gene function. The formation of these structures require that the DNA sequence contains adequately spaced inverted repeats. To study the potential of DNA regions to form hairpin/cruciform structures, we developed a new procedure to analyse the variation of the concentration of occurrence of inverted repeats at different spacings along the human genome. We apply the method to the human genome and identify regions with atypical high concentration of inverted repeats when compared to a control scenario based on a Markov model of order 7. We found that the potential to form hairpin/cruciform structures is very heterogeneous across different human genome regions. Also, different regions display strikingly different patterns of enrichment of concentration depending on inverted repeats spacing.

Keywords Cruciform · Distance distribution · Inverted repeats · Outliers · Markov model

C. A. C. Bastos (✉) · J. M. O. S. Rodrigues · A. J. Pinho
IEETA-Institute of Electronics and Informatics Engineering of Aveiro, Aveiro, Portugal
e-mail: cbastos@ua.pt
URL: <http://www.ua.pt>

J. M. O. S. Rodrigues
e-mail: jmr@ua.pt

A. J. Pinho
e-mail: ap@ua.pt

Department of Electronics, Telecommunications and Informatics, University of Aveiro,
3810-193 Aveiro, Portugal

V. Afreixo
CIDMA-Center for Research and Development in Mathematics and Applications, Aveiro, Portugal
e-mail: vera@ua.pt

Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal

1 Introduction

Hairpin/cruciform structures are a type of non-B DNA structure, sometimes called unusual or alternative DNA structures, not conforming to the canonical B structure, with importance in biological processes and gene function [8]. DNA motifs that are known to potentially form non-B DNA structures are available at public databases [6, 7]. Hairpins/cruciforms may form dynamically when certain conditions are met, such as the coiling state of DNA, but are less stable than the normal B-DNA conformation. Although their properties and relevance in several biological processes are acknowledged, evidence of their genomic location and mechanism of action are lacking *in vivo* [1, 9].

The stem and loop lengths of hairpin/cruciforms structures seem to vary over a wide range. According to different authors, the stem lengths vary between 6 and 100 nucleotides, while loop lengths may range from 0 to 2000 nucleotides [6, 11, 15]. Short distances could favour the occurrence of these structures, but long distances have also been reported, such as the translocation breakpoints associated with human developmental diseases or infertility [1].

The simultaneous occurrence of inverted repeats in a specific region are a required feature of local cruciform structures. However, some regions can greatly enhance the occurrence of hairpin/cruciforms conformations than others.

A DNA word analysis based on the distribution of the distances between adjacent symmetric words of length seven [13] showed a strong over-representation of distances up to 350, a feature that the authors considered might be associated with the potential for the occurrence of cruciform structures. The same research group later extended their analysis to include distance distributions of non-adjacent inverted repeats, since adjacency is not a required condition for cruciform structures to form [2, 4].

2 Methods

This work aims to find, in the human genome, structures with regularity beyond the already well-known repetition structures published in the literature. Thus, we used pre-masked sequences available from the UCSC Genome Browser webpage [10]. These files contain the GRCh38 assembly sequences, with repeats reported by RepeatMasker [12] and Tandem Repeats Finder [5] masked with N symbols.

Consider the alphabet $\mathcal{A} = \{A, C, G, T\}$ and let w be a symbolic sequence (word) defined in \mathcal{A}^k , where k is the length of w . The pair composed by one word, w , and the corresponding reversed complement word, w' , is called an inverted repeat pair. For example, (ACT, AGT) is an inverted repeat pair. Since DNA sequencing is not perfect, some symbols in real DNA sequences are not in \mathcal{A} . These unknown or ambiguous nucleotides are usually coded with N symbols or other IUPAC ambiguity

characters. We treat these ambiguous nucleotides as separators that split the sequence into a set of unambiguous subsequences.

In this work, we analyse, along the human genome, the *cumulative* distance distribution of all possible inverted repeats, by dividing the complete genome in successive windows containing 100000 nucleotides.

2.1 Distance Between Inverted Repeats

For all words of length k , we compute the frequency distributions of distances, f , between occurrences of each word and all succeeding reversed complements at distances between k and 4000.

For example, consider the sequence $ACTTTGTACTAAAGTTAAG$ of length 19. Only four inverted repeats (w , w') of length $k = 3$ occur in this short sequence. The following lines show all occurrences of these inverted repeats, marked by underlines (w) and overlines (w'):

(ACT, AGT) : $ACTTTGTACTAA\overline{AGTTAAG}$,
 (CTT, AAG) : $\overline{ACTTTGTACTAAAGTTAAG}$,
 (TTT, AAA) : $\overline{ACTTTGTACTAAAGTTAAG}$,
 (TAA, TTA) : $ACTTTGTACTAA\overline{AGTTAAG}$.

The previous sequence includes six distances to all the succeeding reversed complement words (distances: 12, 5, 10, 15, 8, and 5). Thus, the cumulative distribution is $f(5) = 2$, $f(8) = f(10) = f(12) = f(15) = 1$ and $f(d) = 0$ for all other d values.

For each word w we analyse distances up to 4000 nucleotides, but, if a N symbol is found, the search for w' is stopped, because the length of long stretches of N s may be artificial. Considering the stem length of possible cruciform structures we choose to study words of length $k = 7$.

2.2 Measuring the Concentration of Inverted Repeats

In order to evaluate how atypical a window is, the observed values of the $f(d)$ inverted repeat cumulative frequencies are compared to the expected values obtained from a Markov chain reference model of order 7. The method uses the total number of possible words at distance d to adjust the residual values to account for the actual number of ambiguous symbols in each window.

Expected Values Under Higher Order Markov Chain for DNA Sequences. Let $F(d)$ be the random variable that represents the total number of inverted repeats occurrence at distance d in a genomic region of length L and $n(d)$ the corresponding total number of possible word pairs at distance d , where $d = 1, 2, \dots, 4000$ and $d = k$ means that the two words (of length k) are in adjacent positions. Only distances between $d = 7$ and 4000 will be considered in this work, because to form hairpin/cruciform structures the words w and w' cannot overlap.

Let $p(d)$ be the probability of occurrence of inverted repeats at distance d . If we assume the independence between trials, $F(d)$ follows a binomial distribution, $F(d) \sim B(n(d), p(d))$. Asymptotically, $F(d)$ has normal distribution with mean $n(d)p(d)$ and variance $n(d)p(d)(1 - p(d))$, so we can define a z-score

$$Z(d) = \frac{F(d) - n(d)p(d)}{\sqrt{n(d)p(d)(1 - p(d))}} \overset{\cdot}{\sim} N(0, 1). \quad (1)$$

However, the occurrence of inverted repeats at distance d in a genomic sequence cannot be considered independent. For example, for $k = 7$ there is no realization where all the trials are inverted repeats.

