BBCC2020

Bioinformatics and Computational Biology Conference

November 16-18, 2020

organized with the Department of Agricultural Sciences,
Università di Napoli Federico II
Portici, Naples, Italy

Virtual Conference

Program
and
(unofficial)
Abstract Book

BBCC2020 Scientific Committee

Dr. Angelo Facchiano – National Research Council, Institute of Food Sciences, Avellino, Italy
(Chair / BBCC Coordination)
Prof. Maria Luisa Chiusano – University of Naples "Federico II", Italy
(Chair)

Dr. Claudia Angelini – National Research Council, Institute for Applied Mathematics "M. Picone", Italy
Prof. Emmanuelle Becker – University of Rennes, France
Prof. Michele Ceccarelli – Univerisity of Naples "Federico II", Italy
Dr. Alessandro Cestaro – Edmund Mach Foundation, San Michele all'Adige, Italy
Prof. Sergio Cocozza – University of Naples "Federico II", Italy
Prof. Olivier Dameron – University of Rennes, France
Dr. Nunzio D'Agostino – University of Naples "Federico II", Italy
Prof. Paola Festa – University of Naples "Federico II", Italy
Dr. Francesco Giannino – University of Naples "Federico II", Italy
Prof. David Gilbert – Brunel University London, UK
Dr. Mario Rosario Guarracino – National Research Council, High Performance Computing and Networking Institute, Naples, Italy
Prof. Antonio Gomez – Rheumatology Research Group (GRR-VHIR) and Vic's University Barcelona, Spain
Prof. Dr. Dominik Heider – University of Marburg, Germany
Dr. Giuliano Langella – University of Naples "Federico II", Italy
Prof. Anna Marabotti – University of Salerno, Italy
Dr. Bogdan Mirauta – Romanian Bioinformatics Society
Prof. Marius Mihasan – University of Iasi, Romania
Dr. Edoardo Pasolli – University of Naples "Federico II", Italy
Dr. Paolo Romano – Ospedale Policlinico San Martino, Genoa, Italy
Prof. Fabrizio Sarghini – University of Naples "Federico II", Italy
Prof. Roberto Tagliaferri – University of Salerno, Italy

BBCC2020 Organizing Committee

Dr. Nunzio D'Agostino, Prof. Maria Luisa Chiusano, Dr. Angelo Facchiano

Contacts

Web site: http://www.bbcc-meetings.it
E-mail: bbcc.meetings@gmail.com

BBCC2020 is a ISCB affiliated Conference, organized under the patronage of BITS - Bioinformatics Italian Society

Conference Program

Central Europe Time

Monday 16

14:00-14:30 Connection to the Virtual Conference Room and Registration

14:30 **Conference Opening**
Maria Luisa Chiusano and Angelo Facchiano – Chairs of BBCC2020

14:45-15:30 Keynote Lecture
Nikolay V. Dokholyan - Penn State College of Medicine, Hershey, PA, USA
Molecular Design for Research and Therapeutics

Session – Bioinformatics for molecular diseases
Chair: Maria Luisa Chiusano, University of Naples Federico II, Naples, Italy

15:30-15:50 **Maria Monticelli**, Andrea Riccio, Mehdi Totonchi, David B Ascher and Maria Vittoria Cubellis
Analysis of whole-exome sequencing and protein modelling: the lesson from cyclin B

15:50-16:10 **Aditi Deokar**
Stratification of Systemic Lupus Erythematosus Patients with Gene Expression Data Reveals Expression of Distinct Immune Pathways

16:10-16:30 **Antonio Facchiano**, Francesco Facchiano, Angelo Facchiano
Molecular basis of cancer co-morbidities in COVID-19 patients

16:30-16:50 **Sohayb Bekkal Brikci**, Imane Abdelli and Faiçal Hassani
Inhibition of angiotensin converting enzyme 2 SARS-CoV-2's receptor by natural compounds: in silico structure-activity relationship study

16:50-17:10 Simone Ciccolella, Luca Denti, Paola Bonizzoni, Gianluca Della Vedova, **Yuri Pirola** and Marco Previtali
MALVIRUS: an integrated web application for viral variant calling

17:10-17:30 **Flavia Villani**, Francesco Porto, Andrea Guarracino, Robert W.Williams, Pjotr Prins, Gianluca Della Vedova, Erik Garrison and Vincenza Colonna
Population genomics analyses on pangenome graphs

17.30-17.40 ***Short presentations / posters***

Carmen Biancaniello, Antonia D'Argenio, Serena Dotolo, Deborah Giordano, Bernardina Scafuri, Antonio d'Acierno, Anna Marabotti, Roberto Tagliaferri and Angelo Facchiano
Investigating structural and functional properties of menin protein

Francesco Monticolo, Emanuela Palomba and Maria Luisa Chiusano
Identification of novel potential genes involved in programmed cell death by integrated and comparative analyses

17.40-18:00 **Discussion and session closure**

Tuesday 17

9:00-9:20 Connection to the Virtual Conference Room

Invited Session: ELIXIR-IT for COVID-19

Chair: Federico Zambelli – University of Milan, Italy – ELIXIR-IT

9:20-9:30 **Federico Zambelli -** University of Milan, Italy, and ELIXIR-IT
Introduction to ELIXIR infrastructure in Italy

9:30-9:50 **Matteo Chiara -** University of Milan, Italy, and ELIXIR-IT
Open data (and bioinformatics) in the COVID-19 pandemic

9:50-10:10 **Marco Tangaro** – IBIOM-CNR, Bari, Italy and ELIXIR-IT
Galaxy based services for Covid-19 research

10:10-10:30 **Allegra Via** – IBPM-CNR, Roma, Italy, and ELIXIR-IT
Remote teaching and online learning: challenges and opportunities

10:30-10:40 Break

Session: Infrastructures and Services

Chair: Nunzio D'Agostino, University of Naples, Italy

10:40-10:50 **Rosa Siciliano** – ISA-CNR, Avellino, Italy
METROFOOD: an European Research Infrastructure for promoting metrology in Food and Nutrition

10:50-11:10 **Katharina Lauer** – ELIXIR-EU Industry Officer
ELIXIR- promoting public-private partnerships to advance science

Session: Applications of computational methods

Chair: Nunzio D'Agostino, University of Naples, Italy

11:10-11:30 Anna Marabotti, Eugenio Del Prete, **Bernardina Scafuri** and Angelo Facchiano
Assessing the performances of protein stability predictors

11:30-11:50 **Chitaranjan Mahapatra**
Computational Study of Action Potential Generation in Uterine Smooth Muscle cell

11:50-12:10 **Varsha Poondi Krishnan**, Monika Krzak, Shir Toubiana, Sara Selig, Claudia Angelini and Maria Rosaria Matarazzo
Studying the early molecular defects of ICF syndrome in patient-derived and CRISPR-corrected iPSCs using an integrated multi-omic approach

12:10-12:20 ***Short presentations / posters***

Giorgio Maria Vingiani, Pasquale De Luca, Daniele De Luca and Chiara Lauritano
Microalgal RNA-seq analyses to identify enzymes involved in the synthesis of bioactive compounds

Jamal Elhasnaoui, Giulio Ferrero and Michele De Bortoli
A comprehensive evaluation of differential alternative splicing tools for RNA-seq data

12:20-12:30 **Discussion and session closure**

14:10-14:30 Connection to the Virtual Conference Room

Session: Plant Sciences

Chair: Maria Luisa Chiusano, University of Naples Federico II, Naples, Italy
and Alessandro Cestaro, Edmund Mach Foundation, Italy

Invited lectures

14:30-15:15 **Alessandro Cestaro -** Edmund Mach Foundation, Italy
Introduction to Elixir Plant User Community. Apple as model ... lessons learnt

15:15-16:00 **Cyril Pommier -** French National Institute for Agriculture, Food and Environment – URGI, INRA, Université Paris-Saclay, Versailles, France
Plant data management, Elixir services and recommendations for FAIR data publication and findability

16:00-16:45 **Astrid Junker -** Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Germany
FAIR Phenomics Data Management

16:45-17:30 **Kristina Gruden -** Department of Biotechnology and Systems Biology, National Institute of Biology, Slovenia
Tools for visualisation of heterogeneous datasets in plant sciences

17:30-17:50 **Francesco Loreto –** Department of Biology, Agriculture and Food Sciences, The National Research Council of Italy
EMPHASIS: The European infrastructural programme on plant phenothyping

Oral contributions proposed by participants

17:50-18:10 **Georgia Tooulakou**, Paraskevi Manolaki, Caroline Urup Byberg, Franziska Eller, Brian Keith Sorrell, Tenna Riis and Maria Klapa
Integrated eco-physiological and metabolomic analyses of the amphibious plant Butomus umbellatus under light limitation and nutrient varying conditions

18:10-18:20 ***Short presentations / posters***
Pierre Larmande
The AgroLD project: A Knowledge Graph Database for plant functional genomics
Emanuela Palomba, Francesco Monticolo, Stefano Mazzoleni and Maria Luisa Chiusano
Integrated bioinformatics to investigate novel biological processes in model species

18:20-18:40 **Alessandro Cestaro and Maria Luisa Chiusano – Session chairs**
Discussion and session concluding remarks

Wednesday 18

9:00-9:20 Connection to the Virtual Conference Room

Session: Methods for biological data analysis

Chair: Claudia Angelini, National Research Council, Naples, Italy

9:20-9:40 **Marco Anteghini**, Edoardo Saccenti and Vitor A.P. Martins Dos Santos

Exploiting deep learning embeddings for sub-peroxisomal localisation

9:40-10:00 **Achal Dhariwal,** Roger Junges, Tsute Chen, and Fernanda Petersen

ResistoXplorer: a web-based tool for visual, statistical and exploratory data analysis of resistome data

10:00-10:20 **Pietro Hiram Guzzi**, Giuseppe Tradigo and Pierangelo Veltri

A novel algorithm for extracting common modular communities from Dual Networks.

10:20-10:40 **Giacomo Baruzzo**, Ilaria Patuzzi and Barbara Di Camillo

Zero-imputation in 16S rRNA gene studies: do we need it?

10:49-11:00 **Massimo Bellato**, Lorenzo Pasotti, Giuseppe Serio, Michela Casanova, Barbara Di Camillo and Paolo Magni

Novel multi- input logic gates for Synthetic Biology: analysis of the interaction between transcription factors and CRISPR interference

11:00-11:10 ***Short presentations/Posters***

Claudia Angelini, Daniela De Canditiis and **Anna Plaksienko**

jewel: a novel method for data integration with applications to omics data analysis

Giulia Babbi, Pier Luigi Martelli and Rita Casadio

Identifying biological functions underlying phenotypes using PhenPath

Invited lecture

11:10-12:00 **Ivo Grosse -** Institute of Computer Science, Martin Luther University Halle-Wittenberg, Germany, and German Center of Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Germany

Hourglass Patterns of Embryonic and Postembryonic Development in Animals and Plants and the Emergence of Biodiversity on our Planet

12:00 **Maria Luisa Chiusano and Angelo Facchiano – Chairs of BBCC2020**

Meeting announcements and closing remarks

Abstracts of oral presentations

Analysis of whole-exome sequencing and protein modelling: the lesson from cyclin B

Maria Monticelli [1], Andrea Riccio [2,3], Mehdi Totonchi [4,5], David B Ascher [6,7,8], Maria Vittoria Cubellis [1]

[1] Dept. Biology, University of Naples "Federico II," 80126 Napoli, Italy

[2] Department of Environmental Biological and Pharmaceutical Sciences and Technologies (DiSTABiF), Università degli Studi della Campania "Luigi Vanvitelli", Caserta, Italy

[3] Institute of Genetics and Biophysics (IGB) "Adriano Buzzati-Traverso", Consiglio Nazionale delle Ricerche (CNR), Naples, Italy

[4] Department of Genetics, Reproductive Biomedicine Research Center, Royan Institute for Reproductive Biomedicine, ACECR, Tehran, Iran

[5] Department of Stem Cells and Developmental Biology, Cell Science Research Center, Royan Institute for Stem Cell Biology and Technology, ACECR, Tehran, Iran

[6] Structural Biology and Bioinformatics, Department of Biochemistry, Bio21 Institute, University of Melbourne, Melbourne, Victoria, Australia.

[7] Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia

[8] Department of Biochemistry, University of Cambridge, Cambridge, UK

Email of Corresponding author: maria.monticelli@unina.it

Abstract

A *CCNB3* (cyclin-B3) missense mutation is here described as causative of recurrent pregnancy loss (RPL) in homozygosity. CCNB3 is a cyclin with unknown structure and partner-kinase interaction, apparently possible both with CDK1 and CDK2. The complex clinical case required the use of structural bioinformatics to unravel the apparent paradox of a mutation in a non-conserved amino acid. CCNB3 model was built using SWISS-MODEL and its interaction with the kinases CDK1 or CDK2 modelled with pyDock and investigated with PISA. V1251D mutation, identified by whole-exome sequencing (WES), was investigated using DynaMut, which revealed its destabilizing effect on the protein.

Valine 1251 is not conserved among mammals but frequently replaced by a threonine, the two amino acids belonging to different classes. DynaMut analysis showed the common role of Val and Thr in the building of a hydrophobic interaction with tyrosine 1192, underlying the isostericity of the apparently different amino acids. Aspartic acid replacement of valine in V1251D disrupts this hydrophobic interaction, destabilizing CCNB3 and probably also reducing its affinity for the kinases.

We strongly believe this case study can be illustrative of the importance of structural biology in support of the exome analysis for diagnostic purposes.

Introduction

Recurrent pregnancy loss (RPL) affects up to 5% of clinical pregnancies (1) and triploidy accounts for about 13% of these cases (2). The causes of RPL can be different and in certain cases depend on genetic variants present in the mother in genes expressed in the oocytes. In this paper Here, we describe a case due to a variation in the gene *CCNB3* that encodes cyclin-B3 (CCNB3). Two Iranian sisters with RPL who were born from a consanguineous marriage between first cousins were affected by RPL. Their case was analysed by whole-exome sequencing, alignment and variant calling by standard and a homozygous missense variant g.50346749T>A (p.V1251D) (ChrX, GRCh38/hg38; SCV000886503) in exon 10 of *CCNB3* was identified (21).

Cyclins are regulatory subunits that bind and activate their catalytic partner serine-threonine kinases (cyclin-dependent kinases, CDKs). They play an important role in male and female meiosis (3). The partner kinase of human CCNB3 is unknown (4). The structure of cyclin-B3 is yet unknown. To validate the role of the missense variation V1251D in CCNB3 as the cause of RPL, we carried an extensive analysis in silico, including protein modelling and protein docking. We identified the critical role of the methyl group of Valine 1251 in the stabilization of human CCNB3.

Methods

The effect of the variants was predicted using the sequence-based tool PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2) (5), whose scores correlate with the residual activity of the protein affected by the mutation (6), Sift (7) and Mutation Taster (8).

The alignment of orthologous cyclin B3 sequences was obtained by retrieving the sequences from the KEGG databank (ORTHOLOGY: K21771) and aligning them using Clustal Omega (9).

Homology modelling was carried out to obtain a 3D structural model of CCNB3. The isoform 1 of CCNB3 sequence was retrieved from the UniProt database (UniProt ID: Q8WWL7). A 3D-model of CCNB3 was built with SWISS-MODEL (10) using as templates 6gu2 (11) or 2jgz (12) (chain B). The complexes of CCNB3 cyclin domain with CDK1 or CDK2 were built with the suite docking programs called pyDock (13) using as a receptor the structure of the kinase deposited in 6gu2 (chain A) (11) or 2jgz (chain A) (12). No spatial or biological restrictions were used during simulations, which allowed a complete sampling of the docking landscape around the kinase. The interactions between CCNB3 cyclin domain-CDK1 or CCNB3 cyclin domain-CDK2 with the lowest energy models obtained with pyDOCK were investigated by the server PISA.

The effect of V1251D substitution on CCNB3 was determined with DynaMut (14), a suite that runs Bio3D (15), ENCoM (16), mCSM (17), SDM (18), and DUET (19).

Results and Discussion

The missense mutation observed in the patients results in the replacement of a hydrophobic amino acid with a negatively-charged one. The variant V1251D in human CCNB3 is rare and is not reported gnomAD and dbSNP databases. SIFT classifies it as deleterious, PolyPhen2 as probably damaging and Mutation Taster as disease. V1251 is not conserved and is frequently substituted by Thr in the mouse and other mammals.

To evaluate the impact of V1251D on the function and/or stability of CCNB3 we modelled the human protein because its structure has not been obtained experimentally yet. Approximately 1000 amino acid residues at the N-terminus are predicted as disordered regions and only the destruction box (residues 60-68), and cyclin boxes (1132-1257 and 1259-1375) can be aligned to the other cyclins. Residues 1126-1388 of

CCNB3 were aligned to G2/mitotic-specific cyclin-B1 (CCNB1). V1251 is found at the end of a long alpha-helix (residues 1238-1251), is relatively exposed to solvent but interacts via the methyl group on Cbeta carbon with Tyr1192 (Figure 1).

We modelled the structure of mouse CCNB3 and found that the interaction is conserved in the mouse protein, in which T1257 replaces human V1251, and Y1198 replaces human Y1192. In the mouse it is the methyl group on Cbeta that interacts with the aromatic ring exactly as observed in the wild type human CCNB3. Aspartic acid cannot play the same role because on Cbeta there is only a carboxylic group.

The importance of the interaction methyl-aromatic ring, conserved in mammals notwithstanding the change of the amino acid that requires two nucleotide changes, is confirmed by programs that assess the effect of mutations on protein stability.