Let M be the transition matrix of the Markov process, where each state corresponds to one genomic word of length k and the indexes $(1, \dots, 4^k)$ are in lexicographic order. M is a sparse matrix without absorbing states. The probability of one specific inverted repeat at distance d is given by

$$P(wx_1x_2\dots x_{d-k-1}w') = P(w)P(x_1x_2\dots x_{d-k-1}w'|w),$$

where $x_i \in \{A, C, G, T\}$. The word probability, $P(w)$, is estimated by the word frequency in the corresponding chromosome. The conditional probabilities, $P(x_1x_2\dots x_{d-k-1}w'|w)$, are obtained through the Markov transition matrix M^d , where $M^d = [p_{ij}(d)]$, with $1 \leq i, j \leq 4^k$, holds the probabilities of getting from initial state i and finishing in state j in d transitions.

Therefore, the probability of occurrence of inverted repeats at distance d is given by

$$p(d) = \sum_i P(i)p_{ii}(d).$$

Number of Possible Word Pairs. Given a sequence of L unambiguous symbols, the number of pairs of non-overlapping words of length k that can occur at a distance d is at most

$$n(d, k, L) = \begin{cases} L - k - d + 1, & \text{if } k \leq d \leq L - k \\ 0, & \text{otherwise.} \end{cases}$$

Sequences containing ambiguous symbols (represented by the N symbol) can be split into a set of unambiguous subsequences of lengths L_1, L_2, \dots , so the number of word pairs at a distance d is given by

$$n(d) = \sum_i n(d, k, L_i).$$

For example, with $k = 4$, the number of possible word pairs at distance $d = 6$ in a sequence of length 40 containing only unambiguous symbols is $n(6) = n(6, 4, 40) = 31$. However, if a length 40 sequence contains a single N symbol at position 11, the number of possible word pairs at distance $d = 6$ is only $n(6) = n(6, 4, 10) + n(6, 4, 29) = 1 + 20 = 21$.

Accordingly, the statistic defined in Eq. (1) is adjusted by the ratio $n(d)/n(d, k, l)$,

$$T(d) = \frac{n(d)}{n(d, k, L)} Z(d). \quad (2)$$

In order to measure the concentration of inverted repeats for a set of successive distances we compute the sum of all T values between two bounds (d_1 and d_2)

$$S_{[d_1, d_2]} = \sum_{d \in \{d_1, \dots, d_2\}} T(d). \quad (3)$$

Control Scenario. We use a simulation process to obtain a control scenario composed of a set of sequences with features similar to those of the masked human genome: 24 sequences with the same size of each of the human chromosomes and the same number and positions of the ambiguous symbols (N s). The control sequences were generated under a k -order Markov process using the statistics of each chromosome to estimate the probabilities of the words and the transition matrices.

We use the results of the simulation procedure to obtain a critical value for the S statistic (Eq. 3) under the assumption that the DNA sequence was generated by a k -order Markovian procedure. Thus, we could overcome the lack of independence between trials and to take into account the relative weight of unambiguous symbols of each window.

To compute the critical values, we use a one-way test, since our main purpose is the identification of genomic regions with atypical (by excess) concentration of inverted repeats. Assuming a significance level of 5%, we compute the 0.95 quantile (cv) of the S values of all windows in each simulated chromosome.

2.3 Windows Selection

In order to locate the sequence windows with highest concentration of inverted repeats, we obtain the sum of all T values $S_{[k,4000]}$. We use the Tukey method [14], and the critical values obtained from the control scenario to identify, in each chromosome, the windows with atypical concentration of inverted repeats. The Tukey's threshold used for outlier detection is $T_{thr} = Q_3 + 1.5|Q_3 - Q_1|$, with Q_1 and Q_3 the sample quartiles.

In order to identify regions (windows) with atypical inverted repeats concentrations for a range of distances, we subdivided the set of distances under analysis into 8 intervals of distances: 7–500; 501–1000; 1001–1500; 1501–2000; 2001–2500; 2501–3000; 3001–3500; 3501–4000. Using the results from all chromosomes, we used the $S_{[d_1, d_2]}$ statistic (Eq. 3) and the procedure described previously to identify genomic regions with atypical concentration of inverted repeats for each range of distances.

3 Results

Table 1 shows order statistics for each of the chromosomes, as well as, the critical values obtained in the control scenario, the Tukey thresholds and the percentage of windows where S surpasses those thresholds. As expected, the statistical behaviours of the concentration $S_{[7,4000]}$ of inverted repeats in the human genome and in the control scenario (Markov model) are significantly different. Almost all chromosomes have a majority of windows with a value of $S_{[7,4000]}$ above the critical values obtained from the simulation procedure.

The S values of the windows in the human chromosomes reveal a strong positive asymmetry, which is not present in the control scenario. This behaviour confirms the existence of some genomic regions with atypical (reinforced) concentration of inverted repeats, potentiating the formation of hairpin/cruciform structures.

Figure 1 shows the variation of concentration, $S_{[7,4000]}$, as a function of window number within chromosome 13. There is clearly higher enrichment of the concentration values in the central region of the chromosome. Also, most atypical values are located in the same region. Heterogeneity of concentration values is observed in all other chromosomes as well.

Table 1 Order statistics of scores, $S_{[7,4000]}$ over all the windows in each of the chromosomes and thresholds for finding windows with atypical $S_{[7,4000]}$ (critical values and Tukey thresholds)

chr	cv	% > cv (%)	min	Q_1	med	Q_3	max	T_{thr}	% > T_{thr} (%)
1	23	71	-2311	19	56	132	2423	301	7
2	27	68	-834	16	63	137	7079	317	6
3	25	62	-196	6	47	123	5219	298	6
4	26	60	-880	0	50	125	3595	312	4
5	26	60	-480	3	48	136	14139	336	4
6	29	61	-273	6	51	119	11399	287	5
7	26	68	-377	14	65	149	2624	352	6
8	24	68	-336	13	53	128	11210	302	6
9	20	72	-128	14	53	127	1638	297	7
10	22	79	-1542	27	65	129	4377	282	5
11	22	80	-84	29	69	141	5401	310	7
12	25	66	-269	13	56	131	4339	307	6
13	29	49	-412	-1	25	140	2976	351	3
14	22	62	-91	0	46	130	5581	326	5
15	24	65	-63	10	45	105	3366	247	7
16	19	79	-371	23	59	134	3128	301	9
17	22	86	-325	40	91	184	6954	399	6
18	25	60	-236	0	47	120	2528	301	5
19	19	83	-31	30	68	129	1815	278	8
20	19	85	-602	29	61	109	3107	227	9
21	27	62	-67	0	64	185	1894	463	4
22	19	62	-111	0	39	136	2010	339	8
X	20	57	-846	3	28	79	7690	193	8
Y	13	33	-1107	0	0	43	4722	107	13

Table 2 shows the windows with the 0.1% highest $S_{[7,4000]}$ values in the complete genome. For these windows, which are the ones with the most enriched concentration of inverted repeats, we studied the behaviour of S values at sub-intervals of distances. As observed in a previous paper [3], there are different patterns of enrichment of inverted repeats along the distances. The values presented in Table 2 also reveal the existence of different patterns of enrichment. The first eight windows show a higher enrichment for shorter distances, while window nine (chr6, win# 1609) increases for higher distances.