DynaMut (14) predicts that the variant V1251D destabilizes and increases the vibrational entropy of human CCNB3 (Figure 1). According to DynaMut, the major changes in molecular flexibility in mutant CCNB3 occur in residues 1149-1157, 1190-1200, 1245-1257. Thus, the destabilizing effect of V1251D is directly exerted on the interacting Tyr 1192, and indirectly on the region 1149-1157. The destabilizing effect of V1251D on human CCNB3 was confirmed by SDM (18), and DUET (19).

The partner kinase of human cyclin B3 in vivo is unknown (4). The modelled CCNB3 was docked onto the structures of human CDK1 (11) or CDK2 (12). Several low energy poses clusters and the interaction of human CCNB3 with the kinases closely resembles those of CCNB1 with the kinases. V1251 is located between two regions (L1249-N1250 and K1253-I1258) that are in contact with the kinases.

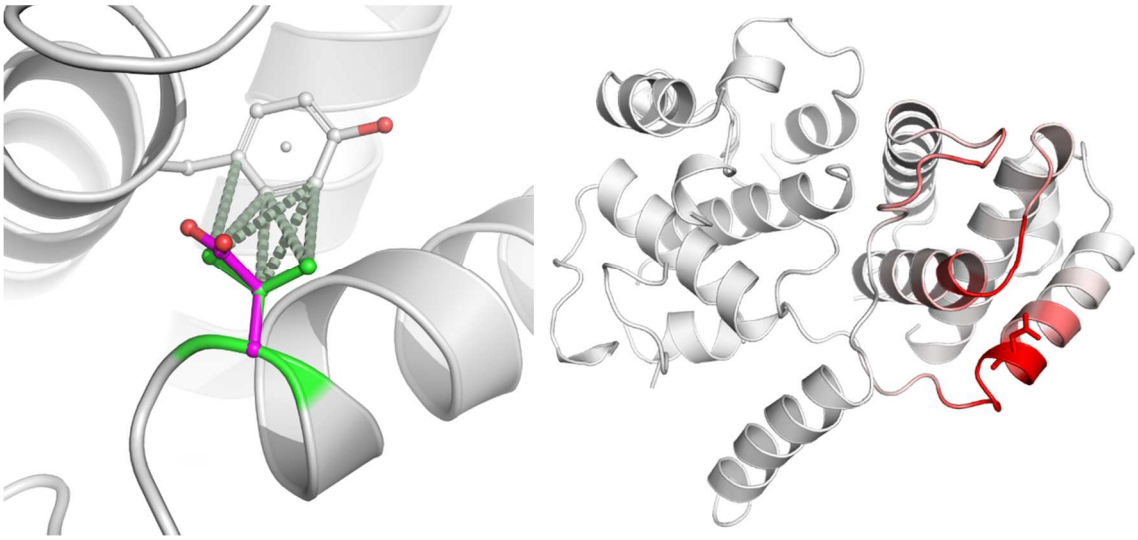


Figure 1: Intramolecular interactions made by V1251 (green) (left). Hydrophobic interactions, calculated using Arpeggio (20), are shown in green dashed lines. The introduction of an Asp (magenta) will alter these interactions, and lead to an increase in molecular flexibility in the region to accommodate the larger charged residue (right), with the cartoon coloured by the change in vibrational energy between the wildtype and mutant structures from blue (rigidification of the structure) to red (gain in flexibility).

Conclusion

Our experiments aimed to prove the causality of a variant in a human pathology. The case was particularly difficult because valine that is present in the wild type protein, is not conserved in mammals, but is often substituted by a threonine. According to the most popular classifications of the amino acids valine and threonine are not “similar” amino acids. Hence the doubt about the deleterious effect of a substitution on valine by another polar amino acid. The case required the analysis of the structure, which unfortunately was not available. Homology modelling and protein docking were instrumental to prove that V1251D can destabilize human CCNB3, thus reducing its intracellular concentration. In addition to this, the mutation might affect the affinity of the cyclin with kinases. We believe that this example proves the role of structural bioinformatics in the very “hot” field of exome analysis for diagnostic purposes.

Acknowledgements

This work was supported in part by a grant from Royan Institute, Iran (MT), Italian MIUR-PRIN 2015 JHLY35 (AR and MVC); Telethon-Italia GGP15131 (AR), Associazione Italiana Ricerca sul Cancro IG 2016 N.18671 (AR); “Progetti per la ricerca oncologica della Regione Campania” Grant: I-Cure (AR) and “Progetti competitivi intra-Ateneo” Programma VALERE (VAnviteLli pEr la RicErca) 2019 Grant: MIRIAM, Università degli studi della Campania “Luigi Vanvitelli” (AR and AS). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Sierra S, Stephenson M. Genetics of recurrent pregnancy loss. Seminars in Reproductive Medicine. 2006.

2. Soler A, Morales C, Mademont-Soler I, Margarit E, Borrell A, Borobio V, et al. Overview of Chromosome Abnormalities in First Trimester Miscarriages: A Series of 1,011 Consecutive Chorionic Villi Sample Karyotypes. Cytogenet Genome Res. 2017;

3. Wolgemuth DJ. Function of cyclins in regulating the mitotic and meiotic cell cycles in male germ cells. Cell Cycle. 2008.

4. Bouftas N, Wassmann K. Cycling through mammalian meiosis: B-type cyclins in oocytes. Cell Cycle. 2019.

5. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature Methods. 2010.

6. Cimmaruta C, Citro V, Andreotti G, Liguori L, Cubellis MV, Hay Mele B. Challenging popular tools for the annotation of genetic variations with a real case, pathogenic mutations of lysosomal alpha-galactosidase. BMC Bioinformatics. 2018;

7. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;

8. Schwarz JM, Cooper DN, Schuelke M, Seelow D. Mutationtaster2: Mutation prediction for the deep-sequencing age. Nature Methods. 2014.

9. Sievers F, Higgins DG. Clustal omega, accurate alignment of very large numbers of sequences.

Methods Mol Biol. 2014;

10. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: Homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;

11. Wood DJ, Korolchuk S, Tatum NJ, Wang LZ, Endicott JA, Noble MEM, et al. Differences in the Conformational Energy Landscape of CDK1 and CDK2 Suggest a Mechanism for Achieving Selective CDK Inhibition. Cell Chem Biol. 2019;

12. Brown NR, Lowe ED, Petri E, Skamnaki V, Antrobus R, Johnson LN. Cyclin B and cyclin A confer different substrate recognition properties on CDK2. Cell Cycle. 2007;

13. Cheng TMK, Blundell TL, Fernandez-Recio J. PyDock: Electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins Struct Funct Genet. 2007;

14. Rodrigues CHM, Pires DEV, Ascher DB. DynaMut: Predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res. 2018;

15. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. Bio3d: An R package for the comparative analysis of protein structures. Bioinformatics. 2006;

16. Frappier V, Najmanovich RJ. A Coarse-Grained Elastic Network Atom Contact Model and Its Use in the Simulation of Protein Dynamics and the Prediction of the Effect of Mutations. PLoS Comput Biol. 2014;

17. Pires DEV, Ascher DB, Blundell TL. MCSM: Predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics. 2014;

18. Pandurangan AP, Ochoa-Montaño B, Ascher DB, Blundell TL. SDM: A server for predicting effects of mutations on protein stability. Nucleic Acids Res. 2017;

19. Pires DEV, Ascher DB, Blundell TL. DUET: A server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res. 2014;

20. Jubb HC, Higueruelo AP, Ochoa-Montaño B, Pitt WR, Ascher DB, Blundell TL. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. J Mol Biol. 2017;

21. Fatemi N, Salehi N, Pignata L, Palumbo P, Cubellis MV, Ramezanali F, Ray PF, Varkiani M, Reyhani-Sabet F, Biglari A, Sparago A, Acurzio B, Palumbo O, Carella M, Riccio A, Totonchi M. Biallelic variant in cyclin B3 is associated with failure of maternal meiosis II and recurrent digynic triploidy. Journal of Medical Genetics. 2020.

Stratification of Systemic Lupus Erythematosus Patients with Gene Expression Data Reveals Expression of Distinct Immune Pathways

Aditi Deokar[1]
[1]Boston University Academy
Email of Corresponding author: adeokar@bu.edu

Abstract

Systemic lupus erythematosus (SLE) is the tenth leading cause of death in females 15-24 years old in the US. The diversity of symptoms and immune pathways expressed in SLE patients causes difficulties in treating SLE as well as in new clinical trials. This study used unsupervised learning on gene expression data from adult SLE patients to separate patients into clusters. The dimensionality of the gene expression data was reduced by three separate methods (PCA, UMAP, and a simple linear autoencoder) and the results from each of these methods were used to separate patients into six clusters with k-means clustering. The clusters revealed three separate immune pathways in the SLE patients that caused SLE. These pathways were: (1) high interferon levels, (2) high autoantibody levels, and (3) dysregulation of the mitochondrial apoptosis pathway. Mitochondrial apoptosis has not been investigated before to our knowledge as a standalone cause of SLE, independent of autoantibody production, and mitochondrial proteins could be investigated as a therapeutic target for SLE in the future.

Introduction

Systemic lupus erythematosus (SLE) is the tenth most common cause of death among females 15-24 years old in the US (1). SLE is one of many autoimmune diseases, which are diseases in which a patient's immune system mistakes parts of their own body as foreign, attacking their body instead of protecting it. In SLE specifically, the immune system attacks healthy organs and tissue. Patients can manifest very different symptoms including fatigue, swelling in the joints, the characteristic lupus butterfly-shaped rash, sunlight sensitivity, and many more (2).

SLE can be driven by defects in the innate immune system and/or the adaptive immune system. SLE patients are often characterized by high levels of interferon-1, which causes inflammation in the innate immune system in response to viruses. In SLE, high interferon levels can be caused by a variety of factors, including autoantibody complexes and neutrophil extracellular traps (3). Most SLE patients also have high levels of autoantibodies, which are created by mature B cells (plasma cells) that were not eliminated by the tolerance mechanisms that usually prevent maturation of self-reactive B cells (4). Autoantibodies cause a much more targeted adaptive immune response against specific self-antigens, but SLE patients can have a wide range of autoantibodies - one study found over 180 autoantibodies expressed in SLE patients (5). Some patients with lupus do not even have autoantibodies, and many of these autoantibodies are also found in other rheumatic diseases (6).

The heterogeneity of lupus symptoms and immune pathways affected makes it difficult to treat, because different drugs work well on different patients. This heterogeneity also causes difficulties in clinical trials of more targeted biologic drugs. Merrill et al. (7) found that certain standard drugs (anti-rheumatic drugs and immunosuppressants) affect immune pathways differently in interferon-low and interferon-high patients. While there is still debate on whether SLE is one disease or many (8), it is clear that subdividing SLE patients into categories will help treatment of patients and clinical trials.

Previous studies have tackled this problem by dividing patients based on antibody levels (9), gene expression (10), and immune molecule levels (11). However, none of these studies have reached a consensus on the best subdivision of SLE that holds true across multiple studies. Guthridge et al. (12) used all three of these factors to divide SLE patients into seven clusters with unsupervised machine learning. Using only Guthridge et al.'s gene expression data, we used different machine learning methods to create another set of clusters of SLE patients. These clusters were then used in comparison with Guthridge et al.'s

clusters to determine if gene expression data alone reveals similar patterns in immune pathway expression as does its combination with antibody levels and immune molecule levels.

Methods

We used a gene expression dataset available on GEO (accession number GSE138458) containing data collected by Guthridge et al. (12). The data included 336 samples in total, with 24 control patients and 198 SLE patients. 108 of the SLE patients had two or more samples taken. Data pre-normalized by Guthridge et al. (12) was used, which had gone through bgAdjust background correction, vst variance stabilizing transformation, and rank invariant normalization. Six outliers were removed by Guthridge et al. in the normalized data, including one control patient and five SLE patients. While Guthridge et al. used modular co-expression scores followed by unsupervised random forest clustering, then reduced the dissimilarity matrix from the random forest clustering into three principal components with t-SNE (t-Distributed Stochastic Neighbor Embedding), and selected the first two components as input for k-means clustering, we chose to instead use three separate dimensionality reduction techniques followed by k-means clustering.

Dimensionality Reduction

Prior to creating clusters from the gene expression data, the dimensionality of the 47,323 gene data was reduced using three separate methods: *Principal Component Analysis* (PCA), *Uniform Manifold Approximation and Projection* (UMAP), and a simple *autoencoder*. These techniques all reduced the 47,323 genes to fewer features in different ways to minimize the effects of random variation on the unsupervised clustering model.

PCA linearly transforms data to a lower-dimensional space in a way that will maximize the variance in the data. 200 principal components (the new features created by PCA that are linear combinations of the original genes) were selected for our model, which explain 96.29 % of the cumulative variance in the original 47,323 genes. The second technique we used, UMAP, is a nonlinear model (unlike PCA) which, similar to t-SNE, can be used for visualization, but is also used for nonlinear dimension reduction (13). For UMAP dimension reduction, we also reduced the 47,323 genes to 200 features. The third technique was a simple autoencoder. Autoencoders are neural networks which consist of two parts: an encoder, which reduces the dimensionality of the data, in this case from 47,323 genes to 1000 features, and a decoder, which expands these 1000 features back into 47,323 dimensions. In our study, we used the encoded data for later clustering similar to the 200 PCA and UMAP components. A simple autoencoder, used in this study, includes only one layer in the encoder and decoder. The autoencoder aims to reduce the loss of information between the original inputs (genes) and the decoded output of the same dimension by minimizing the loss function, in this case the mean squared error, and other hyperparameters such as the activation function. An autoencoder, like other neural networks, has an activation function which determines how the inputs of each node are converted to an output. Two of the most common activation functions are linear and sigmoid. We trained autoencoders with both of these activations and found that the linear autoencoder performed much better after 100 epochs (validation loss 0.068) than the sigmoid autoencoder (validation loss 48.28). The linear autoencoder was then run again for 50 epochs with no change in validation loss, and the 1000 encoded components were used for clustering.

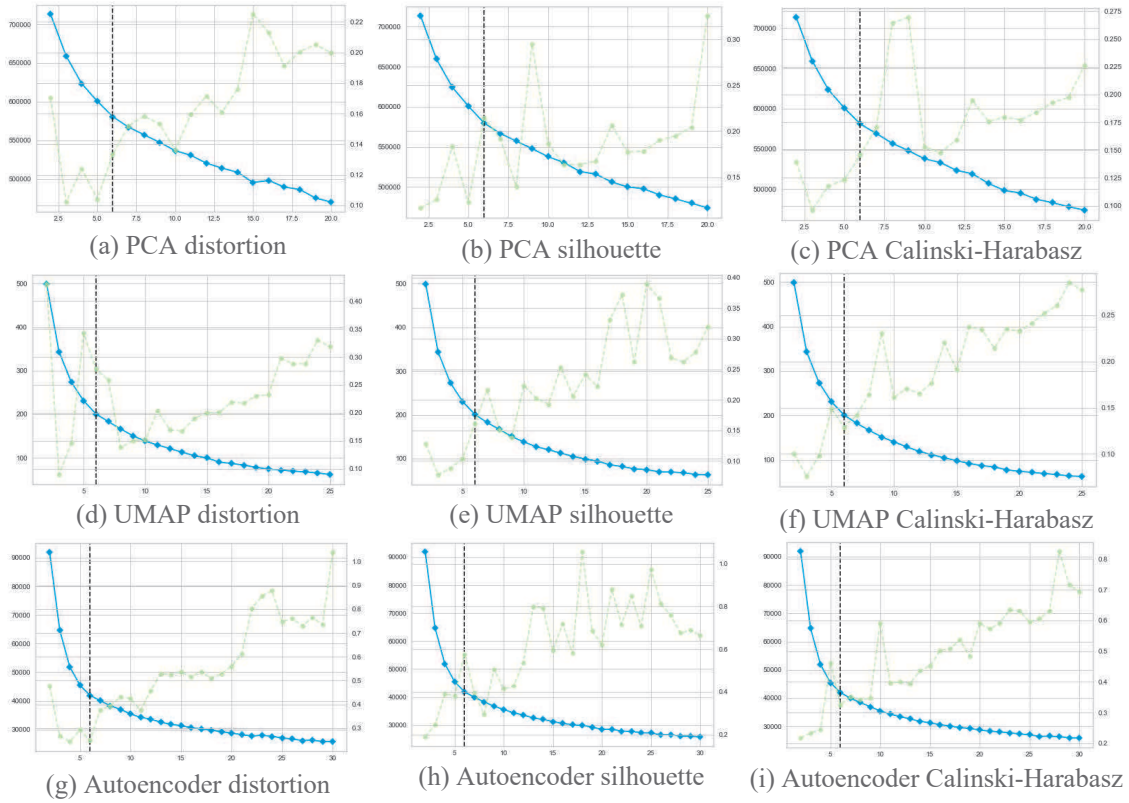
Clustering

The three sets of data with reduced dimensions (200 dimensions for PCA and UMAP and 1000 dimensions for the simple linear autoencoder) were then used for k-means clustering. In order to determine the appropriate number of clusters, we used the YellowBrick Python package, which creates visualizations quantifying the “elbow method” used on the metrics distortion score, silhouette score, and Calinski-Harabasz score (refer to Figure 1). All of these metrics, for all three dimensionality reduction methods, converged on 6 clusters, which were used for k-means clustering on those three reduced datasets.