Figure 2 shows as examples the absolute frequencies of the distances for two windows with atypical concentrations of inverted repeats and with the two distinct patterns mentioned previously.

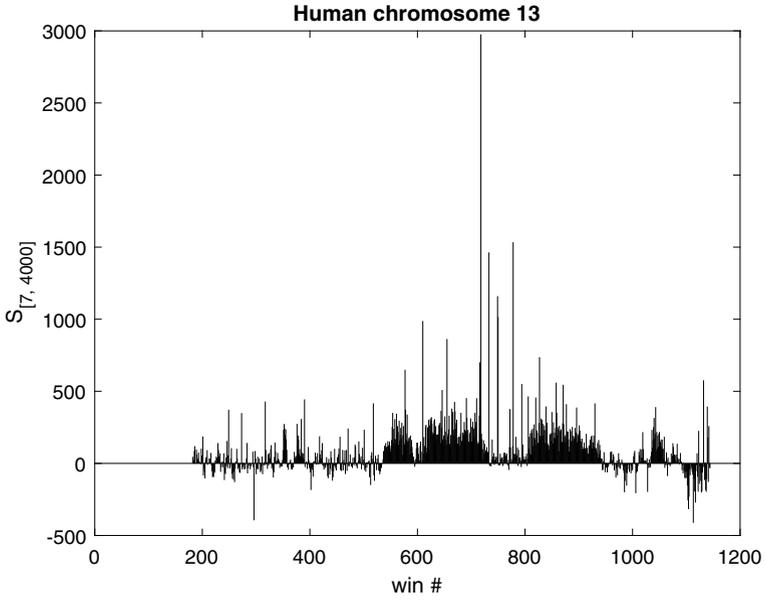


Fig. 1 Plot of $S_{[7,4000]}$ values for chromosome 13, showing concentration variation within the chromosome. win #—represents the sequential number of each windo

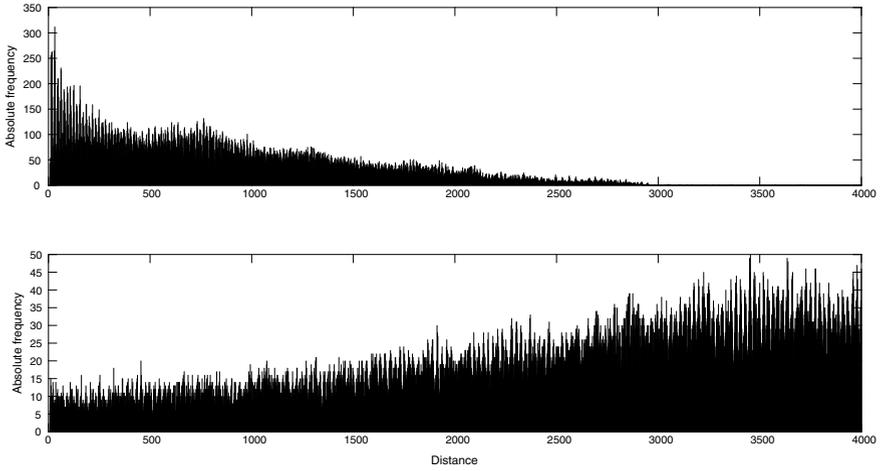


Fig. 2 Cumulative distance frequencies of two windows: top, chr5:67400001:67500000; bottom, chr6:160900001:161000000

Table 2 Values for the windows with the 0.1% highest $S_{[7,4000]}$ values

chr	win #	I_t	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
5	674	14139	6541	3968	2113	1032	378	107	0	0
6	315	11399	6091	3336	1466	444	61	0	0	0
8	576	11210	7684	3030	495	-1	1	0	0	0
8	1295	7847	5001	1823	759	248	15	0	0	0
X	531	7690	5288	1959	443	0	0	0	0	0
X	91	7175	4089	2109	723	252	1	0	0	0
2	1948	7079	5832	1209	28	8	2	0	0	0
17	135	6954	4247	1876	781	50	0	0	0	0
6	1609	6801	292	480	599	771	924	1127	1281	1327
X	1157	6588	4708	1658	221	0	0	0	0	0
14	859	5581	3417	1519	548	97	0	0	0	0
2	685	5420	3342	1296	546	203	32	0	0	0
11	1147	5401	848	740	693	671	643	628	607	572
3	1888	5219	4068	1091	59	1	0	0	0	0
5	1262	4737	3263	1229	242	3	0	0	0	0
Y	223	4722	3660	990	77	0	-2	0	-1	-2
X	2	4507	584	613	575	606	538	530	555	506
2	2265	4412	2578	1122	626	86	0	0	0	0
10	575	4377	2410	1173	538	219	36	2	0	0
Y	2	4368	566	595	557	588	521	514	538	490
12	866	4339	3381	918	40	0	0	0	0	0
8	38	4222	2588	1194	409	31	2	-2	-1	0
14	827	4002	3354	621	21	5	0	0	0	0
8	353	3698	2360	933	387	17	0	0	0	0
3	1119	3678	1948	1134	537	61	-1	-1	0	0
2	439	3627	922	690	509	419	351	278	245	214
4	1763	3595	2407	926	261	1	0	0	0	0
5	894	3390	2638	712	38	1	2	-1	0	0
15	231	3366	766	560	440	383	346	316	289	266
16	891	3128	1814	903	316	79	9	5	2	1

$I_t = S_{[7,4000]}$, $I_1 = S_{[7,500]}$, $I_2 = S_{[501,1000]}$, $I_3 = S_{[1001,1500]}$, $I_4 = S_{[1501,2000]}$,
 $I_5 = S_{[2001,2500]}$, $I_6 = S_{[2501,3000]}$, $I_7 = S_{[3001,3500]}$, $I_8 = S_{[3501,4000]}$.

4 Discussion and Conclusion

Motivated by the potential connection between the occurrence of inverted repeat pairs and the possible formation of hairpin/cruciform structures, we introduced a new measure to quantify the concentration of inverted repeats along the human genome.

In order to assess the relevance of our findings, a control scenario was created using a Markov model of order 7. The results of the human genome are clearly different to those of the control scenario, showing several regions with significant enrichment of the occurrence of inverted repeats. Thus, identifying regions with high potential for the formation of hairpin/cruciform structures.