Gene Modules

For visualization and interpretation of the clusters, we used 28 pre-existing modules created by Chaussabel et al. (14). One of these modules did not contain genes in our data, so 27 modules were used to calculate module scores for the datasets from each of the three dimensionality reduction techniques (PCA, UMAP, and simple linear autoencoder). Module scores for each cluster represented the percentage of genes in each

module that were significantly upregulated or downregulated by a two-tailed t-test ($p < 0.05$) in that cluster as compared to the controls.



(a) PCA distortion (b) PCA silhouette (c) PCA Calinski-Harabasz

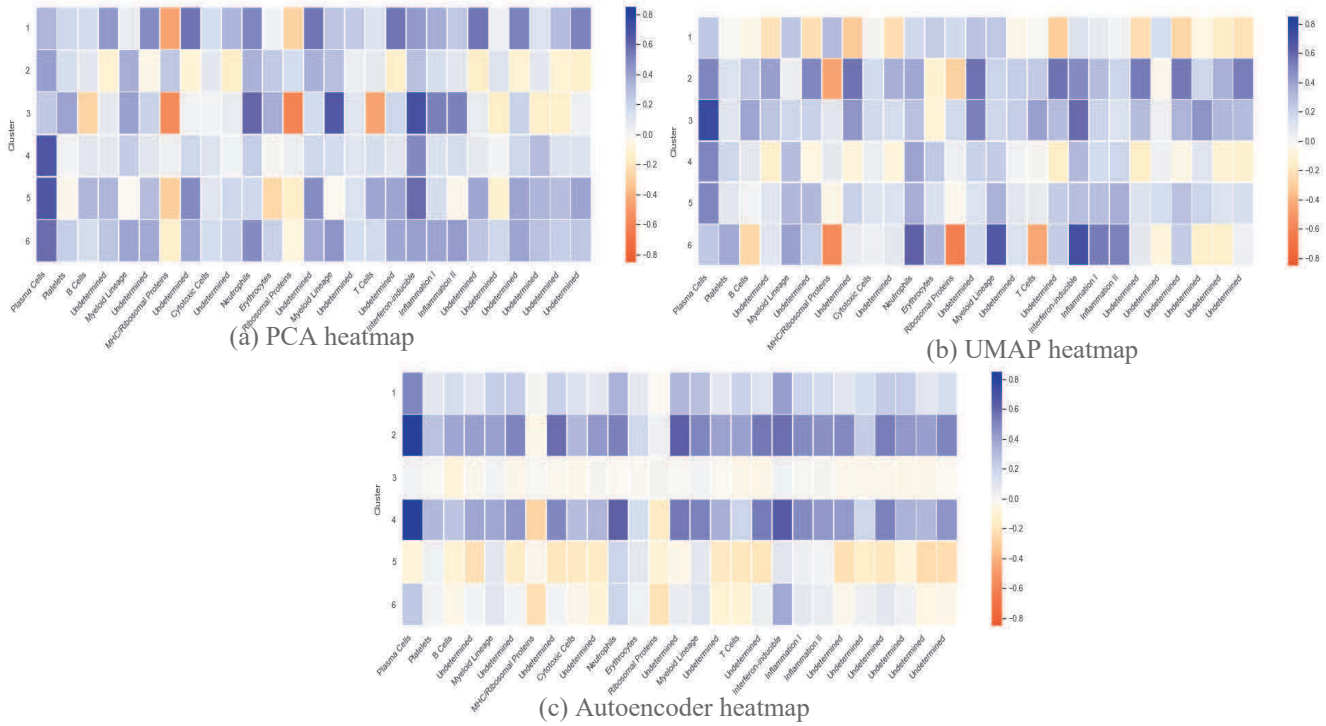
(d) UMAP distortion (e) UMAP silhouette (f) UMAP Calinski-Harabasz

(g) Autoencoder distortion (h) Autoencoder silhouette (i) Autoencoder Calinski-Harabasz

Figure 1: Metrics used to determine optimal number of clusters for k-means clustering

Results

Module scores were used to create heatmaps for data from each of the dimensionality reduction techniques (PCA, UMAP, and simple linear autoencoder) (Figure 2). In the heatmaps, the clusters originating from the PCA and UMAP dimensionality reductions showed very similar patterns in the upregulated and downregulated modules.



(a) PCA heatmap

(b) UMAP heatmap

(c) Autoencoder heatmap

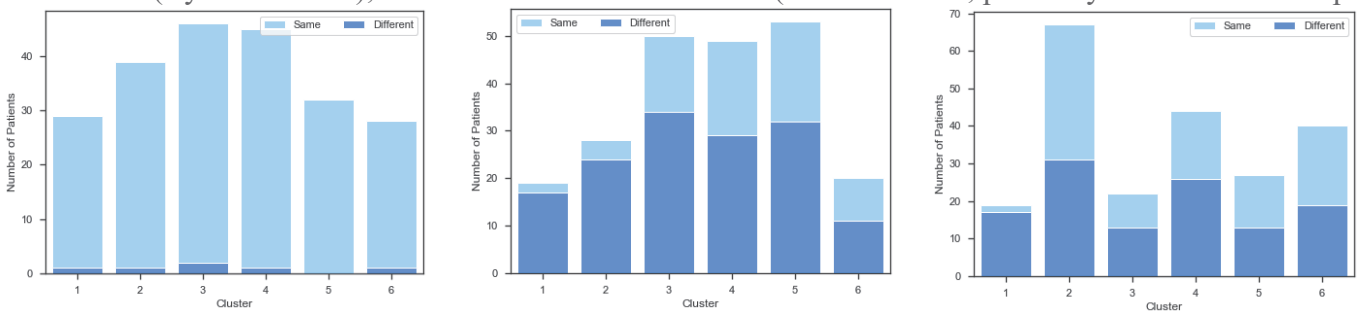
Figure 2: Heatmaps displaying the percentage of genes underexpressed (orange) or overexpressed (purple) as compared to the controls for each gene expression module (columns) in each cluster (rows).

The clusters originating from the autoencoder dimensionality reduction mostly showed a consistent level of increased or decreased gene expression across all modules. Autoencoder clusters 2 and 4 showed consistent substantial upregulations across almost all genes, except for MHC class 1 proteins and ribosomal proteins (the MHC genes in the module labels MHC/Ribosomal Proteins almost entirely consist of class 1 genes as described in Chaussabel et al. (14)), which were downregulated in cluster 4 and unchanged in cluster 2. Autoencoder clusters 1 and 6 both showed only upregulation of plasma cells and interferon-inducible genes; cluster 1 additionally had some upregulation of neutrophils and cluster 6 additionally had some downregulation of MHC class 1 proteins and ribosomal proteins. Autoencoder cluster 3 had no substantial gene expression differences at all as compared to the control group in the genes present in the modules. Autoencoder cluster 5 is described below alongside PCA clusters 1 and 4 and UMAP cluster 2.

MHC class 1 proteins and ribosomal proteins were downregulated in PCA clusters 2 and 6 and UMAP clusters 1, 3, 5, and 6. Plasma cells were substantially upregulated in PCA clusters 2, 3, 4 and 5 and UMAP clusters 4, 5, and 6. PCA cluster 5 and UMAP cluster 4 were very similar in their upregulation of plasma cells; UMAP cluster 4 also showed some upregulation of interferon-inducible genes. Interferon-inducible genes were upregulated in PCA clusters 2, 3 and 6 and UMAP clusters 3, 4, 5, and 6.

PCA cluster 6 and UMAP cluster 3 both displayed substantial upregulation of neutrophils, myeloid lineage proteins, interferon-inducible genes, and proteins related to inflammation, as well as substantial downregulation of MHC class I proteins, ribosomal proteins, B cells, and T cells.

PCA clusters 1 and 4, UMAP cluster 2, and Autoencoder cluster 5 displayed a different pattern from many of the other clusters, with many of the modules labeled as Undetermined underexpressed. These Undetermined modules represent modules which contained different types of genes and so did not have any clear uniting factor as in the other modules. The underexpressed genes in these modules (1.4, 1.6, 1.8, 2.9, 2.11, 3.4, 3.6, 3.8, 3.9) are listed in more detail in Table 1 of Chaussabel et al. (14), but included genes coding for components of the cAMP-signaling pathway, repressors of NF-KB activation, other signaling molecules, metabolic enzymes, other enzymes, proteins involved in DNA replication, kinases, phosphatases, RAS-family members, cytoskeleton-related molecules, T cell-expressed genes, and mitochondrial ribosomal proteins and elongation factors. Autoencoder cluster 5 patients additionally underexpressed genes from module 2.1 (Cytotoxic cells), and from modules 2.7 and 3.5 (Undetermined, primarily unknown transcripts).



(a) PCA clusters (b) UMAP clusters (c) Autoencoder clusters

Figure 3: Patients in each cluster, of those who had multiple samples taken, who were put in the same cluster or different cluster.

Since 108 of the SLE patients had two or more samples taken, so different samples from the same patient were included in clustering, we graphed in Figure 3 whether each of these 108 patients' samples were placed in the same cluster or different clusters for all of their samples. Each patient was counted once for each sample, so a patient who was placed in clusters 1 and 3 after PCA would be counted as "Different" in both cluster 1 and cluster 3 in the PCA graph, and a patient who was placed in cluster 4 after PCA for both samples taken would be counted as "Same" twice in cluster 4 in the PCA graph. In clustering following PCA dimensionality reduction most patients who had multiple samples taken were placed in different clusters, and in clustering following UMAP about half of these patients were placed in the same cluster and half in different clusters. In clustering following dimensionality reduction with the autoencoder, however, almost all patients who had multiple samples taken were placed in the same cluster for all their samples.

Discussion

The aim of this study was to use unsupervised machine learning techniques on gene expression data from SLE patients to stratify patients into clusters representative of their active molecular pathways. These

clusters could then be used in future clinical practice to identify the immune pathways responsible for a patient's SLE disease and prescribe treatment accordingly. We identified three sets of six clusters from gene expression data of adult SLE patients. These sets of clusters followed from three separate dimensionality reduction techniques used prior to clustering: PCA, UMAP, and a simple linear autoencoder.

For patients who had multiple samples taken, k-means following the autoencoder classified them into the same cluster 97.3 % of the time, while k-means following PCA and UMAP classified them into the same cluster 32.9 % and 45.7 % of the time respectively (Figure 3). While gene expression data is correlated with SLE disease activity (10,15), Petri et al. found that the majority of gene expression signatures were stable in patients over time (16). This suggests that the autoencoder's dimensionality reduction may have emphasized the stable gene expression signatures, causing them to be a major factor in the clustering, but that PCA and UMAP, which aimed to preserve more of the variance in the data, did not maintain the data from genes whose expression was stable over time. Many of these more stable genes might not have been related to the immune system, so they were not included in Chaussabel et al.'s coexpression modules (14). Thus, the more variable modules that were in the heatmap would have shown a lot of variation between the patients in each cluster, causing the clusters in the autoencoder heatmap to show a more consistent level of expression across all genes in the modules. Further analysis should be done to determine the level of variation in gene expression in the modules for the autoencoder clusters in comparison to the PCA and UMAP clusters, and to determine whether the PCA and UMAP clusters correlated to disease activity more than the autoencoder clusters, which this would imply.

The patients in the clusters created from the PCA and UMAP dimensionality reduction techniques and Autoencoder cluster 5 can be designated as belonging to one of three groups: interferon-driven SLE, antibody-driven SLE, and SLE caused by mitochondrial apoptosis.

Interferon-driven SLE

PCA cluster 6 and UMAP cluster 3 in Figure 2 both displayed substantial upregulation of interferon-inducible genes, proteins related to inflammation, neutrophils, and myeloid lineage proteins and downregulation of T cells and B cells. These two clusters correlate to Guthridge et al.'s clusters 1, 4 and 6, which also displayed these same patterns (12). All of the upregulated genes are related to the innate immune response. In lupus, type 1 interferon levels are often elevated, which can lead to inflammation and tissue damage (17). This pathway caused by elevated interferon levels is most likely the main cause of SLE in the patients in PCA cluster 6 and UMAP cluster 3 because of several reasons. Those patients' have overexpressed interferon-inducible genes, and also overexpressed myeloid lineage proteins (myeloid lineage cells include pDCs and neutrophils, both of which can cause interferon levels to increase). They also have underexpressed B and T cells and normal expression of plasma cells, which would all be overexpressed if dysregulation of self-reactive B cells was the main reason for autoimmunity, rather than interferon levels.

Antibody-driven SLE

Many of the other PCA and UMAP clusters displayed upregulation of plasma cells, indicative of increased antibody production; particularly PCA clusters 2, 3, 4 and 5 and UMAP clusters 4, 5, and 6. Guthridge et al. (12) observed a similar trend, where their clusters 2, 3, and 5 had higher T cell, B cell, and plasma cell related expression. While autoantibodies are known to be common in SLE, the diversity of autoantibodies (as discussed by Yaniv et al. in (5)) means that attributing the SLE of the patients in these clusters to autoantibodies in general is not enough, and there is still work to be done understanding what is different among the four PCA clusters and three UMAP clusters. Some of these differences might come from the genes used to create the clusters that were not included in the modules used for the heatmap.

Brant et al. (18), who grouped lupus patients based on their correlation between gene expression and disease activity, found three clusters, one where neutrophil levels correlated to disease activity, one where lymphocyte levels correlated to disease activity, and one more heterogeneous group. Since neutrophil extracellular traps are one way that interferon levels become elevated, their neutrophil-correlated group might correspond to our high-interferon group, and their lymphocyte-correlated group might correspond to our antibody-driven group. Once again, however, more analysis needs to be done on disease activity correlation in our data to confirm this.

SLE caused by mitochondrial apoptosis

As noted in the Results, PCA clusters 1 and 4, UMAP cluster 2, and Autoencoder cluster 5 displayed a different pattern from many of the other clusters. In these clusters, many of the modules labeled as

Undetermined were underexpressed. A closer look at the genes in these Undetermined modules reveals that they included mitochondrial ribosomal proteins, mitochondrial elongation factors, and proteins in the cAMP-signaling pathway. Mitochondrial ribosomal proteins (MRPS/MRPL), in addition to their ribosomal functions, are involved in apoptotic pathways (19), and cAMP signaling regulates mitochondrial apoptosis (20). Apoptosis is known to be a factor in SLE, but mainly because ineffective clearance of apoptotic cells can expose B and T cells to intracellular material, leading to the creation of autoantibodies against this intracellular material. SLE patients also tend to have higher rates of apoptosis generally (21).

We suggest that for the patients in PCA clusters 1 and 4, UMAP cluster 2, and Autoencoder cluster 5, dysregulation of mitochondrial pathways or signaling from outside molecules (possibly lymphocytes) could cause mitochondrial apoptotic pathways to become activated in healthy cells, destroying healthy cells as is characteristic of SLE. These healthy cells would have a range of gene expression of mitochondrial proteins, including MRPS/MRPL, and the cells with higher expression of the proteins would activate the apoptotic pathway. Only cells with lower expression levels would survive, so lower expression levels were found in our study. These lower expression levels would also impair mitochondrial functions, which has been observed to be true in SLE patient (22).

Guthridge et al.'s cluster 7 (12) also had low expression of mitochondrial respiration and mitochondrial stress/proteasome genes (which were not discussed in their study). The discovery of this cluster of patients using two completely different machine learning approaches corroborates the idea that the mitochondrial apoptotic pathway is a novel cause for SLE.

A note on MHC class 1 and ribosomal proteins

In most of the clusters, MHC class 1 proteins and ribosomal proteins were downregulated. Both of these proteins are important in autoimmunity. MHC class 1 proteins are usually produced by all nucleated, healthy cells, which use them to display self-antigens that indicate they are healthy. This is a regulatory measure that is meant to prevent autoreactive T cells from recognizing healthy cells, and the underexpression of MHC class 1 proteins blocks this regulation (23).

Ribosomal proteins, in addition to their transcriptional role also have a role in the innate immune system. Certain ribosomal proteins, namely RPL13A and RPS3, regulate pathways that mediate inflammation (24). Their underexpression in SLE patients would thus allow for uncontrolled inflammation.

Conclusion

In this study, we separated SLE patients into clusters based on their gene expression data using unsupervised learning. The data was collected by Guthridge et al. (12), who clustered patients using antibody levels and immune phenotyping in addition to gene expression levels. We used only gene expression data and used entirely different methods from their study, in order to determine whether we would find similar clusters of patients. The dimensionality of the gene expression data was first reduced by three separate methods (PCA, UMAP, and a simple linear autoencoder) and the results from each of these methods were used to separate patients into six clusters with k-means clustering. The clusters revealed three separate immune pathways in the SLE patients that caused SLE. These pathways were 1) high interferon levels, 2) high autoantibody levels, and 3) dysregulation of the mitochondrial apoptosis pathway. All three of these pathways were present in Guthridge et al.'s clusters (12), but to our knowledge this study is the first to propose mitochondrial apoptosis as a standalone cause of SLE, independent of autoantibody production. Future studies should investigate to a further extent the mitochondrial apoptotic pathway in SLE patients as a reason for destruction of self cells in addition to a way that autoantibodies are produced.

Acknowledgements

I would like to thank Anmol Warman for providing advice during this project.

References

1. Yen EY, Singh RR. Lupus – An Unrecognized Leading Cause of Death in Young Women: Population-based Study Using Nationwide Death Certificates, 2000–2015. Arthritis Rheumatol. 2018;70(8):1251–1255.

2. Lupus Foundation of America. What is lupus? [Internet]. 2020 [cited 2020 Jul 31]. Available from: https://www.lupus.org/resources/what-is-lupus
3. Bengtsson AA, Rönnblom L. Role of interferons in SLE. Best Pract Res Clin Rheumatol. 2017;31(3):415–28.
4. Dema B, Charles N. Autoantibodies in SLE: Specificities, Isotypes and Receptors. Antibodies. 2016;5(1):2.
5. Yaniv G, Twig G, Shor DBA, Furer A, Sherer Y, Mozes O, et al. A volcanic explosion of autoantibodies in systemic lupus erythematosus: A diversity of 180 different antibodies found in SLE patients. Autoimmun Rev [Internet]. 2015;14(1):75–9. Available from: http://dx.doi.org/10.1016/j.autrev.2014.10.003
6. Egner W. The use of laboratory tests in the diagnosis of SLE. J Clin Pathol. 2000;53(6):424–32.
7. Merrill JT, Immermann F, Whitley M, Zhou T, Hill A, O'Toole M, et al. The Biomarkers of Lupus Disease Study: A Bold Approach May Mitigate Interference of Background Immunosuppressants in Clinical Trials. Arthritis Rheumatol. 2017;69(6):1257–66.
8. Agmon-Levin N, Mosca M, Petri M, Shoenfeld Y. Systemic lupus erythematosus one disease or many? Autoimmun Rev [Internet]. 2012;11(8):593–5. Available from: http://dx.doi.org/10.1016/j.autrev.2011.10.020
9. Artim-Esen B, Çene E, Şahinkaya Y, Ertan S, Pehlivan Ö, Kamali S, et al. Cluster analysis of autoantibodies in 852 patients with systemic lupus erythematosus from a single center. J Rheumatol. 2014;41(7):1304–10.
10. Toro-Domínguez D, Martorell-Marugán J, Goldman D, Petri M, Carmona-Sáez P, Alarcón-Riquelme ME. Stratification of Systemic Lupus Erythematosus Patients Into Three Groups of Disease Activity Progression According to Longitudinal Gene Expression. Arthritis Rheumatol. 2018;70(12):2025–35.
11. Hamilton JA, Wu Q, Yang P, Luo B, Liu S, Li J, et al. Cutting Edge: Intracellular IFN-β and Distinct Type I IFN Expression Patterns in Circulating Systemic Lupus Erythematosus B Cells. J Immunol. 2018;201(8):2203–8.
12. Guthridge JM, Lu R, Tran LTH, Arriens C, Aberle T, Kamp S, et al. Adults with systemic lupus exhibit distinct molecular phenotypes in a cross-sectional study. EClinicalMedicine [Internet]. 2020;20:100291. Available from: https://doi.org/10.1016/j.eclinm.2020.100291
13. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018; Available from: http://arxiv.org/abs/1802.03426
14. Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, Baldwin N, et al. A Modular Analysis Framework for Blood Genomics Studies: Application to Systemic Lupus Erythematosus. Immunity. 2008;29(1):150–64.
15. Kegerreis B, Catalina MD, Bachali P, Geraci NS, Labonte AC, Zeng C, et al. Machine learning approaches to predict lupus disease activity from gene expression data. Sci Rep [Internet]. 2019;9(1):1–12. Available from: http://dx.doi.org/10.1038/s41598-019-45989-0
16. Petri M, Fu W, Ranger A, Allaire N, Cullen P, Magder LS, et al. Association between changes in gene signatures expression and disease activity among patients with systemic lupus erythematosus. BMC Med Genomics. 2019;12(1):1–9.
17. Crow MK. Type I Interferon in the Pathogenesis of Lupus. J Immunol. 2014;192(12):5459–68.
18. Brant EJ, Rietman EA, Klement GL, Cavaglia M, Tuszynski JA. Personalized therapy design for systemic lupus erythematosus based on the analysis of protein-protein interaction networks. PLoS One [Internet]. 2020;15(3):1–16. Available from: http://dx.doi.org/10.1371/journal.pone.0226883
19. Kim H-J, Maiti P, Barrientos A. Mitochondrial ribosomes in cancer. Semin Cancer Biol [Internet]. 2017 Dec;47(3):67–81. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1044579X17300962
20. Valsecchi F, Ramos-Espiritu LS, Buck J, Levin LR, Manfredi G. cAMP and mitochondria. Physiology. 2013;28(3):199–209.
21. Mevorach D. Systemic Lupus Erythematosus and Apoptosis. Clin Rev Allergy Immunol. 2003;25:49–59.
22. Leishangthem BD, Sharma A, Bhatnagar A. Role of altered mitochondria functions in the pathogenesis of systemic lupus erythematosus. Lupus. 2016;25(3):272–81.
23. McPhee CG, Sproule TJ, Shin D-M, Bubier JA, Schott WH, Steinbuck MP, et al. MHC Class I Family Proteins Retard Systemic Lupus Erythematosus Autoimmunity and B Cell Lymphomagenesis. J Immunol. 2011;187(9):4695–704.
24. Zhou X, Liao WJ, Liao JM, Liao P, Lu H. Ribosomal proteins: Functions beyond the ribosome. J Mol Cell Biol. 2015;7(2):92–104.

Molecular basis of cancer co-morbidities in COVID-19 patients

Antonio Facchiano [1], Francesco Facchiano [2], Angelo Facchiano [3]

[1] Istituto Dermopatico dell'Immacolata, IDI-IRCCS, Rome, Italy.

[2] Department of Oncology and Molecular Medicine, Istituto Superiore di Sanità, Rome, Italy.

[3] CNR - Istituto di Scienze dell'Alimentazione, Avellino, Italy

Email of Corresponding author: a.facchiano@idi.it

Abstract

To investigate comorbidities in COVID-19 patients, we analyzed the role of five human genes, known to encode coronavirus receptors/interactors (ACE2, TMPRSS2, CLEC4M, DPP4 and TMPRSS11D), in normal and cancer tissues. We integrated different tools and resources (i.e., DisGeNet, Genemania, DAVID, GEPIA2 and GENT2, Chilibot) to identify human diseases associated with these genes, finding relationships with the most frequent COVID-19 comorbidities, i.e. acute respiratory syndrome, cardiovascular diseases, diabetes, and cancer. In particular, their expression levels were found to be significantly altered ($P < 0.0001$) in colon, kidney, liver, testis, thyroid and skin cancers. These results suggest that three genes are relevant markers of kidney, liver, and thyroid cancer. Further investigation into their role is needed, in order to better understand molecular basis underlying comorbidities and follow-up of patients who have recovered from SARS-CoV-2 infection, and possibly to improve the recovery of COVID-19 patients.

Inhibition of angiotensin converting enzyme 2 SARS-CoV-2's receptor by natural compounds: *in silico* structure-activity relationship study

BEKKAL BRIKCI S[1], ABDELLI I [23], HASSANI F[1]

1: Ecology and Management of Naturals Ecosystems Laboratory, Department of Ecology and Environment-Faculty

SNV-STU- University- Tlemcen-Algeria

2: Higher School of Applied Sciences- Tlemcen-Algeria

3: Laboratory of Natural and bio-actives Substances, Faculty of Science- University- Tlemcen -Algeria

Email *: sohaybtlemcen@gmail.com*

Abstract

Essential oils (EOs) extracted from medicinal plants gained interest in research due to their potential effectiveness as antimicrobial compounds that can substitute chemical drugs for treatment of different disease. This study aims to use *Tetraclinis articulata* EO to block the activity of the angiotensin converting enzyme 2 (ACE2) as a receptor for SARS-CoV-2, his infection is triggered by binding of the spike protein of the virus to ACE2, which is highly expressed in the heart and lungs. Barbary thuja (*Tetraclinis articulata*), is an aromatic species that is essentially confined to the western Mediterranean region. It plays an important socioeconomic role in North Africa; it constitutes a pasture land for livestock and provides products for domestic use, and it widely used in traditional medicine for its multiple therapeutic virtues. This study reveals that some natural compounds extracted from *Tetraclinis articulata* give the best molecular docking scores, compared to the co-crystallized inhibitor β-D-mannose of the enzyme ACE2, to Chloroquine, and Hydroxychloroquine antiviral drugs also involved in other mechanisms as inhibition of ACE2 cellular receptor. A set of tests: the druglikeness property test, ADME/T test, PASS & P450 site of metabolism prediction, pharmacophore mapping and Molecular dynamics, were performed to determine the safety and efficacy of these ligands. This study revealed for the first time that the components of *Tetraclinis articulata* essential oils can be used as potential inhibitors to the ACE2 receptor of SARS-CoV-2.

Keywords : Essential oils, *Tetraclinis articulata*, SARS-CoV-2, Molecular docking.

MALVIRUS: an integrated web application for viral variant calling

Simone Ciccolella[1,†], Luca Denti[1,2,†], Paola Bonizzoni[1], Gianluca Della Vedova[1],
Yuri Pirola[1,*], Marco Previtali[1,*]

[1] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy

[2] Institut Pasteur, C3BI - USR 3756, 25-28 rue du Docteur Roux, Paris, France

[†] Joint First Authors

[*] Joint Last Authors

Email of Corresponding author: yuri.pirola@unimib.it

Abstract

Being able to efficiently call variants from the increasing amount of sequencing data daily produced from multiple viral strains is of the utmost importance, as demonstrated during the COVID-19 pandemic, in order to track the spread of the viral strains across the globe.

We present MALVIRUS, an accurate, easy-to-install, and easy-to-use web application that assists users in (1) computing a variant catalog consisting in a set of population SNP loci from the population sequences and (2) efficiently calling variants of the catalog from a read sample.

Tests on Illumina and Nanopore samples prove the efficiency and the effectiveness of MALVIRUS in genotyping SARS-CoV-2 strain samples with respect to GISAID data.

Introduction

The SARS-CoV-2 pandemic has put the global health care services to the test and many researchers are racing to face its swift and rapid spread. Since the outbreak of the virus in China and in other European countries, several studies are using sequencing technologies to track the geographical origin of SARS-Cov-2 and to analyze the evolution of sequence variants (1, 2). In this context, the availability of efficient approaches to analyze variations from the growing amount of sequencing data daily produced is of the utmost importance.

The typical pipelines for the analysis of variations within viral samples consists in aligning reads against a reference genome (3), then analyzing the alignments to discover the variants (4, 5). However, the increasing number of viral assemblies available in public databases such as GISAID (6), GenBank (7), and the COVID-19 Data Portal allows to build a complete catalog of variants of a viral population. Such a catalog can be used to reduce the complexity of comparative analysis of genetic variants of sequencing samples. Clearly, to this aim, it is crucial that users are assisted by an efficient and easy-to-use method for building and updating the catalog and for calling variants that are in this catalog. In this paper, we introduce MALVIRUS, a web application for quickly genotype newly sequenced viral strains, including but not limited to the SARS-CoV-2 strains. The application is distributed as a multi-platform Docker container and it can be easily accessed using any modern Internet browser. As use case, we show that MALVIRUS is accurate at genotyping newly sequenced SARS-CoV-2 strains on both short and long read data.

Methods

To efficiently genotype a viral sample from an individual with respect to the current knowledge, we propose MALVIRUS a web application based on five state-of-the-art tools.

The application is divided into two logically distinct modules: the creation of the catalog containing the SNP loci of the viral species under investigation and the variant calling from the read sample. Fig. 1 shows the MALVIRUS pipeline.

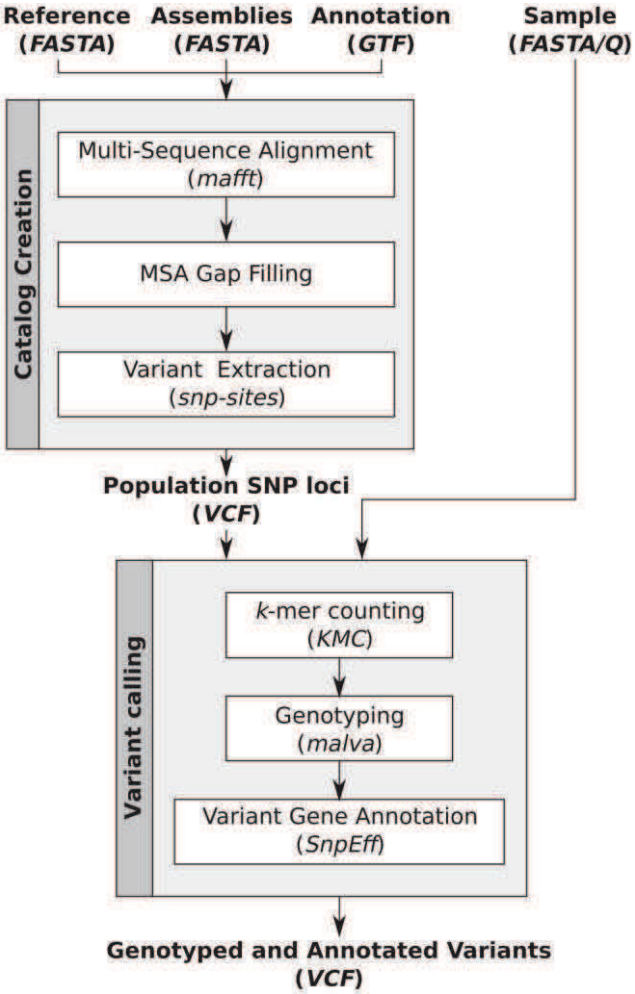


Figure 1. Schematic representation of the pipeline integrated in MALVIRUS.

The first module requires as input the reference genome of the species under investigation, the assemblies of a set of strains of that species, and, if available, the annotation of the genes. The output of this module is the set of population SNP loci in VCF format. MALVIRUS first builds the full-length sequence alignment of the input sequences to the input reference genome using MAFFT (8), then extracts the set of population SNP loci from the multiple alignment using snp-sites (9). Since snp-sites is not able to output variants in positions with gaps, MALVIRUS fills the gaps in the alignment with the corresponding portions of the reference. Although this step might induce some artificial variants, it allows to preserve real ones that might be lost due to incomplete assemblies. If the population under investigation is well characterized and/or the user wants a finer control over the variant catalog, it is possible to upload a custom catalog of SNP loci in VCF format instead of relying on the automatic computation from a set of assemblies.

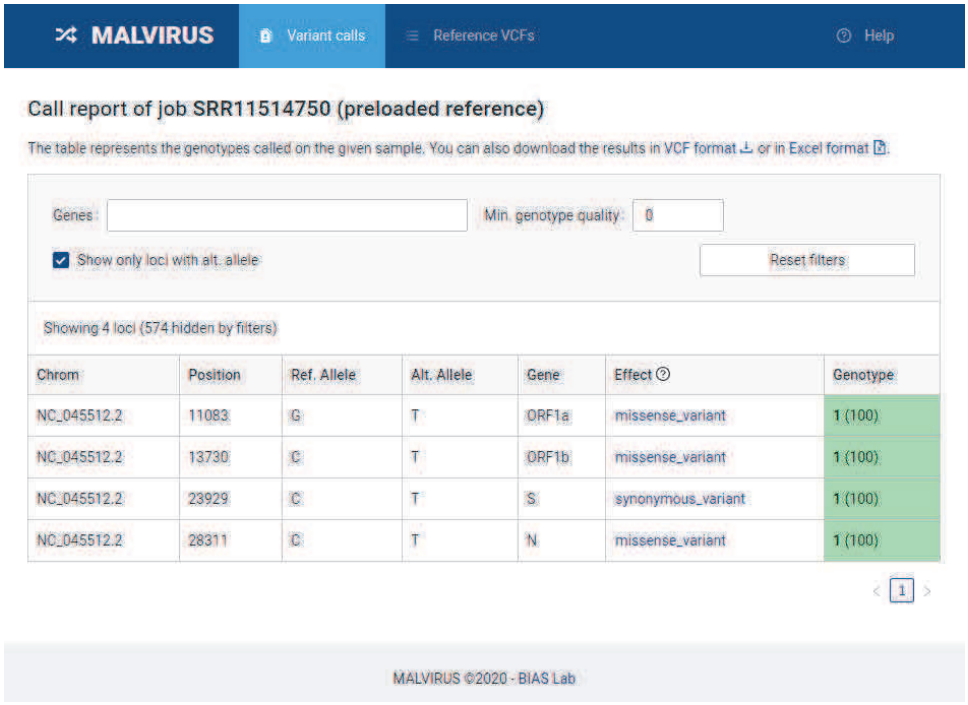


Figure 2. Example of the final report of MALVIRUS.

The second module requires as input a sample of reads in FASTA/Q format and a catalog of population SNP loci along with the corresponding reference genome chosen among the ones computed or uploaded in the first module. The output of the second module is a VCF containing the genotype information of the sample and their qualities. To call the genotype of each variant, this module counts the *k*-mers in the sample using KMC3 (10), then it genotypes the variants using MALVA (11): an efficient and accurate mapping-free approach for genotyping a set of known SNPs and indels initially developed for human individuals. We extended MALVA to support haploid organisms and high-coverage samples. Additionally, if gene annotation is available, the module also annotates the functional effects of each variant using SnpEff (12). Finally, the results of each analysis can be visualized as a table (see Fig. 2 for an example) or downloaded in VCF format or as a spreadsheet for further analysis.

MALVIRUS is available as a self-hosted web application distributed as a Docker container image that can be installed and run on multiple platforms, from personal laptops to large cloud infrastructures. For user convenience, the application is distributed with a set of precomputed catalogs of variants for SARS-CoV-2 based on the assemblies available on GenBank (7), therefore the user can immediately run MALVIRUS on a locally available (*e.g.*, private) viral sample. The precomputed catalogs can be easily updated from the application itself with a single click.