The regions of enriched concentration of inverted repeats cannot be explained simply by well-known repetitive structures, such as those reported by RepeatMasker [12] and Tandem Repeats Finder [5], since we observed this phenomenon even in sequences of the human genome in which those repeats were masked.

The S measure allows the identification of regions with high potential of connection between inverted repeats. However, it is still not known if such potential translates directly to more frequent formation of hairpin/cruciform structures *in vivo*. We expect that the actual formation of those non-B structures depends not only on the higher overall potential of connection between inverted repeats, but also on the actual frequency distribution of distances between inverted repeats, which we observed to differ markedly between regions, even when they have similar overall potential (see Fig. 2).

Acknowledgements This work was supported by the Institute of Electronics and Informatics Engineering of Aveiro (IEETA) and Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia) references UIDB/00127/2020, UIDB/04106/2020, UIDP/04106/2020.

References

1. Bacolla A, Wells RD (2004) Non-B DNA conformations, genomic rearrangements, and human disease. *J Biol Chem* 279(46):47411–47414. <https://doi.org/10.1074/jbc.R400028200>
2. Bastos CAC, Afreixo V, Rodrigues JMOS, Pinho AJ (2018) An analysis of symmetric words in human DNA: adjacent vs non-adjacent word distances. In: PACBB 2018–12th International Conference on Practical Applications of Computational Biology & Bioinformatics. Toledo, Spain, June 2018. https://doi.org/10.1007/978-3-319-98702-6_10
3. Bastos CAC, Afreixo V, Rodrigues JMOS, Pinho AJ (2020) Detection and characterization of local inverted repeats regularities. In: Fdez-Riverola F, Rocha M, Mohamad MS, Zaki N, Castellanos-Garzón JA (eds) PACBB 2019, vol 1005. AISC. Springer, Cham, pp 113–120. https://doi.org/10.1007/978-3-030-23873-5_14
4. Bastos CAC, Afreixo V, Rodrigues JMOS, Pinho AJ, Silva R (2019) Distribution of distances between symmetric words in the human genome: analysis of regular peaks. *Interdiscip Sci Comput Life Sci* 11(3): 367–372. <https://doi.org/10.1007/s12539-019-00326-x>
5. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27(2):573. <https://doi.org/10.1093/nar/27.2.573>
6. Cer RZ et al (2010) Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res* 39(suppl_1):D383–D391. <https://doi.org/10.1093/nar/gkq1170>
7. Cer RZ et al (2012) Non-B DB v2. 0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res* 41(D1):D94–D100. <https://doi.org/10.1093/nar/gks955>

8. Du Y, Zhou X (2013) Targeting non-B-form DNA in living cells. *Chem Rec* 13(4):371–384. <https://doi.org/10.1002/tcr.201300005>
9. Inagaki H, Kato T, Tsutsumi M, Ouchi Y, Ohye T, Kurahashi H (2016) Palindrome-mediated translocations in humans: a new mechanistic model for gross chromosomal rearrangements. *Front Genet* 7:125. <https://doi.org/10.3389/fgene.2016.00125>
10. Kent W et al (2002) The human genome browser at UCSC. *Genome Res* 12(6):996–1006. <https://doi.org/10.1101/gr.229102>, <http://genome.ucsc.edu>
11. Kolb J et al (2009) Cruciform-forming inverted repeats appear to have mediated many of the microinversions that distinguish the human and chimpanzee genomes. *Chromosome Res* 17(4):469–483. <https://doi.org/10.1007/s10577-009-9039-9>
12. Smit AFA, Hubley R, Green P: RepeatMasker Open-4.0 (2013–2015). <http://www.repeatmasker.org>
13. Tavares AH et al (2017) DNA word analysis based on the distribution of the distances between symmetric words. *Sci Rep* 7(1):728. <https://doi.org/10.1038/s41598-017-00646-2>
14. Tukey JW (1977) *Exploratory Data Analysis*. Addison-Wesley, Boston
15. Wang Y, Leung FC (2006) Long inverted repeats in eukaryotic genomes: recombinogenic motifs determine genomic plasticity. *FEBS Lett* 580(5):1277–1284. <https://doi.org/10.1016/j.febslet.2006.01.045>

EvoPPI 2: A Web and Local Platform for the Comparison of Protein–Protein Interaction Data from Multiple Sources from the Same and Distinct Species



Miguel Reboiro-Jato, Jorge Vieira, Sara Rocha, André D. Sousa, Hugo López-Fernández, and Cristina P. Vieira

Abstract The understanding of the molecular basis of cellular processes and ultimately disease, requires knowledge on protein structures, interactions, and functions. Protein–protein interaction data (PPI) is available in the publicly available main PPI databases that show little overlap due to the use of different criteria. Therefore, web platforms that aggregate the data from multiple sources, such as EvoPPI (<http://evoppi.i3s.up.pt>), where the existing databases have been updated and new ones were added, as here described, and APID (<http://cicblade.dep.usal.es:8080/APID/init.action>) are useful. Still, in both EvoPPI 1.0 and 2, here presented, we have made

Hugo López-Fernández and Cristina P. Vieira are co-senior authors.
Miguel Reboiro-Jato and Jorge Vieira are contributed equally to this work.

M. Reboiro-Jato · H. López-Fernández (✉)
CINBIO, Department of Computer Science, ESEI-Escuela Superior de Ingeniería Informática,
Universidade de Vigo, 32004 Ourense, Spain
e-mail: hlfernandez@uvigo.es

M. Reboiro-Jato
e-mail: mrjato@uvigo.es

SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur),
SERGAS-UVIGO, 36213 Vigo, Spain

J. Vieira · S. Rocha · A. D. Sousa · H. López-Fernández · C. P. Vieira
Instituto de Investigação E Inovação Em Saúde (I3S), Universidade Do Porto, Rua Alfredo Allen,
208, 4200-135 Porto, Portugal
e-mail: jvvieira@ibmc.up.pt

S. Rocha
e-mail: sara.rocha@i3s.up.pt

A. D. Sousa
e-mail: andre.sousa@ibmc.up.pt

C. P. Vieira
e-mail: cgvieira@ibmc.up.pt

J. Vieira · H. López-Fernández · C. P. Vieira
Instituto de Biologia Molecular E Celular (IBMC), Rua Alfredo Allen, 208, 4200-135 Porto,
Portugal

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
F. Fdez-Riverola et al. (eds.), *Practical Applications of Computational Biology and Bioinformatics, 16th International Conference (PACBB 2022)*, Lecture Notes in Networks and Systems 553, https://doi.org/10.1007/978-3-031-17024-9_10

a special effort to make it flexible in what concerns the choice of the databases to be compared. Moreover, interacting protein pairs tend to be evolutionarily conserved, and thus the information available for one species might be used to predict the incompleteness of the network in another one, and identify putative missing interactions. This approach is now available in EvoPPI 2 for *Homo sapiens* and the model species *Mus musculus*, *Caenorhabditis elegans*, and *Drosophila melanogaster*, using either Ensembl (<https://www.ensembl.org>) or DIOPT Ortholog Finder (https://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl) orthologies/paralogies. Moreover, since not all available PPI data is present in the main databases (e.g. PPI observed in patient tissues and mutant animal species, where PPI might be aberrant, are usually not included in the main databases, although in several studies this has been shown not to be the case), we provide the needed tools (including a Ubuntu-based virtual machine where all software is already installed and ready-to-run) to run a local EvoPPI 2 instance and create a custom database from the existing ones. This way the user can add new data for any species and from any source database, creating custom interactomes. Administrator tools are provided to help in the automatic processing and conversion of files from various sources into the custom EvoPPI database format.

Keywords PPI · Web platform · Local platform · Docker · Database

1 Introduction

In order to understand the molecular basis of cellular processes and ultimately disease, protein structures, interactions, and functions must be elucidated [1, 2]. The identification of protein–protein interactions (PPI) in itself provides opportunities to explore biological functions (see for instance, [3]), which, in turn, can lead to a better understanding of the molecular basis of multiple diseases, and to the identification of possible therapeutic targets. The complete PPI network, the interactome, is however difficult to obtain due to the heterogeneity of the nature of the interactions. Indeed, some PPI are obligate, others are non-obligate. In the latter group, interactions can be permanent or transient. The interaction strength can also be strong or weak, and some proteins must undergo chemical modifications in order to be able to interact with their partners. Moreover, gene expression, and thus protein levels, vary among different tissues, and therefore not all PPIs may be observed in every tissue.

PPIs can be detected with high-throughput experimental techniques such as yeast-two-hybrid (Y2H) system [4], affinity purification followed by mass spectrometry (AP-MS) [5], luminescence-based mammalian interactome mapping (LUMIER) [6, 7] or literature-derived low-throughput experiments. All these techniques have known drawbacks that contribute, on one hand, to the identification of false positives, which for Y2H can be as high as 45% [8], but, on the other hand, to the incomplete network of interactions (see review [9, 10]), since in these methodologies only binary interactions are identified. These results are available in the main PPI databases such as (BioGRID [11, 12], CCSB [13], DroID [14], FlyBase [15],

HIPPIE [16], HitPredict [17], HomoMINT [18], INstruct [19], Interactome3D [20], Mentha [21], MINT [22, 23], and PINA [24]). Since different criteria have been used to build these databases, the overlap is low among them. Therefore, it is useful to have a web application such as EvoPPI [25], that shows, for a given protein, the results for the different databases side by side, but that also allows their integration as if they were a single database. The databases used by EvoPPI 1.0 version (whose database is currently named 2018.03) were updated, leading to increments in the number of unique interactions in the range of 6.42% up to 119.95%, depending on the species. Moreover, since interacting protein pairs tend to be evolutionarily conserved [26], the PPI information that is acquired for one species can, in principle, be transposed to another one, as long as orthologous gene pairs can be identified, using, for instance, BLAST, Ensembl¹ orthologies, or the DIOPT Ortholog Finder webpage.² In EvoPPI 1.0 version, there is already a “Distinct species” option that allows the use of a BLAST approach (where the user can specify the maximum number of target genes, the minimum expect value, minimum length of aligned block, and percentage of minimum identity). Nevertheless, this is a time consuming process as BLAST queries can last several hours.

In order to address this issue, in the new EvoPPI 2 version here reported, the predicted interactomes of *Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans*, and *Drosophila melanogaster*, based on the updated databases for these model organisms has been pre-computed, and are available for search as a regular interactome. Gene orthologies were derived using either Ensembl or DIOPT Ortholog Finder. The use of pre-computed predicted interactomes can give hints on whether a given interaction may be real or not, as well as on the incompleteness of the network for a given protein. There are, however, many other computational methods that can be used, as well, to infer interactomes [10]. The result of such methods can be uploaded to a local version of EvoPPI 2 as a list of Entrez Gene ID pairs describing the inferred interactions. It should, however, be noted that such computational methods can generate false-positive interactions, similar to the high throughput techniques [27–29].

2 Methods

2.1 Data

The EvoPPI architecture, data structure, interactome comparison algorithms, and basic user interface has been previously described [25]. The current version of the EvoPPI database (2022.04) includes 101 interactome datasets for ten animal species (*Bos taurus*, *Caenorhabditis elegans*, *Danio rerio*, *Drosophila melanogaster*, *Gallus*

¹ <https://www.ensembl.org/>.

² https://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl.

gallus, Homo sapiens, Mus musculus, Oryctolagus cuniculus, Rattus norvegicus, and Xenopus laevis), obtained from main PPI databases (BioGRID [11, 12], CCSB [13], DroID [14], FlyBase [15], HIPPIE [16], HitPredict [17], HomoMINT [18], INstruct [19], Interactome3D [20], Mentha [21], MINT [22, 23], and PINA [24]), and prepared as previously described [25]. Moreover, as detailed in the next section, 462 Predicted Interactome files were created for H. sapiens, M. musculus, C. elegans, and D. melanogaster, based on the gene orthologies established in DIOPT Ortholog Finder (the options used are: all “Ortholog Sources” and the “Exclude low score (score > 1, unless only match score is 1)” filter), and Ensembl Genome Browser, for these same species.

The EvoPPI philosophy is to give the user the possibility to see the results based on individual databases, as well as the databases of choice as an aggregate. Therefore, for HINT database, for instance, we downloaded the interactions as binary physical interactions (direct biophysical interaction between two proteins) and co-complex associations (provide information about co-membership in a complex), separately. The same applies to those obtained by literature-curation (LC), high-throughput experiments (HT), or sub-interactomes assigned as high-quality (hq). For CCSB Interactome Database, we downloaded the Human Reference Protein Interactome (HuRI, also known as HI-III-19 [13]), HI-union (an aggregate of all PPIs identified in HI-I-05, HI-II-14, HuRI, Venkatesan-09, Yu-11, Yang-16, and Test space screens-19), CCSB Yang-16 (where the extent to which different protein isoforms perform different functions within the cell is assessed; [30]), and CCSB Test_space_screens-19 (independent, reciprocal Y2H assay screens on a search space of $\sim 1,800 \times \sim 1,800$ genes; [13]). The same rationale was applied to DroID database from where we downloaded the “Genetic Interaction Data” (named DroIDfly at EvoPPI), the “PPI curated by Flybase” (DroID_PPI_curated_by_FlyBase), and the “PPI from other databases” (DroID_PPI_from_other_DBs) interaction datasets. The remaining databases were treated as before [25] and updated (a summary of the between versions change can be seen at EvoPPI’s site³).