Extensive documentation and a detailed tutorial are available at https://algolab.github.io/MALVIRUS.

Results

To test the effectiveness of MALVIRUS, we considered 10 strains from the GISAID database for which a sample of raw reads is available on the Sequence Read Archive (SRA). These were the only samples that we were able to cross-reference between GISAID and SRA at the moment of writing, furthermore for 5 of such strains we analyzed reads sequenced using both Illumina and Oxford Nanopore technologies, showing that MALVIRUS achieves similar results on both data types.

For simulating a real case scenario, where the goal is to genotype a newly-sequenced strain, before analyzing a sample, we removed it from the set of complete SARS-CoV-2 strains available on GISAID (accessed on July 17, 2020) and we ran MALVIRUS on the remaining 42709 strains for building the variant catalog. From the 42709 strains, the first module of MALVIRUS produced a VCF containing 13709/13710 variants (depending on which strains were removed). Then, we genotyped such a catalog using the second module of MALVIRUS starting from the corresponding read samples.

To evaluate the overall accuracy of MALVIRUS we computed its precision and recall in genotyping the set of known variants produced by its first module. To compute precision and recall, we used the first module of MALVIRUS to build the variant catalog with respect to the considered strain (*i.e.*, the strain we removed) and we used it as truth set. We then classified each variant as a *reference variant* if its real genotype is 0, *i.e.* the reference allele, and as an *alternate variant* if its genotype is not 0. Finally, we compute the precision and recall of MALVIRUS and reported the results of this analysis in Table 1.

MALVIRUS scored a perfect precision (100%) on both reference and alternate alleles, while recall on the reference is almost perfect (99.9-100%) with some loss of recall on the alternate alleles. This loss of recall on the alternate alleles is caused by the fact that, especially on ONT data, some SNPs exhibit an unexpected and extremely low coverage that together with the high error rate makes them harder to correctly genotype. A careful inspection of these cases showed that a different choice of parameters (especially the *k*-mer size) improves its accuracy, allowing it to correctly genotype most of these low-covered SNPs at the cost of slightly lower precision. However, we believe that the default parameters of MALVIRUS allow to achieve the best trade-off between precision and recall. Finally, a single SNP (5508:T>C) is unique to the specific strain considered (GISAID ID *EPI_ISL_416410*) and cannot be present in the variant catalog built by the first module of MALVIRUS. Therefore, that variant could not be genotyped by the second module of MALVIRUS. However, since the rapidly increasing number of available complete sequences will broaden the variant catalog, we can expect that this situation will be uncommon in the next few months. On the other hand, such an increasing amount of data does not significantly challenge MALVIRUS since each step of the pipeline is efficient.

We ran MALVIRUS using 8 threads and the analysis of each sample completed in 50/60 minutes requiring less than 7GB of RAM. Such amount of resources is nowadays available on any computer, allowing MALVIRUS to run even on laptops and desktop machines. The first module of our application (catalog creation) required less than 15 minutes and less than 12GB of RAM. Anyway, we point out that the catalog creation needs to be run only when new strains are available, that each catalog can be reused multiple times, and that the software is distributed with a precomputed variant catalog built using the sequences available on NCBI.

Table 1. Results on real data. For each considered strain (GISAID ID, for ease of presentation we removed the EPI_ISL_ prefix) and the corresponding SRA sample, we report the Precision and Recall obtained by MALVIRUS on calling reference variants (i.e., those variants whose real genotype is the reference allele, REF) and alternate variants (i.e., those variants whose real genotype is the alternate allele, ALT). For each sample, we also report the technology used (ONT for Oxford Nanopore and ILLU for Illumina) and its coverage (in terms of number of bases).

GISAID ID	SRA ID	Seq. Tech.	# of bases	Precision REF	Recall REF	Precision ALT	Recall ALT
416410	SRR11397727	ONT	331	1	0.999	1	0.5
416410	SRR11397730	ILLU	178	1	0.999	1	0.5
416411	SRR11397726	ONT	363	1	1	1	1
416411	SRR11397729	ILLU	109	1	1	1	1
416412	SRR11397721	ILLU	126	1	1	1	1
416412	SRR11397725	ONT	236	1	0.999	1	0.57
416413	SRR11397720	ILLU	82	1	1	1	1
416413	SRR11397724	ONT	221	1	0.999	1	0.83
416415	SRR11397718	ILLU	112	1	0.999	1	0.6
416415	SRR11397722	ONT	381	1	0.999	1	0.4
416514	SRR11397717	ILLU	89	1	1	1	1
416515	SRR11397716	ILLU	81	1	1	1	1
416516	SRR11397715	ILLU	96	1	1	1	1
430819	SRR11667145	ILLU	13	1	0.999	1	0.75
430820	SRR11667146	ILLU	25	1	0.999	1	0.714

Conclusions

In this work, we presented MALVIRUS, a web application for quickly genotyping viral strains. As shown by our tests, MALVIRUS is able to efficiently and accurately genotype a newly sequenced SARS-CoV-2 strain both from short (Illumina) and long (Oxford Nanopore) reads. Since MALVIRUS benefits from comprehensive variant catalogs, the constantly increasing number of available strains will broaden the completeness of the current variant knowledge, thus boosting the overall accuracy of our pipeline.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 872539.

References

1. F. Gudbjartsson *et al.*, "Spread of SARS-CoV-2 in the Icelandic population," New England Journal of Medicine, vol. 382, pp. 2302–2315, 2020.
2. M. Bohmer *et al.*, "Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: a case series," The Lancet, vol. 20, no. 8, pp. 920–928, 2020.
3. H. Li, "Minimap2: pairwise alignment for nucleotide sequences," Bioinformatics, vol. 34, no. 18, pp. 3094–3100, 2018.
4. McKenna *et al.*, "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," Genome research, vol. 20, no. 9, pp. 1297–1303, 2010.
5. Wilm *et al.*, "LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets," Nucleic acids research, vol. 40, no. 22, pp. 11 189–11 201, 2012.
6. Y. Shu and J. McCauley, "GISAID: Global initiative on sharing all influenza data–from vision to reality," Eurosurveillance, vol. 22, no. 13, p. 30494, 2017.
7. W. Sayers *et al.*, "GenBank," Nucleic Acids Research, vol. 48, no. D1, pp. D84–D86, 2019.
8. K. Katoh and D. Standley, "MAFFT multiple sequence alignment software version 7: improvements in performance and usability," Molecular biology and evolution, vol. 30, no. 4, pp. 772–780, 2013.
9. J. Page *et al.*, "SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments," Microbial genomics, vol. 2, no. 4, 2016.
10. M. Kokot, M. Długosz, and S. Deorowicz, "KMC 3: counting and manipulating k-mer statistics," Bioinformatics, vol. 33, no. 17, pp. 2759–2761, 2017.
11. L. Denti, M. Previtali, G. Bernardini, A. Schonhuth, and P. Bonizzoni, "MALVA: genotyping by Mapping-free ALlele detection of known VAriants," iScience, vol. 18, pp. 20–27, 2019.
12. Pablo Cingolani *et al*. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3," Fly, vol. 6, no. 2, pp. 80-92, 2012.

Population genomics analyses on pangenome graphs

Flavia Villani[1], Francesco Porto[2], Andrea Guarracino[3], Robert W. Williams[4], Pjotr Prins[4], Gianluca Della Vedova[2], Erik Garrison[5], Vincenza Colonna[1]

[1]National Research Council, Institute of Genetics and Biophysics Adriano Buzzati-Traverso, Napoli, Italy; [2]Department of Informatics, Systems, and Communication, University of Milano-Bicocca, Italy; [3]Centre for Molecular Bioinformatics, Department of Biology, University Of Rome Tor Vergata, Rome, Italy; [4]Department of Genetics, Genomics and Informatics, College of Medicine, UTHSC [5]University of California, Santa Cruz, US

Introduction

Population genomics is the study of the causes and the consequences of genetic variability within and among populations. Population genomics is based on the study of variable sites. The accuracy of the inferences made by population genomics analyses is strictly correlated to the amount of information on genetic variation. For this reason, the field of population genomics has been particularly active in the last ten years, due to the unprecedented availability of genomic sequences that made possible the identification of millions of novel genetic variants [1, 2, 3].

Nevertheless, most of the population genomics studies are based on genomic variants which are simple to detect like single nucleotide variants and and very few studies deal with complex structural variants so far. This is mostly due to the inability to have reliable data set of complex structural variants, a limitation that is now being tackled by the use of long-reads sequence technology and pangenomes.

Standard approaches in sequence analysis relate sequences to a single linear reference genome. Sequence fragments produced by NGS technologies are mapped and assembled against a reference genome, and genetic variants are identified through comparison with it. While this is an efficient way of processing sequence information, the approach has a fundamental problem; substantial differences from the reference sequence are hard to observe and describe. As pangenome we refer to the entire set of genomic elements in a given species or clade. Pangenomic methods allow us to overcome limitations of the use of the reference genomes, relating all genomes directly to each other; sequences and variations are combined [4]. In pangenome variation graphs, genetic variants appear as bubbles. These sites have a common starting context (a single inbound node), a common exit point (a single outbound node), and a diversity of possible paths that connect the two, each of which represents an allele [5]. We consider these bubbles in the context of a data model developed to represent the basic components of variation graphs, the handlegraph abstraction [6]. This data structure breaks down the elements of a variation graph and proposes a programming interface based on them. We use this data model as the basis for algorithms to find bubbles, their alleles, and the frequencies of these alleles among the genomes embedded in the graph.

Pangenomes are large, and unwieldy to work with as raw collections of sequences. One possible approach to processing them considers the collection of genomes and their mutual alignment in a compact, graphical model. As a lossless representation of the pangenome and its embedded sequence variations, these variation graphs should, at least in principle, support any kind of population genetic analysis that would be completed on simpler representations of the genomes and their variations. But, because this pangenomic approach is quite recent, the software for population genetic analyses currently available are still mostly based on genomic data in the linear format.

Here we respond to this need by implementing VGPOP, a set of tools for population genetics

on genome graphs. At a high level, our work has two parts. We first uncover genomic variation embedded in pangenomic variation graphs by developing and implementing straightforward algorithms for bubble detection on variation graphs. We then demonstrate the calculation of basic population genetic parameters over variation graphs.

Methods

Implementation of the VGPOPlibrary We developed a library named VGPOP to conduct standard population genetics analyses using pangenomic data models. Typically represented in the Graphical Fragment Assembly (GFA) format [7], these models can represent whole genome alignments in a compact graphical structure. The library is written in the Python programming language under MIT license; the code is publicly available on GitHub (`https://github.com/Flavia95/VGpop`). Currently VGPOP has three sets of functions detailed in the following paragraphs.

1. Functions for the identification of variable sites The first mandatory step for any further population genetics analysis is to extract from the graphs the information about variable sites, i.e. the regions where more that one type of sequence is present. Any population genetic analysis is indeed based on the information contained in the variable segments of the sequence and their occurrence in the population under investigation. Because of their appearance in the pangenome graph, variable sites are referred to as bubbles. We implemented two main functions for bubble detection, namely BUBBLEPOP and BUBBLECALL.

The **BUBBLEPOP** function takes as input a GFA file and gives as output a dictionary, i.e. a table of correspondences between region of the graph and sequence variants. It explores the graph using the two recursive algorithms, the Depth First Search (DFS) [8, 9] and the Breadth First Search (BFS) [10, 11].

In BUBBLEPOP, we run BFS on the tree obtained by DFS. Starting from the tree root, the DFS explores the tree until it finds a bubble, that is a pair of nodes whose distance from the root is the same. When this happens it calculates the distance from the root of all the nodes in the bubble.

At the beginning of the bubbles all paths share the same identical node, and this is true also at the end of the bubble.

Once the pangenome has been decomposed with BUBBLEPOP in a tree whose information on the node distance from the root is stored in dictionary, **BUBBLECALL** explicits the content of the bubbles and its position in relation to a chosen reference sequence in three steps:

1. Choosing the reference path - We consider all the possible paths that connect the initial node and the final node of each bubble, the first path in the GFA file is chosen as reference (REF)

2. Variant identification - In this step BUBBLECALL iterates over all available nodes to analyze paths traversing the node and compare them with the reference node. BUBBLECALL considers all the possible paths pairs(x,y) in which x is the REF and y is any other path. A node is called as a variant if:

    - (i) it is supported by at least one path;
    - (ii) the node sequence is different from the sequence of the corresponding reference node;
    - (iii) if its distance from the root is the same as the one of the reference node, than the variant is classified as a Single Nucleotide Variants (SNV);
    - (iv) if its distance from the root is smaller that the one of the reference node then the variant is classified as a Deletion;
    - (v) if its distance from the root is greater that the one of the reference node then the variant is classified as an Insertion.

3. Variant positioning - This step defines the position of a variants with respect to the reference sequence. When the two paths were used to call the variants, the length of the sequences was taken into account in order to map the variants on the individual paths.

Re-implementation in Rust Population genomics analysis requires the study of a large number of individuals of any species, therefore the pangenomic approach has to be implemented in a way that is applicable to graphs of any complexity. For this reason, we decided to re-implement the core functions of our library in Rust; this project is publicly available on GitHub at `https://github.com/HopedWall/rs-gfatovcf`. Rust is a programming language which allows us to build reliable and efficient programs when compared to other languages, such as Python, which was used for our original implementation.

In order to achieve scalability, the following changes were made:

1. we employed a non-recursive strategy for building the spanning tree, since the original procedure required an excessive amount of memory on large graphs. The Rust implementation, instead, uses a queue-based approach, which prevents this type of problem.

2. we introduced as a parameter the maximum amount of edges to traverse during the BUBBLECALL step. This is required since finding all paths between two given nodes is a problem which is known to be NP-hard, hence it may take exponential time. This change limits the running time but might result in missing some paths.

3. introduced the ability to set only specific paths as references, avoiding the variant identification with respect to all the paths in the graph. This should increase performances when simpler analyses are required.

2. Functions for format conversion

The **GFA2VCF** function of VGPOP takes as input a graph in the GFA format and outputs a corresponding linear representation in the VCF format, i.e. the file format that is currently used to store sequence information on variable sites. To do this *gfa2vcf* uses first BUBBLEPOP to decompose the pangenome in a tree and then BUBBLECALL to identify the variable sites. Finally the dictionary of the variable site is formatted according to the vcf specifications.

3. Functions for population genetics

GFA2ALLELEFREQ - The frequency of an allele is an indication of how common the allele is in a population. It is calculated by counting how many times the allele appears in the population, divided by the total number of copies of the gene. The code we developed for the GFA2ALLELEFREQ function of VGPOPtakes as input a graph in GFA format and a metadata file (with information on paths, individuals, and populations), and outputs a file that contains the allele frequencies for variable loci per each population. In GFA2ALLELEFREQ the allele frequency corresponds to the number of paths that support a node (i.e. a variant) divided by the total number of paths actually realized. The frequencies of monomorphic nodes (i.e. frequency = 1) are not reported. GFA2ALLELEFREQ first uses bubblepop and bubblecall to read the graph, and than applies calculation of frequencies.

GFA2FST - The Wright's fixation index (Fst) is a measure of population differentiation due to genetic structure [12]. It is estimated from genetic polymorphism data, such as SNV or microsatellites. Several formulae exists for its calculation among which the one that estimates it as the standardized variance of allele frequencies among sub-populations. The code we developed for the GFA2FST function of VGPOP, takes as input an allele frequencies file, and as output a file that contains the calculation of Fst. GFA2FST first uses BUBBLECALL and BUBBLEPOP to read the graph, and than applies GFA2ALLELEFREQ to calculate allele frequencies and than calculates F_{st} as the standardized variance of allele frequencies among subpopulations: F_{st}= s_2/p (1-p) with s_2 and p being the variance and mean, respectively, of the allele frequencies.

GFA2TAJIMASD - The test statistic developed by Tajima [13] allows to identify non-random evolution of DNA sequences and consists in the ratio between two estimate of the effective population size (i.e. a measure of genetic diversity [14]): the number of segregating sites and the nucleotide diversity. The code we developed for GFA2TAJIMASD takes as input a GFA and outputs the corresponding value of the test statistic.

Results

***Calculation of F_{st} on simulated data using* GFA2FST**. To test if the calculation made by VGPOPare accurate, we applied VGPOPfunctions to data for which we can predict ranges of expectations for the parameters calculated by VGPOP. In particular we used sequence data produced by simulation under a known demographic scenario of two populations separating from a common ancestral population to measure the degree of separation calculated as F_{st}.

As simulation scenario we considered a model adapted from [15] with two diploid populations separating without subsequent migration. The first population is bigger in size compared to the second, and through time develops maintaining constant size until 5k generations ago when it starts to exponentially expand. The second population develops through time maintaining constant size. We considered three possible scenarios for separation time: 5k (T1), 10k (T2), and 15k (T3) generations ago. The expectation is that the longer the separation time, the higher will be the F_{st}, with scenario T3 having the higher F_{st} compared to T2 and T1.

We used the software ms [15] to produce 100 replicates of simulated variable sites in a 10kb region for eighty individuals under each of the three scenario. The variable sites were transformed in sequences that include also the invariable part (using Seq-Gen [16]) and the sequences were used to reconstruct the pangenome of the simulated data that was then processed with the GFA2FST function of the VGPOPlibrary. F_{st} calculation was validated using vcftools REF, that uses a different F_{st} formula.

We found that the F_{st} trend vary according to expectation of the three simulated scenario, i.e. the lowest value is found at T1 and the highest at T3. We observe the same trend when calculating F_{st} with a different formula as a control. Nevertheless, the absolute values of F_{st} obtained from VGPOPare lower that those obtained from vcftools, suggesting that a further comparison would be required to fully clarify the discordance and improve the VGPOPlibrary.