2.2 Web Interface Updates

The large increase in the number of EvoPPI databases, as well as the new resources made available, implied several changes to its web interface in order to ease the user interaction and the visualization of the results, namely:

- a) A select button has been added to both the “Distinct species” and “Same species” query pages that allows selecting or deselecting all databases as well as searching by keyword names. In the case of “Predicted Interactomes”, for convenience, the user can also choose from a list the species upon which the predicted interactomes are based, as well as whether homologies are based on data from Ensembl or DIOPT Ortholog Finder.

³ <http://evoppi.i3s.up.pt/help>.

- b) Under the “Interactions table” of the “Same Species Results” section, the user can now choose to collapse the results for the different interactomes, or see the result for a subset of the interactomes/predicted interactomes included in the query. It should be noted that the “Show chart” option will still give the results for all the interactomes/predicted interactomes, since this image is created upon the initial request. In order to improve readability, uncollapsed results are shown for a maximum of ten interactomes/predicted interactomes.
- c) In the new “Species” tab, the user can explore the available proteomes and download for each species a FASTA file containing the protein sequences associated with each Gene ID.
- d) In the new “Interactomes” and “Predicted interactomes” tabs, the user can explore the available interactomes/predicted interactomes and download the corresponding PPI files for each species and database. Such PPI files are two-column TSV files where each line represents an interaction between two proteins (Gene IDs).
- e) In the new “Databases” tab, the user can see the basic information about the data used by the different EvoPPI database versions. There is also a tutorial on how to use EvoPPI locally with previous or custom database versions. Next subsection provides further details about this.

2.3 Using EvoPPI Locally

So far, EvoPPI 1.0 could only be used through the public instance maintained by us. With this update, we have released the *evoppi-docker* project,⁴ which allows the deployment of a local EvoPPI 2 instance using a Docker environment for the database, the backend, and the frontend. Thus, researchers can now use their own EvoPPI instances for performing their analysis and select the specific database version that will be used. In addition, using this local version also allows researchers to register new species (as long as a GBFF file is available for them), new interactomes, and new predicted interactomes easily, as well as to manage the existing ones.

We also provide a virtual Ubuntu disk image, where all software is already installed and running, for users without informatics expertise. However, it should be noted that performance may be lower than in a local or Docker installation, due to the use of a virtual machine.

3 Results and Discussion

As shown in the EvoPPI 2 website⁵ under the “Database” tab, 51 new databases were added to EvoPPI. Moreover, 39 databases were updated (on average, each of the databases shows an increase of 1.36 times), and 462 predicted interactome files

⁴ <https://github.com/sing-group/evoppi-docker>.

⁵ <http://evoppi.i3s.up.pt/>.

were also created for *H. sapiens*, *M. musculus*, *C. elegans*, and *D. melanogaster* based on the available databases for these species. The increase in the number of unique interactions from database version 2018.03 to 2022.04 is shown in Fig. 1 for *H. sapiens* and three model species. Such increase is a warning on how incomplete the interactomes for the different species are. It should be noted that PPI obtained from tissues of patients or mutant animal model species used to study PPI in several diseases, for instance, are not included in these main databases [31]. This is the case of the neurodegenerative diseases caused by the expansion of the polyglutamine (polyQ) stretch [32], where it is well established that the polyQ expansion alters the native PPI, implying different accessibility at specific interacting residues, post-translational modifications, RNA binding regions, or chaperone binding regions, needed for the normal protein activity, and not because of novel complexes formed by misfolded or aggregating proteins [25, 33, 34].

Interacting protein pairs tend to be evolutionarily conserved [34]. Therefore, the PPI information that is acquired for one species can, in principle, be transposed to another one, as long as orthologous gene pairs can be identified, using bidirectional BLAST, for instance. Nevertheless, when transposing data from protostomian species to deuterostomian species or vice versa (such as predicting *H. sapiens* PPIs, based on

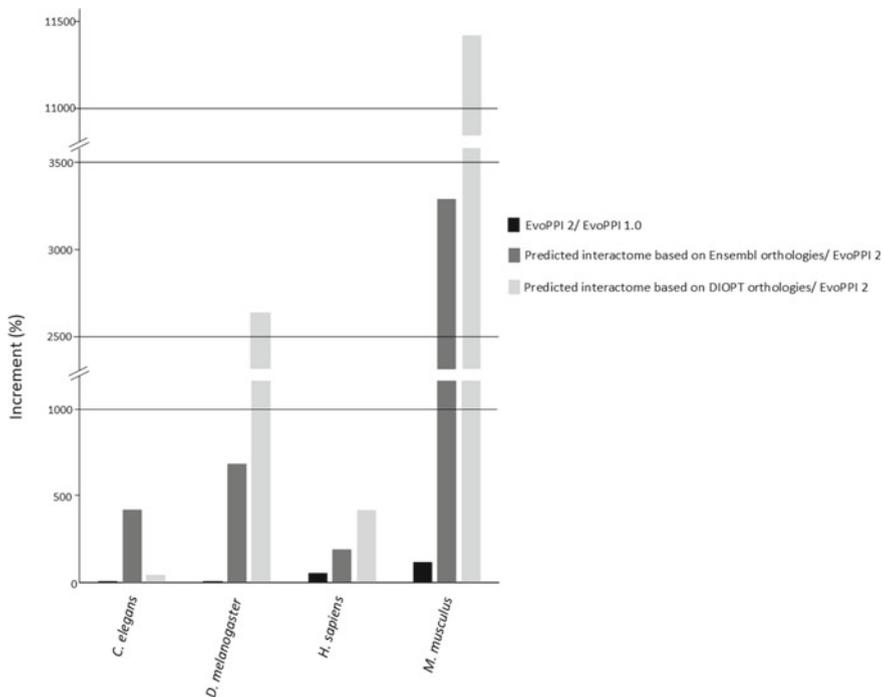
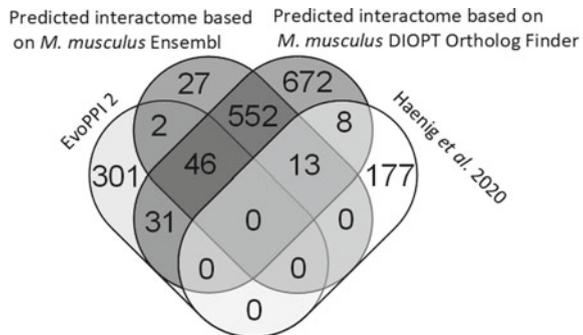


Fig. 1 Increment (%) in the number of interactions reported in EvoPPI 2 relative to EvoPPI 1.0, for *H. sapiens* and the model species *M. musculus*, *C. elegans*, and *D. melanogaster*

data acquired for *D. melanogaster*), only paralogous gene pairs can be established, due to the two whole genome duplications events that happened early in vertebrate evolution (see for instance, [35]). In this case, this approach likely works, although with a lower degree of success. In EvoPPI 2, we provide predicted interactomes for *H. sapiens*, *M. musculus*, *C. elegans*, and *D. melanogaster* based on the data available for these species at a given PPI database, using Ensembl or DIOPT Ortholog Finder orthologies.