Allele frequencies at variable loci of the human HLA region using GFA2ALLELEFREQ The HLA region is located on the short arm of chromosome 6 from 6p21.1 to p21.3 in a region spanning 7Mb. The class II region includes genes for the α and β chains of the MHC class II molecules HLA-DR, HLA-DP and HLA-DQ. In addition, the genes encoding the DMα and DM β chains, as well as the genes encoding the α and β chains of the DO molecule (DO α and DO β, respectively), are also located in the MHC class II region [17].

In the latest version of the human reference genome (GRCh38), there are alternate loci highly polymorphic where the sequence variation is too complex to be represented with a single sequence [18]. These loci are known to co-segregate with disease and are therefore of great interest in population genetics. Sequence reads alignment in the HLA region, is known to be particularly difficult, particularly in regions originating from highly polymorphic regions and regions absent from the reference genome.

We considered three genes of the HLA region, *HLA-E*, and *HLA-DMA*, and *HLA-C*. For these three genes we started from eleven (*HLA-DMA*), nine (*HLA-E*), and ten (HLA-C) sequences downloaded from GenBank. We used the sequences to reconstruct the pangenomes. The pangenomes of *HLA-E* and *HLA-DMA* are less complex compared to the pangenome of *HLA-C*, suggesting less diversity in these two genes compared to *HLA-C*. We used the pangenomes to detect of variable sites (bubbles) with BUBBLEPOP, and then the allele frequencies are calculated as the number of paths supporting the variant node divided by total number of paths using GFA2ALELLEFREQ.

Variant identification in Sars-CoV-2 using rs-gfatovcf

Since the main motivation for re-implementing GFA2VCF in Rust was the ability to use it on larger graphs, we considered the Sars-CoV-2 pangenome available at `http://covid19.genenetwork.org/` in GFA format. This pangenome is composed of sequences of approximately 1.2 GBytes and with 78571 fragments, obtained from 15127 genomes. rs-gfatovcf is capable of obtaining a VCF file from it in 16 minutes on a machine with 256GB RAM, we found 294626 variants. While this result is satisfactory, we want to exploit concurrent and parallel computing to reduce its running time.

Conclusions

We have presented the results of our project to develop VGPOP, a library for population genetic analyses based on pangenome graphs.

The use of pangenomes and variation graphs is one of the major changes in genomics. Because this approach is quite recent, there has been little focus on developing software for population genetic analyses from pangenomes, and in fact almost all the already available software is based on genomic data in the linear format. With our project we contributed to fill this gap by writing software for population genetic analyses able to deal with pangenomes.

Two function of VGPOP, BUBBLEPOP and BUBBLECALL, have the primary function to parse the pangenome and identify the variable sites (bubbles). These two functions are exploited by some of the others, like GFA2ALLELEFREQ and GFA2FST that instead produce population genetics summary statistics. Finally, other functions, e.g. GFA2VCF are utility to convert file formats. This set of functions does not cover all possible needs for population genetic analyses, but it shows that several types of function are required to cover all possible tasks.

We first tested VGPOP on simulated data where we could rely on known expectations. We demonstrated that with VGPOP we can reliably estimate genetic distance between a pair of populations in three scenarios of increasing genetic diversity, using as a measure of diversity F_{st}, one of the basic summary statistics in population genetics. We also demonstrated that VGPOPcan calculate allele frequencies in regions of the genome with complex genetic variability, such as the HLA region, a complex variable region due to the high degree of similarity and polymorphism of its genes. The range of complexity in the variability of the HLA region made it also possible to test the limitations of VGPOP. Finally, we focused on the Sars-CoV-2 pangenome, which we chose for its current international relevance. This pangenome also acts as a benchmark for what the Rust version of VGPOPcan do, as it targets bigger graphs, which would cause memory problems in the original Python implementation.

Overall, with our project we demonstrated that VGPOP can calculate the basic statistics for population genomics inference directly from pangenomes. VGPOP is able to process pangenomic data, therefore putatively access complex variants scantily considered so far in population genomics. Even if in its current form VGPOP is only effective with simple variants, it has the potential to be adapted also for more complex ones. To our knowledge, this is the first such exploration that has been undertaken in the scope of this representation. Our work suggests a series of follow-up studies to extend related population genetic metrics to pangenome models. We hope to explore the development of haplotype-based scans for genetic selection (e.g nSL [19], iHS [20], and xp-ehh [21]) to pangenome graphs, as well as other measures of frequency differentiation between populations could be applied to alleles in bubbles in the graph (e.g. PBS [22]).

We are also aware of the current limitations of VGPOP, namely (1) the inability to detect complex bubbles and (2) its overall running time on larger graphs. In order to address (1), we are looking into a new bubble detection algorithm [5]. In order to address (2), we plan on exploiting parallel computing, which we hope will drastically improve the the running time of our functions.

References

[1] . G. P. Consortium *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, 2015.

[2] N. A. Rosenberg, "Standardized subsets of the hgdp-ceph human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives," *Annals of human genetics*, vol. 70, no. 6, pp. 841–847, 2006.

[3] K. Karczewski and L. Francioli, "The genome aggregation database (gnomad)," *MacArthur Lab*, 2017.

[4] J. M. Eizenga, A. M. Novak, J. A. Sibbesen, S. Heumos, A. Ghaffaari, G. Hickey, X. Chang, J. D. Seaman, R. Rounthwaite, J. Ebler, *et al.*, "Pangenome graphs," *Annual Review of Genomics and Human Genetics*, vol. 21, 2020.

[5] B. Paten, J. M. Eizenga, Y. M. Rosen, A. M. Novak, E. Garrison, and G. Hickey, "Superbubbles, Ultrabubbles, and Cacti," *J. Comput. Biol.*, vol. 25, pp. 649–663, 07 2018.

[6] J. M. Eizenga, A. M. Novak, E. Kobayashi, F. Villani, C. Cisar, S. Heumos, G. Hickey, V. Colonna, B. Paten, and E. Garrison, "Efficient dynamic variation graphs," *Bioinformatics*, 2020.

[7] "Gfaformat."

[8] R. E. Korf, "Depth-first iterative-deepening: An optimal admissible tree search," *Artificial intelligence*, vol. 27, no. 1, pp. 97–109, 1985.

[9] Wikipedia contributors, "Depth-first search — Wikipedia, the free encyclopedia," 2020. [Online; accessed 12-June-2020].

[10] S. Beamer, K. Asanovic, and D. Patterson, "Direction-optimizing breadth-first search," in *SC'12: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, pp. 1–10, IEEE, 2012.

[11] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. MIT Press, 2nd ed., 2001.

[12] R. R. Hudson, "Generating samples under a wright–fisher neutral model of genetic variation," *Bioinformatics*, vol. 18, no. 2, pp. 337–338, 2002.

[13] F. Tajima, "Statistical method for testing the neutral mutation hypothesis by dna polymorphism.," *Genetics*, vol. 123, no. 3, pp. 585–595, 1989.

[14] D. L. Hartl and A. G. Clark, *Principles of population genetics*, vol. 116.

[15] R. R. Hudson, "ms a program for generating samples under neutral models," *Bioinformatics*, 2004.

[16] A. Rambaut and N. C. Grass, "Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees," *Bioinformatics*, vol. 13, no. 3, pp. 235–238, 1997.

[17] R. D. Campbell and J. Trowsdale, "Map of the human mhc," *Immunology today*, vol. 14, no. 7, pp. 349–352, 1993.

[18] H. P. Eggertsson, H. Jonsson, S. Kristmundsdottir, E. Hjartarson, B. Kehr, G. Masson, F. Zink, K. E. Hjorleifsson, A. Jonasdottir, A. Jonasdottir, *et al.*, "Graphtyper enables population-scale genotyping using pangenome graphs," *Nature genetics*, vol. 49, no. 11, p. 1654, 2017.

[19] A. Ferrer-Admetlla, M. Liang, T. Korneliussen, and R. Nielsen, "On detecting incomplete soft or hard selective sweeps using haplotype structure," *Molecular biology and evolution*, vol. 31, no. 5, pp. 1275–1291, 2014.

[20] B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard, "A map of recent positive selection in the human genome," *PLoS Biol*, vol. 4, no. 3, p. e72, 2006.

[21] P. C. Sabeti, P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll, R. Gaudet, *et al.*, "Genome-wide detection and characterization of positive selection in human populations," *Nature*, vol. 449, no. 7164, pp. 913–918, 2007.

[22] X. Yi, Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo, J. E. Pool, X. Xu, H. Jiang, N. Vinckenbosch, T. S. Korneliussen, *et al.*, "Sequencing of 50 human exomes reveals adaptation to high altitude," *Science*, vol. 329, no. 5987, pp. 75–78, 2010.

Assessing the performances of protein stability predictors

Anna Marabotti[1], Eugenio Del Prete[2], Bernardina Scafuri[1], Angelo Facchiano[3]

[1]: Dept. Chemistry and Biology “A. Zambelli”, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy.
[2]: CNR-IAC, National Research Council, Institute for Calculus Applications "Mauro Picone", Via P. Castellino, 111, 80131 Naples, Italy
[3]: CNR-ISA, National Research Council, Institute of Food Science, Via Roma 64, 83100 Avellino, Italy

.

amarabotti@unisa.it

angelo.facchiano@isa.cnr.it

Abstract

The prediction of the changes in protein’s thermodynamic stability caused by mutations is a hard task to perform and, despite decades of development of dedicated predictors, there are still doubts about their reliability. Moreover, also the creation of a reference database to assess their performances is difficult because of the paucity of high quality, reliable data. In this work we present the creation of a reliable dataset of proteins and the assessment we have made of five predictors available online, representing different approaches developed so far, describing also the problems we had to face and the solutions we have adopted. Results show that these tools are surely more reliable than the past, but still far from ideal, and that some main issues are still present, such as the bias towards the destabilizing mutations. In general, the binary interpretation “less/more stable” must be taken with caution when the predicted ΔΔG is within the interval ±0.5 kcal/mol. The combination of several predictions is a rough, but effective way to increase the reliability of the results.

Introduction

The knowledge of the impact of a mutation on the thermodynamic stability of a protein is crucial to understand how this mutation can perturb the structure-function-dynamics relationships of that protein, and to identify possible approaches to prevent or counteract this problem. However, it is often hard to perform in vitro testing of the protein’s destabilization following mutations, especially when hundreds of mutations are associated to a single protein, such as in the case of rare diseases [1]. Therefore, in the past decades, many methods to predict the effect of a mutation on the thermodynamics stability of a protein have been developed (reviewed in [2]). Several assessments questioned the reliability of these predictors (for a review, see [2]) because the risk of obtaining a wrong prediction using these tools appears to be high. Most of these assessments were performed on methods and tools that have been updated later, or on methods based on a single approach. Moreover, we noticed that the experimental data related to the impact of mutations on protein stability are often of low quality, with noise and errors, and the efforts made to improve these data [3], although very valuable, are still not sufficient to guarantee high data quality.

Therefore, the focus of this work was, on one side, to create a dataset including only high reliable data in terms not only of thermodynamic experiments, but also of structural data, and on the other side, to use this dataset to assess five popular, Web-accessible methods for predicting the impact of mutations on protein stability, representative of the different approaches used so far to develop predictors in this field.

Methods

Starting from VariBench dataset [3], we performed a further selection in order to include in our dataset only high-quality reference proteins, in terms of both thermodynamic experimental data and quality of structures

associated to them. Since the data available are also biased in terms of quantity of destabilizing mutations towards stabilizing ones, we created a balanced dataset in order to take into account this issue. We also took into account the fact that most predictors are not able to handle directly multimeric proteins, by creating two different datasets, one for monomeric and one for multimeric proteins, and assessing separately the predictions made on these two groups.
The predictors assessed were: INPS-3D [4] (https://inpsmd.biocomp.unibo.it/inpsSuite/default/index3D), a machine-learning method tailored to face the problem of anti-symmetric property; PoPMuSiC [5] (https://soft.dezyme.com/query/create/pop), a method formed by a linear combination of statistical potentials, tailored to correct the bias toward destabilizing mutations; DynaMut [6] (http://biosig.unimelb.edu.au/dynamut/), one of the most recent Web servers developed, based on Normal Mode to take into account the contribution of protein flexibility; DUET [7] (http://biosig.unimelb.edu.au/duet/), a consensus predictor; MAESTROweb [8] (https://pbwww.che.sbg.ac.at/maestro/web), the only Web server able to manage both multimeric proteins and compound heterozygous multiple mutations.
To assess the reliability of the predictors, we evaluated if the sign of the ΔΔG predicted by the different tools was in agreement with the sign of the experimental measure associated to the same mutation, and we calculated for each method its accuracy, sensitivity, specificity, precision towards the different datasets created. We calculated the Receiver Operating Characteristic (ROC) curve and the Precision-Recall curve (PRC), and calculated the Area Under the Curve (AUC) for both curves to compare the various prediction methods. All statistics were carried out in **R** language, using in particular the **precrec** package for confusion matrix [9], and **ggplot2** package for graphs [10].

Results and Discussion

Our final “gold standard” dataset for benchmarking includes 48 proteins, of which 10 are homodimeric, 1 is homotetrameric and the others are monomeric. Mutations affecting monomeric proteins are 759, mutations affecting multimeric proteins are 265, for a total of 1024 mutations. 585 mutations are considered destabilizing ($\Delta\Delta G < -0.5$ kcal/mol), 168 slightly destabilizing ($-0.5 \leq \Delta\Delta G < 0$ kcal/mol) and 103 slightly stabilizing ($0 < \Delta\Delta G \leq 0.5$ kcal/mol), 147 stabilizing ($\Delta\Delta G > 0.5$ kcal/mol), and 21 mutations with $\Delta\Delta G=0$. Since mutations with $\Delta\Delta G<0$ are three times those with $\Delta\Delta G>0$, we derived two subsets of mutations (one for monomeric and one for multimeric proteins) well balanced for the distribution of ΔΔG, by keeping all stabilizing mutations, and picking an equal number of destabilizing mutations, balanced for type of protein, type of mutation, and ΔΔG distribution.
Statistics calculated on those mutations causing a ΔΔG energy variation outside the range of the experimental error of 0.5 kcal/mol, calculated on the total reference datasets of monomeric and multimeric proteins (Figure 1) show that in general, DynaMut underperforms the other methods, and that predictions made on monomers are more reliable than those made on multimeric proteins. Statistics calculated on those mutations causing a ΔΔG energy variation inside the range of the experimental error of 0.5 kcal/mol show that all the methods in this range are unreliable (data not shown). The analyses made on the balanced dataset, however, show that the MCC of the predictions of all predictors, including those that claimed to be tailored expressly to take into account the antisymmetric property, are clearly low (Table 1). In particular, DynaMut is the predictor with the highest precision (true positive rate), indicating that it is the less biased towards destabilizing mutations. Using a consensus of predictors, the MCC is higher with respect to the MCC calculated for every single predictor. The consensus of 3/5 predictors performs better in the full dataset of monomeric proteins, whereas the consensus of 2/3 predictors performs better in the balanced dataset of monomeric proteins and in the full dataset of multimeric proteins (Table 2).
From these data, it appears that: i. there is a general improvement in the reliability of predictors developed in the last years with respect to the older ones [2], but there is still room for improvement; ii. despite the efforts of their developers, most predictors are still biased towards destabilizing mutations; iii. most predictors available are not able to manage multimeric proteins, and in general, it is more difficult to derive an overall result in the case of multimeric proteins; iv. when the experimental ΔΔG value of a mutation is close to the experimental error, all the predictors return essentially random predictions for this mutation; v. using the results of single predictors to perform an "in house" consensus procedure, it is possible to increase the reliability compared with the single best performing method.