The predicted interactomes can be useful to understand how incomplete a given species PPI network might be, and identify the missing putative interactions, which can then be validated by performing hypothesis-driven experiments using techniques such as Y2H interaction assays, or searching the literature for results not included in the main PPI databases. For instance, ATXN1 causes Spinocerebellar ataxia type 1 (SCA1) when it shows an expanded number of trinucleotide repeats in the polyglutamine tract [36], and thus, it is of interest to fully characterize its network. When using all EvoPPI 2 databases, there are 380 and 665 ATXN1 interactors for *H. sapiens* and *M. musculus*, respectively (Fig. 2). Since human disease genes are largely conserved between the two genomes (99.5%; [37]), and PPI are largely conserved between species [26], the above observation suggests that the human ATXN1 network is likely very incomplete. Using the human predicted interactomes based on the 15 *M. musculus* databases, 640 and 1321 PPI are predicted, depending on whether Ensembl or DIOPT Ortholog Finder are used to establish orthologies. Therefore, by performing such analyses, the human ATXN1 network increases to 972 when using Ensembl, and 1652 when using DIOPT Ortholog Finder orthologies. When using proteomic data not present in the main PPI databases (Fig. 2), further evidence is obtained for the incompleteness of the ATXN1 network. For instance, there are 198 ATXN1 interactors reported in [38] that are not listed in any of the main PPI databases, and thus, are not present in EvoPPI. Out of these 198 interactors, 13 could have been predicted by the EvoPPI predicted interactomes from *M. musculus*, using either Ensembl or DIOPT Ortholog Finder, and eight could have been predicted using the latter approach only. Taken together, this suggests that the ATXN1 interaction network is very large, although the presence of a large number of false positive interactions could also be an explanation.

Fig. 2 Comparison of the 198 ATXN1 interactors reported in [38] not present in EvoPPI 2 main interactome databases, with the human predicted interactomes based on *M. musculus* and Ensembl or DIOPT Ortholog Finder orthologies/paralogies



Finally, the possibility that is now given of creating a local EvoPPI 2 instance using the *evoppi-docker* project mentioned before has many advantages: a) it allows the use of local computing resources, which for time consuming large scale analyses may be advantageous; b) unpublished data or data subject to legal restrictions can be analysed without leaving the institution where it was obtained; c) by using the EvoPPI 2 RESTful API, the data stored on a local EvoPPI database can be accessed by a pipeline/script running in headless servers, which is useful for the case of large scale analyses; and d) it allows the reproducibility of the analyses made with previous database versions. In addition, it should be noted that EvoPPI 2 includes the tools required to convert and prepare new interactomes databases in the required format (two-column TSV files where each line represents an interaction between two proteins represented as Gene IDs). Such tools perform the required steps, such as the removal of file headers, isolation of the columns containing the interactions data, the removal of prefixes and suffixes that may be associated with the gene/protein identifiers, the extraction of data for a single species from multispecies files, and most importantly, the conversion of most commonly used gene/protein identifiers to Gene ID (using the UniProt database identifier mapping service⁶). Tools are also given to prepare new species files from genome GBFF files, obtained from the NCBI Assembly database,⁷ for instance. With these tools, researchers can easily build custom databases, dedicated to the study of a single disease, for instance, where they can incorporate all PPI data that is relevant for such a study. The possibility to create a local custom database is a unique feature of EvoPPI.

4 Conclusion

EvoPPI 2 (<http://evoppi.i3s.up.pt>) is a flexible web resource, where the user can easily choose and retrieve the available PPI data for a given protein, or proteins that are less than a number of steps away from that protein. The publicly available EvoPPI 1.0 interactome databases have been updated in EvoPPI 2, and new ones were added, as well. Predicted interactome files were also created for *H. sapiens*, *M. musculus*, *C. elegans*, and *D. melanogaster* based on the available interactome databases for these species, using two alternatives for the establishment of gene orthologies/paralogies (those from Ensembl and DIOPT Ortholog Finder), which enables the fast and easy comparison of interactome data from different species. Nevertheless, there is also the possibility (already available in EvoPPI 1.0) of performing tailored made BLAST searches to identify orthologous genes, although this is a much more time consuming option. Since not all available PPI data is present in the main databases, EvoPPI 2 can also be deployed as a local instance, giving the user the opportunity to use species and interactome databases (including unpublished ones) not included in EvoPPI 2.

⁶ <https://www.uniprot.org/help/api%5Fidmapping>.

⁷ <https://www.ncbi.nlm.nih.gov/assembly>.

Tools are provided (including the conversion between different types of gene/protein identifiers) to convert the files of interest into the EvoPPI format.

Acknowledgements This work was funded by: (i) National Funds through FCT—Fundação para a Ciência e a Tecnologia, I.P., under the project UIDB/04293/2020, and (ii) Consellería de Educación, Universidades e Formación Profesional (Xunta de Galicia) under the scope of the strategic funding ED431C2018/55-GRC Competitive Reference Group. H. López-Fernández is supported by a “María Zambrano” post-doctoral contract from Ministerio de Universidades (Gobierno de España).

References

1. Blackstock WP, Weir MP (1999) Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol* 17(3):121–127
2. (1999) The promise of proteomics. *Nature* 402(6763):703
3. Peng W, Wang J, Cai J, Chen L, Li M, Wu F-X (2014) Improving protein function prediction using domain and protein complexes in PPI networks. *BMC Syst Biol* 8(1):1–13
4. Yu H, Braun P, Yıldırım MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N (2008) High-quality binary protein interaction map of the yeast interactome network. *Science* 322(5898):104–110
5. Dunham WH, Mullin M, Gingras AC (2012) Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics* 12(10):1576–1590
6. Petschnigg J, Snider J, Stagljar I (2011) Interactive proteomics research technologies: recent applications and advances. *Curr Opin Biotechnol* 22(1):50–58
7. Stynen B, Tourneau H, Tavernier J, Van Dijck P (2012) Diversity in genetic in vivo methods for protein-protein interaction studies: from the yeast two-hybrid system to the mammalian split-luciferase system. *Microbiol Mol Biol Rev* 76(2):331–382
8. Huang H, Jedynak B, Bader J (2007) Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput Biol* 3(11):e214
9. Szilagyí A, Grimm V, Arakaki AK, Skolnick J (2005) Prediction of physical protein–protein interactions. *Phys Biol* 2(2):S1
10. Keskin O, Tuncbag N, Gursoy A (2016) Predicting protein–protein interactions from the molecular to the proteome level. *Chem Rev* 116(8):4884–4909
11. Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, Kolas N, O’Donnell L, Leung G, McAdam R (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res* 47(D1):D529–D541
12. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 34(suppl_1):D535–D539
13. Luck K, Kim D-K, Lambourne L, Spirohn K, Begg BE, Bian W, Brignall R, Cafarelli T, Campos-Laborie FJ, Charlotteaux B (2020) A reference map of the human binary protein interactome. *Nature* 580(7803):402–408
14. Murali T, Pacifico S, Yu J, Guest S, Roberts GG, Finley RL (2011) DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for *Drosophila*. *Nucleic Acids Res* 39(suppl_1):D736–D743.
15. Attrill H, Falls K, Goodman JL, Millburn GH, Antonazzo G, Rey AJ, Marygold SJ, Consortium F (2016) FlyBase: establishing a Gene Group resource for *Drosophila melanogaster*. *Nucleic Acids Res* 44(D1):D786–D792
16. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH: HIPPIE v2. 0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic acids research* 2016:gkw985.