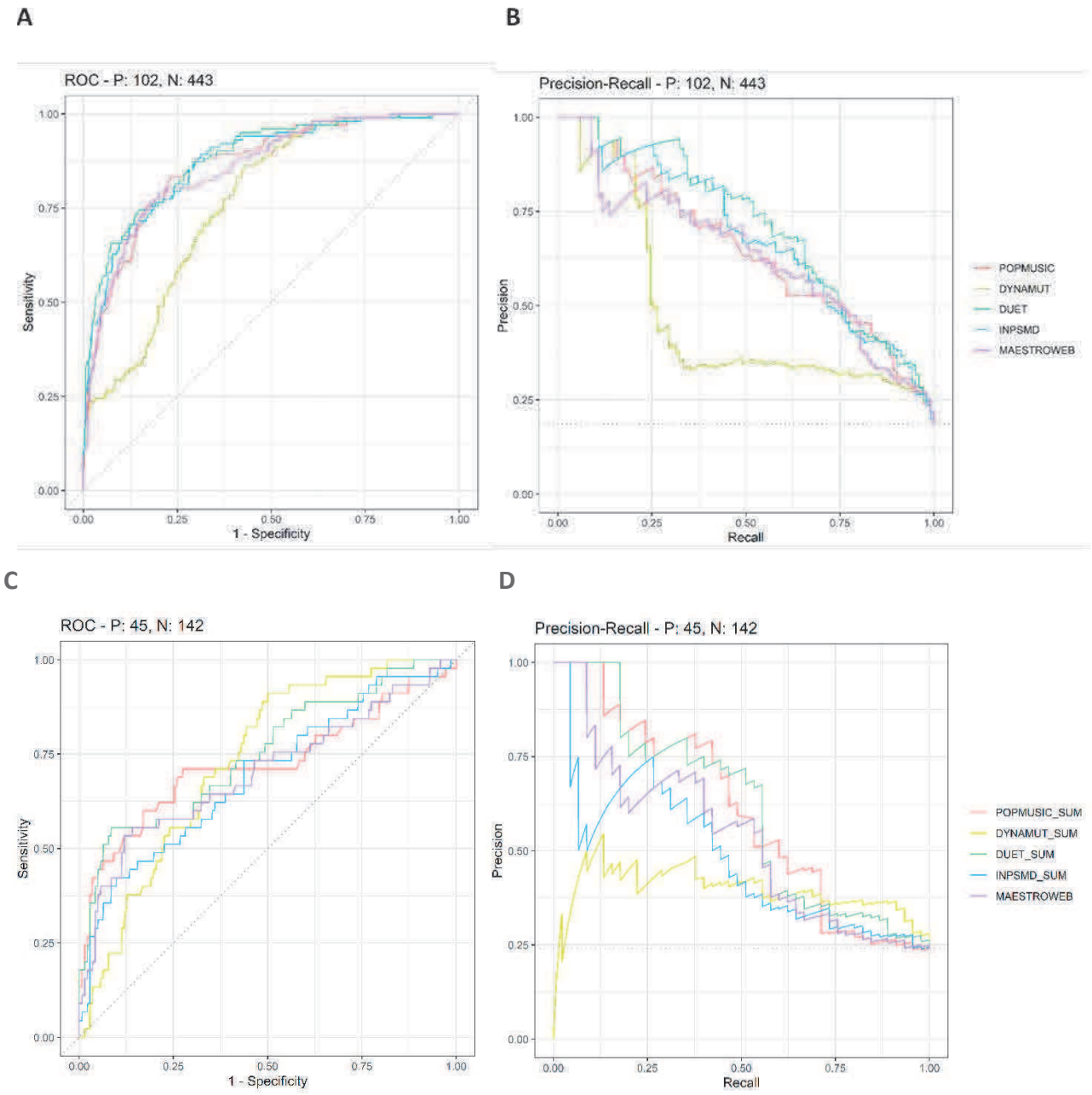


Figure 1: Panels (A) and (B) show, respectively, the ROC and PRC for predictions made on the dataset of monomeric proteins, taking into account only those mutations with a ΔΔG value outside the range of the experimental error. Panels (C) and (D) show, respectively, the ROC and PRC for predictions made on the dataset of multimeric proteins, taking into account only those mutations with a ΔΔG value outside the range of the experimental error.

Table 1: general results from the assessment on the balanced dataset of monomeric and multimeric proteins

True negative and true positive values have been considered as those predictions that correctly predicted a negative and a positive sign for destabilizing and stabilizing mutations, respectively. For PoPMuSiC and MAESTROweb that assume a positive sign for destabilizing mutations, we inverted the sign of their output.

Values calculated for mutations in the balanced dataset of monomeric proteins causing a ΔΔG >|0.5|kcal/mol					
	PoPMuSiC	DynaMut	DUET	INPS-MD	MAESTROweb
Accuracy	0.68	0.65	0.75	0.71	0.72
Specificity	0.62	0.70	0.70	0.64	0.68
Sensitivity	0.85	0.61	0.83	0.87	0.79
Precision	0.43	0.78	0.63	0.48	0.61
MCC	0.40	0.31	0.51	0.46	0.46
Values calculated for mutations in the balanced dataset of multimeric proteins causing a ΔΔG >|0.5|kcal/mol					
	PoPMuSiC	DynaMut	DUET	INPS-MD	MAESTROweb
Accuracy	0.69	0.61	0.67	0.59	0.64
Specificity	0.60	0.58	0.61	0.53	0.57
Sensitivity	0.91	0.64	0.78	0.82	0.76
Precision	0.47	0.67	0.55	0.31	0.49
MCC	0.46	0.22	0.38	0.29	0.32

Table 2: consensus of the predictors.

True negative and true positive values have been considered as those predictions that correctly found a negative and a positive sign for destabilizing and stabilizing mutations, respectively. For PoPMuSiC and MAESTROweb that assume a positive sign for destabilizing mutations, we inverted the sign of their output. Data have been reported for mutations causing a ΔΔG energy variation outside the range of the experimental error, in the full and the balanced datasets of both monomeric and multimeric proteins. [a]: results obtained using DynaMut, DUET and INPS-MD; [b]: results obtained using PoPMuSiC, DUET and MAESTROweb; [c]: results obtained using PoPMuSiC, DUET and INPS-MD.

Values calculated the full dataset				
	Monomeric proteins		Multimeric proteins	
	3/5 methods	2/3 methods[a]	3/5 methods	2/3 methods[b]
Accuracy	0.90	0.88	0.83	0.84
Specificity	0.92	0.93	0.85	0.86
Sensitivity	0.79	0.67	0.73	0.72
Precision	0.62	0.72	0.49	0.53
MCC	0.64	0.62	0.50	0.53
Values calculated for the balanced dataset				
	Monomeric proteins		Multimeric proteins	
	3/5 methods	2/3 methods[a]	3/5 methods	2/3 methods[c]
Accuracy	0.76	0.78	0.69	0.69
Specificity	0.70	0.75	0.60	0.60
Sensitivity	0.86	0.83	0.91	0.91
Precision	0.62	0.72	0.47	0.47
MCC	0.54	0.57	0.46	0.46

Conclusion

Despite the high numbers of predictors available, and despite the continuous work made by developers in the past decades, the reliability of these tools is still far from ideal and the bias towards destabilizing mutations is still present even in the most recent predictors. We advise especially “naive” users to interpret with caution those predicted results falling in the range of the experimental error, preferring in that case to define the effect of the mutation on the stability of the protein as "uncertain".

Acknoweldgements

This work was supported by the University of Salerno, Fondi di Ateneo per la Ricerca di base [grant numbers ORSA170308, ORSA180380, ORSA199808 to A.M.]; and by the Italian Ministry of University and Research, FFABR 2017 program, and PRIN 2017 program [grant number: 2017483NH8 to A.M.]. The work was made in the frame of ELIXIR-IIB (elixir-italy.org), the Italian Node of the European ELIXIR infrastructure (elixir-europe.org).

References

[1] Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 2005;353:459-473.
[2] Marabotti A, Scafuri B, Facchiano A. Predicting the stability of mutant proteins by computational approaches: an overview. Brief Bioinform Jun 3:bbaa074 (2020). Online ahead of print. doi: 10.1093/bib/bbaa074. PMID: 32496523.
[3] Nair PS, Vihinen M. VariBench: A benchmark database for variations. Hum Mutat 2013;34:42-49.
[4] Savojardo C, Fariselli P, Martelli PL et al. INPS-MD: a web server to predict stability of protein variants from sequence and structure. Bioinformatics 2016;32:2542–2544.
[5] Pucci F, Bernaerts KV, Kwasigroch J, Rooman M. Quantification of biases in predictions of protein stability changes upon mutations. Bioinformatics 2018;34:3659-3665.
[6] Rodrigues CH, Pires DEV, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 2018;46:W350–W355.
[7] Pires DEV, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations of protein stability using an integrated computational approach. Nucleic Acids Res 2014;42:W314–W319.
[8] Laimer J, Hiebl-Flach J, Lengauer D, Lackner P. MAESTROweb: a web server for structure-based protein stability prediction. Bioinformatics 2016;32:1414-1416.
[9] Saito T, Rehmsmeier M. precrec: fast and accurate precision-recall and ROC curve calculations in R. Bioinformatics 2017;33:145-147.
[10] Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer International Publishing, 2016.

Computational Study of Action Potential Generation in Uterine Smooth Muscle cell

Chitaranjan Mahapatra

University of California San Francisco

Email of Corresponding author: chitaranjan.mahapatra@ucsf.edu

Abstract

Abstract: Stress urinary incontinence is defined by the involuntary loss of urine during the sneezing and coughing. The urethral smooth muscle cell contributes to stress urinary incontinence by generating spontaneous mechanical and electrical activities. It generates spontaneous electrical events in the terms of membrane depolarization and action potentials. Therefore, a complete understanding of the urethral smooth muscle cell's spontaneous action potential biophysics will help in identifying novel pharmacological targets for the stress urinary incontinence. The action potential is evoked by the activation of various ion channels across the cell membrane. This study aims in establishing a computational model of the single urethral smooth muscle cell to simulate the action potential after incorporating all-important ion channels. The ion channels are designed with Hodgkin- Huxley formalism, where the internal kinetics are expressed in terms of the ordinary differential equations. This computational model generates experimental spontaneous action potential and the underlying ionic currents in urethral smooth muscle cell successfully. In summary, this mathematical model contributes an elemental tool to investigate the physiological ionic mechanisms underlying the spikes in the urethral smooth muscle cell, which in turn can shed light on the genesis of stress urinary incontinence.

Introduction

The International Continence Society has defined urinary incontinence (UI) as a condition in which involuntary loss of urine is objectively demonstrable and is a social or hygiene problem [1]. Among different types of UI, stress urinary incontinence (SUI) is one, which is a common syndrome in women that is typically associated with advanced age, obesity, diabetes mellitus, and fertility [2]. Stress urinary incontinence, defined as a "complaint of involuntary loss of urine on effort or physical exertion or on sneezing or coughing" by the International Continence Society [3, 4]. The smooth muscles from the urinary bladder and urethra display spontaneous contractility patterns, which are associated with UI and SUI. The mammalian urethra is known to exhibit spontaneous tonic contraction activity during the urine-storage phase [5]. Although the factors regulating the SUI are not still precisely identified, it is also widely demonstrated that the abnormal urethral smooth muscle (USM) cell contraction phenomena play an important role in regulating these activities [6, 7, 8]. The isolated USM cell from various species shows slow waves, spontaneous depolarization (SD), and spontaneous action potentials (sAPs) as its' intracellular electrical activity [7, 9, 10]. The sAPs trigger spontaneous contractions by permitting extracellular calcium (Ca^{2+}) via the voltage-gated Ca^{2+} channels across the membrane and releasing stored Ca^{2+} from the sarcoplasmic reticulum (SR) in the intracellular compartment [5,10,11]. The resting membrane potential (RMP) values of the USM cell are in the range from – 35 mV to – 45 mV [12, 13, 14]. The sAPs can be fired spontaneously or evoked by the external stimulation [13]. The array of ion channels located across the USM cell membrane play a crucial role in regulating both RMP and sAP formation and therefore the overall function of the urethra [15]. Therefore, a better understanding of the ion channel kinetics in forming the USM cell sAP would shed light on developing improved therapies for the SUI.

Biophysical constrained computational models always provide a virtual experimental set up to investigate the underlying ionic mechanisms for the cell's electrical activities. Over the past decades, several computational models have been developed for the neuronal and cardiac cells to investigate individual ion channels' contribution in generating the action potential. However, there are a few numbers of

computational models are developed for smooth muscle electrophysiology. To address this gap, recently, we have developed a biophysically constrained computational model for the detrusor smooth muscle (DSM) AP by incorporating nine ion channels [16, 17, 18 , 19]. As both DSM and USM contractions are related to UI and SUI, this paper presents the first biophysically based model of USM AP which integrates some ionic currents underlying the electrogenic processes in the urethra. This single-cell USM model can be subsequently coupled to other active ionic currents and a syncytium model to examine hypotheses concerning the generation of SUI.

Methods

The first step in developing this computational model is to form a conceptual model expressed by the mathematical equations. The classical Hodgkin-Huxley (HH) approach is implemented to form this conceptual model. According to the HH formalism, the cell membrane can be interpreted into an equivalent parallel conductance circuit consisting of membrane capacitance and several variable conductances representing all ion channels. The USM cell model simulation is performed in "NEURON" [20] software environment. The "NEURON" simulation platform is designed to investigate electrophysiological properties in biological excitable cells at different spatiotemporal levels. For USM cell geometry, a cylindrical morphology is considered with length and diameter of 200 μm and 6μm respectively. The membrane capacitance (C_m), membrane resistance (R_m), and axial resistance (R_a) are basic electrical properties of the excitable cell membrane. For this model, the C_m, R_m, and R_a are taken as 1μF/cm^2, 138MΩ– cm^2, and 181Ω-cm respectively. Figure 1 illustrates the USM cell model as a parallel conductance model. The membrane capacitance (C_m) is shunted by an array of ion channel conductances g_{ion} with respective Nernst potentials E_{ion}. The ion channels in the USM cell model are Ca^{2+} activated Cl^- channel (g_{CaCl}, E_{Cl}), voltage-gated Ca^{2+} channel (g_{CaL}, g_{CaT}, E_{Ca}), voltage-gated K^+ channel (g_{Kv}, E_K), Ca^{2+} activated K^+ channel(g_{Kca}, E_K), ATP-dependent K^+ channel (g_{KATP}, E_K) and leakage currents (g_{Leak}, E_{Leak}). The leakage current is considered as a constant value. Applying Kirchhoff's current law, we will get the following differential equation describing changes in transmembrane potential V_m. The time dependence of the membrane potential is governed by the following differential equation

$$\frac{dV_m}{dt} = -\frac{I_{ion\,(t)}}{C_m} \qquad (1)$$

where both V_m, and I_{ion} represent the transmembrane potential and sum of the ionic currents across the cell membrane. The units of both V_m and I_{ion} are in mV and pA respectively.

$$\frac{dV_m}{dt} = -\frac{1}{C_m}\,(I_{Cl} + I_{Ca} + I_K + I_l) \qquad (2)$$

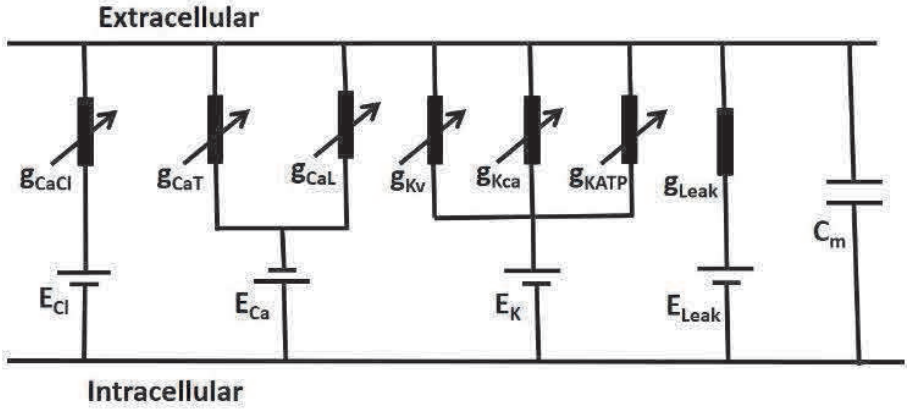


Figure 1. A USM cell parallel conductance Model. It describes all membrane currents and transmembrane potential.

All ionic currents were modeled according to the Hodgkin-Huxley formalism, which is expressed by the following equation.

$$I = \bar{g} m^x h^y (V_m - E_{rev}) \qquad (3)$$

where $\bar{g}$ is the maximum conductance, E_{rev} is the ion's Nernst/reversal potential, m and h are the dimensionless activation and inactivation gating variables.
Both m and h are dependent upon membrane potential and time. First order differential equations are used to express the time dependent properties of both m and h. The following differential equation represents the dynamics of 'm' variable.

$$\frac{dm}{dt} = \frac{(m_\infty - m)}{\tau_m} \quad (4)$$

where m_∞ is the steady-state value of the m and τ_m, is the time constant for reaching the steady-state value. These are also functions of voltage and/or ionic concentrations.
In addition, the steady-state inactivation and activation values for all ion channels are described by the following Boltzman equation.

$$m_\infty = {}^{1}\Big/1 + \exp\left((V_m + V_{\frac{1}{2}})/s\right) \quad (5)$$

Where $V_{1/2}$ is the half activation potential and S is the slope factor. For our model, both $V_{1/2}$ and S are taken from the published experimental data.
The sAPs were induced in the whole-cell model by applying an external stimulus current as brief rectangular pulses or synaptic input.

Results

There is an array of ion channels discovered in USM cell electrophysiology to regulate the cell's excitability. It includes T and L-type voltage-gated Ca^{2+} channels (I_{CaL} and I_{CaT}), ATP-dependent K^+ channel (I_{KATP}), two outward rectifying voltage-gated K^+ channel (I_{KA} and I_{Kv}), $Ca^{2+,}$ and voltage-dependent large-conductance K^+ channel (I_{KCa}), Ca^{2+} dependent Cl^- channel (I_{ClCa}) and the leakage channel (I_{Leak}). The biophysical details of one inward current (I_{CaL}) and one outward current (I_{Kv}) are presented in the following section.

L-type Calcium current (I_{CaL})

Several researcher groups have elucidated the presence of two types of Ca^{2+} channel (Transient and long-lasting type) in USM cell electrophysiology. However, the L-type (Long-lasting) Ca^{2+} channel (I_{CaL}) is responsible for the major inward current in USM cells [5,15]. It is demonstrated that I_{CaL} is activated first between $V_m \approx -35$ and -20mV; the peak magnitude of the current-voltage (I–V) relationship curve appears at $V_m \approx 10$mV. The half-activation potentials for both steady-state activation and inactivation curve are –3.4 mV and –24.8 mV respectively. The Nerst potential E_{CaL} is fixed at 45mV. The equations of I_{CaL} incorporate both activation (m) and inactivation (h) gating variables. The biophysical parameters for the I_{CaL} are extracted from the published experimental data in human USM electrophysiology [21]. Figure 2 (A) shows the steady-state activation and inactivation curve with respect to membrane potential.