17. López Y, Nakai K, Patil A (2015) HitPredict version 4: comprehensive reliability scoring of physical protein–protein interactions from more than 100 species. In: Database 2015
18. Persico M, Ceol A, Gavrila C, Hoffmann R, Florio A, Cesareni G (2005) HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinf* 6(4):1–12
19. Meyer MJ, Das J, Wang X, Yu H (2013) INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* 29(12):1577–1579
20. Mosca R, Céol A, Aloy P (2013) Interactome3D: adding structural details to protein networks. *Nat Methods* 10(1):47–53
21. Calderone A, Castagnoli L, Cesareni G (2013) Mentha: a resource for browsing integrated protein-interaction networks. *Nat Methods* 10(8):690–691
22. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G (2002) MINT: a Molecular INTeraction database. *FEBS Lett* 513(1):135–140
23. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardoza AP, Santonico E (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 40(D1):D857–D861
24. Cowley MJ, Pinese M, Kassahn KS, Waddell N, Pearson JV, Grimmond SM, Biankin AV, Hautaniemi S, Wu J (2012) PINA v2.0: mining interactome modules. *Nucleic Acids Res* 40(D1):D862–D865
25. Vázquez N, Rocha S, López-Fernández H, Torres A, Camacho R, Fdez-Riverola F, Vieira J, Vieira CP, Reboiro-Jato M (2019) EvoPPI 1.0: a web platform for within-and between-species multiple Interactome comparisons and application to nine PolyQ proteins determining neurodegenerative diseases. *Interdisc Sci Comput Life Sci* 11(1):45–56
26. Pagel P, Mewes H-W, Frishman D (2004) Conservation of protein–protein interactions – lessons from ascomycota. *Trends Genet* 20(2):72–76
27. Sittler RMK, Ideker T (2005) From the cover: conserved patterns of protein interaction in multiple species
28. De Las RJ, Fontanillo C (2012) Protein–protein interaction networks: unraveling the wiring of molecular machines within the cell. *Brief Funct Genom* 11(6):489–496
29. Mrowka R, Patzak A, Herzog H (2001) Is there a bias in proteome research? *Genome Res* 11(12):1971–1973
30. Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, Sun S, Yang F, Shen YA, Murray RR (2016) Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* 164(4):805–817
31. Nibbe RK, Chowdhury SA, Koyutürk M, Ewing R, Chance MR (2011) Protein–protein interaction networks and subnetworks in the biology of disease. *Wiley Interdisc Rev Syst Biol Med* 3(3):357–367
32. Silva A, de Almeida AV, Macedo-Ribeiro S (2018) Polyglutamine expansion diseases: more than simple repeats. *J Struct Biol* 201(2):139–154
33. Suter B, Fontaine J-F, Yildirimman R, Rasko T, Schaefer MH, Rasche A, Porras P, Vazquez-Alvarez BM, Russ J, Rau K (2013) Development and application of a DNA microarray-based yeast two-hybrid system. *Nucleic Acids Res* 41(3):1496–1507
34. Rocha S, Vieira J, Vázquez N, López-Fernández H, Fdez-Riverola F, Reboiro-Jato M, Sousa AD, Vieira CP (2019) ATXN1 N-terminal region explains the binding differences of wild-type and expanded forms. *BMC Med Genom* 12(1):1–14
35. Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* 3(10):e314
36. Orr HT, Zoghbi HY (2007) Trinucleotide repeat disorders. *Annu Rev Neurosci* 30:575–621
37. Huang H, Winter EE, Wang H, Weinstock KG, Xing H, Goodstadt L, Stenson PD, Cooper DN, Smith D, Albà MM (2004) Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. *Genome Biol* 5(7):1–15
38. Haenig C, Atias N, Taylor AK, Mazza A, Schaefer MH, Russ J, Riechers S-P, Jain S, Coughlin M, Fontaine J-F (2020) Interactome mapping provides a network of neurodegenerative disease proteins and uncovers widespread protein aggregation in affected brains. *Cell Rep* 32(7):108050

Author Index

A

Afreixo, Vera, [89](#)
Agís-Balboa, Roberto C., [31](#)
Azé, Jérôme, [55](#)

B

Balsano, Clara, [23](#)
Bastos, Carlos A. C., [89](#)
Bringay, Sandra, [55](#)
Butler, Gregory, [1](#)

C

C Canal-Alonso, Ángel, [13](#)
Casano, Nicolò, [23](#)
Corchado, Juan Manuel, [1](#)
Cunha, Emanuel. *See* [79](#)

D

de la Prieta, Fernando, [13](#)
Dias, Oscar, [79](#)
Duarte-Pereira, Sara, [43](#)

F

Faria, Cristiana, [79](#)

G

Gavara, Laurent, [55](#)
Ghazikhani, Hamed, [1](#)

H

Hernández, Marco, [13](#)

K

Kar, Anuradha, [67](#)

L

López-Fernández, Hugo, [31](#), [101](#)
Louet, Maxime, [55](#)

M

Matos, Sérgio, [43](#)

O

Oliveira, Alexandre, [79](#)
Oliveira, José Luís, [43](#)

P

Papastergiou, Thomas, [55](#)
Pérez-Rodríguez, Daniel, [31](#)
Pérez-Rodríguez, Mateo, [31](#)
Pinho, Armando J., [89](#)
Poncelet, Pascal, [55](#)
Prieto, Javier, [13](#)

R

Reboiro-Jato, Miguel, [101](#)
Rocha, Sara, [101](#)
Rodrigues, João M. O. S., [89](#)
Rodríguez, Sara, [13](#)

S

Santini, Silvano Junior, [23](#)
Silva, Migue, [79](#)

Silva, Raquel M., [43](#)
Sinatti, Gaia, [23](#)
Sousa, André D., [101](#)

V

Vieira, Cristina P., [101](#)
Vittorini, Pierpaolo, [23](#)
Vieira, Jorge, [101](#)