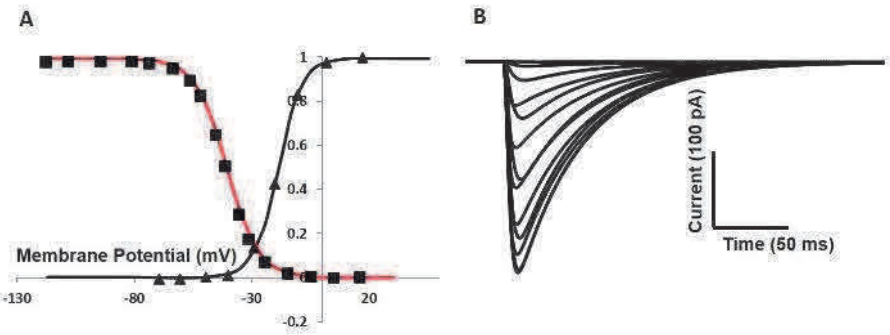


Figure 2. USM I_{CaL} model. A steady state activation and inactivation parameter curve and B shows the current traces from the voltage clamp protocol

The red and black solid lines represent simulated steady-state curves for inactivation and activation parameters respectively. The superimposed filled squares and triangles represent the experimental data [21].

The whole-cell current I_{CaL} is simulated according to the voltage clamp protocol for a duration of 200 ms. The holding potential is –70 mV. Simulated tracings of I_{CaL} are shown in figure 2 (B).

Voltage gated K^+ current (I_{Kv})

Like the DSM cell, a number of different K^+ channels with small and large conductance properties are reported in the USM cell [22]. Among the various K^+ channels, the voltage-gated outward rectifier K^+ channels (I_{Kv}) is identified in many spices [Brading et al., 2006]. The I_{Kv} is partially responsible for containing the outward current during the repolarization phase of the action potential. It is demonstrated that I_{Kv} is significantly found as a part of the outward current after $V_m \approx -20$ mV. The half-activation potentials for both steady-state activation and inactivation curve are –7 mV and –56 mV respectively. The Nerst potential E_K is fixed at –70 mV. The equations of I_{Kv} also incorporate both activation (m) and inactivation (h) gating variables. The biophysical parameters for the I_{Kv} are extracted from the published experimental data in rabbit USM electrophysiology [22]. Figure 3 (A) shows the steady-state activation and inactivation curve with respect to membrane potential. The red and black solid lines represent simulated steady-state curves for inactivation and activation parameters respectively. The superimposed filled squares and triangles represent the experimental data [22]. The whole-cell current I_{Kv} is simulated according to the voltage clamp protocol for a duration of 500 ms. The holding potential is –60 mV. Simulated tracings of I_{Kv} are shown in figure 3 (B).

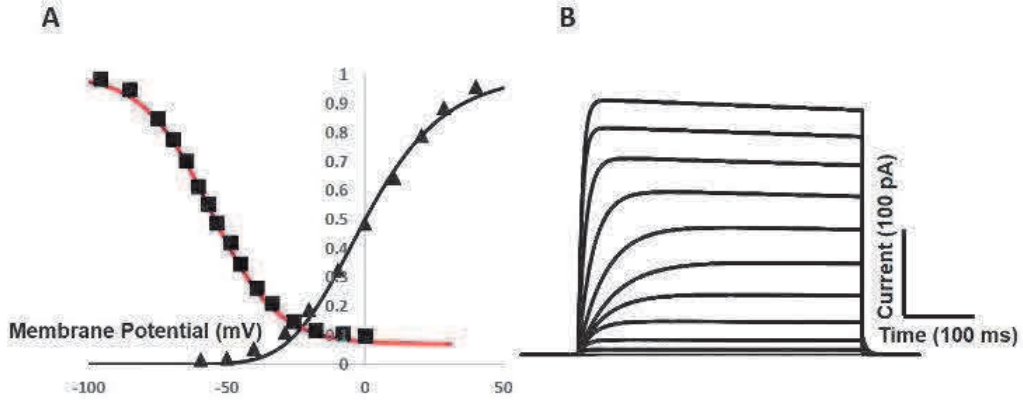


Figure 3. USM I_{Kv} model. A steady state activation and inactivation parameter curve and B shows the current traces from the voltage clamp protocol

AP Simulation

The AP can be evoked either by the external current injection via the inserted electrode or by the induced synaptic input from the neighbor nerve. Seven numbers of ionic conductances are incorporated into this single USM cell model. The USM cell model successively responded to both current and synaptic input stimuli by showing all-or-none AP firing properties. A current input is a step input pulse with different amplitudes and durations. A synaptic input is also mimicked by the alpha function to evoke AP in our model. The voltage threshold is ≈ -35mV. Figure 4 presents the simulated AP after inducing a synaptic input to mimic the experimental AP in [22].

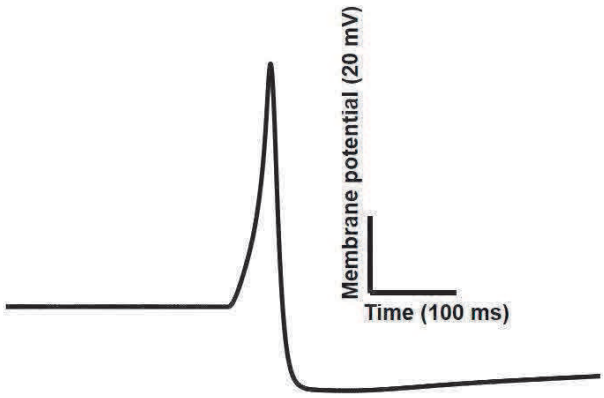


Figure 4. The simulated AP in the USM model.

Table 1 compares the simulated AP with experimental one [22] in terms of RMP, peak amplitude, AP duration, and AHP (afterhyperpolarization potential).

TABLE I Comparison between simulated AP and Experimental AP [22]

	RMP (mV)	Peak (mV)	AHP (mV)	Duration (ms)
Experiment	− 60	11	− 73	97
Simulation	− 60	12	−72	98

Discussion

The primary objective of this study was to develop and validate a computational model of a USM cellular electrophysiology. The model description integrates those ion channels that were significantly contributing to generate the USM cell AP. The ion channel kinetics are characterized by the Hodgkin and Huxley formalism after extracting all parameter values from the literature on USM electrophysiology. The model has demonstrated its' ability by simulating the experimental AP successfully.

The assumptions and simplification approaches are concerned about developing a perfect mathematical model.

A better physiologically realistic model is always based on enough electrophysiological data obtained from a single species. However, due to experimental setup complexity, these data are not always available from the same species. We, therefore, made assumptions driven from values obtained from USM in different species (rat, human, mouse, pig, guinea pig, and rabbit) and under various experimental conditions. Some debate also exists with regard to the ionic conductances that are involved in the repolarizing phase. It has been suggested that more than one K^+ conductance (for example fast A-type K^+ current [15] may carry a portion of the outward current. However, due to a lack of experimental evidence, this model doesn't include this channel. Another question can also be raised towards simulating the experimental AP when the single USM cell is coupled to the other cell.

In the present state, this model is at an elementary stage. Integration of other active channels, Na^+- Ca^{2+} exchanger, Ca^{2+} ATPase pump and sarcoplasmic reticulum Ca^{2+} releasing mechanism will improve this model towards a more comprehensive stage. In addition, the expansion of this single-cell model to syncytium or network level will help to establish a better physiologically realistic computational model for investigating the SUI.

References

1. Abrams P, Cardozo L, Fall M, Griffiths D, Rosier P, Ulmsten U, Van Kerrebroeck P, Victor A, Wein A. The standardisation of terminology in lower urinary tract function: report from the standardisation sub-committee of the International Continence Society. Urology. 2003 Jan 1;61(1):37-49.
2. Chen D, Meng W, Shu L, Liu S, Gu Y, Wang X, Feng M. ANO1 in urethral SMCs contributes to sex differences in urethral spontaneous tone. American Journal of Physiology-Renal Physiology. 2020 Sep 1;319(3):F394-402.
3. Chancellor MB, Yoshimura N. Neurophysiology of stress urinary incontinence. Reviews in urology. 2004;6(Suppl 3):S19.
4. Yono M, Irie S, Gotoh M. TAS-303 effects on urethral sphincter function in women with stress urinary incontinence: phase I study. International Urogynecology Journal. 2020 Oct 10:1-8.
5. Brading AF. Spontaneous activity of lower urinary tract smooth muscles: correlation between ion channels and tissue function. The Journal of physiology. 2006 Jan;570(1):13-22.
6. Greenland JE, Dass N, Brading AF. Intrinsic urethral closure mechanisms in the female pig. Scandinavian journal of urology and nephrology. Supplementum. 1996;179:75.
7. Hollywood MA, McCloskey KD, McHale NG, Thornbury KD. Characterization of outward K+ currents in isolated smooth muscle cells from sheep urethra. American Journal of Physiology-Cell Physiology. 2000 Aug 1;279(2):C420-8.

8. Feng M, Wang Z, Liu Z, Liu D, Zheng K, Lu P, Liu C, Zhang M, Li J. The RyR–ClCa–VDCC axis contributes to spontaneous tone in urethral smooth muscle. Journal of cellular physiology. 2019 Dec;234(12):23256-67.
9. Hashitani H, Van Helden DF, Suzuki H. Properties of spontaneous depolarizations in circular smooth muscle cells of rabbit urethra. British journal of pharmacology. 1996 Aug;118(7):1627.
10. Hashitani H, Edwards FR. Spontaneous and neurally activated depolarizations in smooth muscle cells of the guinea-pig urethra. The Journal of Physiology. 1999 Jan;514(2):459-70.
11. Berridge MJ. Smooth muscle cell calcium activation mechanisms. The Journal of physiology. 2008 Nov 1;586(21):5047-61.
12. Teramoto N, Creed KE, Brading AF. Activity of glibenclamide-sensitive K+ channels under unstimulated conditions in smooth muscle cells of pig proximal urethra. Naunyn-Schmiedeberg's archives of pharmacology. 1997 Aug 1;356(3):418-24.
13. Creed KE, Oike M, Ito Y. The electrical properties and responses to nerve stimulation of the proximal urethra of the male rabbit. British journal of urology. 1997 Apr;79(4):543-53.
14. Hashitani H, Suzuki H. Properties of spontaneous Ca2+ transients recorded from interstitial cells of Cajal-like cells of the rabbit urethra in situ. The Journal of Physiology. 2007 Sep;583(2):505-19.
15. Kyle BD. Ion channels of the mammalian urethra. Channels. 2014 Sep 3;8(5):393-401.
16. Mahapatra C, Brain KL, Manchanda R. Computational study of Hodgkin-Huxley type calcium-dependent potassium current in urinary bladder over activity. In2018 IEEE 8th international conference on computational advances in bio and medical sciences (ICCABS) 2018 Oct 18 (pp. 1-4). IEEE.
17. Mahapatra C, Brain KL, Manchanda R. A biophysically constrained computational model of the action potential of mouse urinary bladder smooth muscle. PloS one. 2018 Jul 26;13(7):e0200712.
18. Mahapatra C, Brain KL, Manchanda R. Computational studies on urinary bladder smooth muscle: Modeling ion channels and their role in generating electrical activity. In2015 7th International IEEE/EMBS Conference on Neural Engineering (NER) 2015 Apr 22 (pp. 832-835). IEEE.
19. Mahapatra C, Manchanda R. Simulation of In Vitro-Like Electrical Activities in Urinary Bladder Smooth Muscle Cells. InJournal of Biomimetics, Biomaterials and Biomedical Engineering 2017 (Vol. 33, pp. 45-51). Trans Tech Publications Ltd.
20. Hines ML, Carnevale NT. The NEURON simulation environment, neural computation.
21. Hollywood MA, Woolsey S, Walsh IK, Keane PF, McHale NG, Thornbury KD. T-and L-type Ca2+ currents in freshly dispersed smooth muscle cells from the human proximal urethra. The Journal of physiology. 2003 Aug;550(3):753-64.
22. Kyle B, Bradley E, Ohya S, Sergeant GP, McHale NG, Thornbury KD, Hollywood MA. Contribution of Kv2. 1 channels to the delayed rectifier current in freshly dispersed smooth muscle cells from rabbit urethra. American Journal of Physiology-Cell Physiology. 2011 Nov;301(5):C1186-200.

Studying the early molecular defects of ICF syndrome in patient-derived and CRISPR-corrected iPSCs using an integrated multi-omic approach

Varsha Poondi Krishnan[1], Monika Krzak[2], Shir Toubiana[3], Sara Selig[3], Claudia Angelini[4], Maria Rosaria Matarazzo[1]

[1] Institute of Genetics and Biophysics "A. Buzzati Traverso", Consiglio Nazionale delle Ricerche,Naples, Italy

[2] Wellcome Sanger Institute,Wellcome Genome Campus, Hinxton,Cambrige, United Kindgom

[3] Molecular Medicine Laboratory, Rambam Health Care Campus and Rappaport Faculty of Medicine, Technion,Haifa, Israel

[4] Istituto per le Applicazioni del Calcolo "Mauro Picone" ,Naples, Italy

Email of Corresponding author:

varsha.poondi@igb.cnr.it, c.angelini@iac.cnr.it, maria.matarazzo@igb.cnr.it

Abstract

DNMT3B is one of the major de novo methyltransferases responsible for the genome-wide methylation during the early stages of embryonic development. Immunodeficiency, Centromeric instability and Facial anomalies syndrome (ICF syndrome) is a rare autosomal recessive disorder where about 60% of the patients carry hypomorphic mutations in DNMT3B gene. The wide spectrum and varying degree of severity of clinical phenotypes can be postulated to the genome-wide effect of DNMT3B dysfunction. To elucidate the early molecular mechanisms involved in the pathogenesis of ICF syndrome , we have performed comparative multi-omic analysis of Whole Genome Bisulfite Sequencing(WGBS) and RNA-Seq datasets from control, patient-derived iPSCs and their CRISPR/Cas9-corrected clones.

The global level of DNA methylation was not dramatically reduced in patients compared to controls. However, we identified about 27,000 differentially methylated regions (DMRs) uniformly distributed across the chromosomes. Approximately 74% of hypomethylated DMRs in patients were rescued in both of their respective corrected clones. A significant percentage of differentially expressed genes in patients compared to controls were associated to DMRs.

To understand the complex interplay between the methylation and expression defects within the chromatin context, ChIP-Seq data for DNMT3B binding and H3K4me3 and H3K36me3 marks were analysed and the integrated results will be discussed in detail.

Integrated eco-physiological and metabolomic analyses of the amphibious plant *Butomus umbellatus* under light limitation and nutrient varying conditions

Georgia Tooulakou 1, Paraskevi Manolaki 2,3, Caroline Urup Byberg 3, Franziska Eller 3, Brian Keith Sorrell 3, Tenna Riis 3, Maria I. Klapa 1*

1 Metabolic Engineering and Systems Biology Laboratory, Institute of Chemical Engineering Sciences, Foundation for Research & Technology-Hellas (FORTH/ICE-HT), Patras, Greece

2 Aarhus Institute of Advanced Studies, AIAS, Aarhus, Denmark

3 Aquatic Biology, Department of Biology, Aarhus University, Aarhus, Denmark

*Corresponding author

email: mklapa@iceht.forth.gr

Abstract

Amphibious plants have to cope with ever changing growth conditions in their habitats with respect to nutrient and light availability, having evolved specialized adaptation mechanisms. Furthering our currently limited understanding of these toleration processes is a major ecophysiology research area. The major objective of this study was to systemically investigate the response of *Butomus umbellatus*, a common amphibious species in Denmark, to nutrient level changes and shading in a large-scale mesocosm experiment, through integrated morphological, eco-physiological and metabolomic analyses with the use of systems biology and bioinformatic tools. Using this methodological approach, we were able to identify when the increase in nutrient levels, initially promoting plant growth, reached a point of saturation for *B. umbellatus* physiological acclimation and tolerance. Moreover, multivariate analysis of the combined morpho-physiological and metabolomic profiles indicated them as discriminatory of the shading compared to the open treatment conditions independently of the nutrient level, while this discrimination is not directly available from the eco-physiological measurements alone. Our results underline the usefulness of the systems biology methodological framework in stream ecology research. The challenge of the ecological research is to adapt these analytical protocols and use the results for open field experiments and studies.

Exploiting deep learning embeddings for sub-peroxisomal localisation

Marco Anteghini[1,2], Edoardo Saccenti[2], Vitor AP Martins dos Santos[1,2]

1 LifeGlimmer GmbH, Berlin, Germany
2 Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen WE, The Netherlands.

anteghini@lifeglimmer.com

Abstract

We are the first investigating methods relying on new deep learning related encoding features (e.g. UniRep) for sub-peroxisomal localisation. This study aims to highlight these feature's performances and understand how they can contribute to the peroxisome-related research and proposing insights for an extension to other organelles. Deep learning embedding methods outperformed the more classical ones when applied to sub-peroxisomal and sub-mitochondrial classification. The combination of SeqVec and UniRep as encoding features, showed promising results with several Machine Learning methods, in particular with Logistic Regression and Support Vector Machines among the analysed ones. Thus, suggesting to adopt the usage of these protein representations for sub-organelle classification purposes as a conventional approach. Also, we present In-Pero and In-Mito, two sub-organelle protein localisation predictors based on our findings

Introduction

The simple structure of peroxisomes, ubiquitous organelles surrounded by a single biomembrane, is the reason why we can define a binary classification problem for peroxisomal proteins. More precisely, its proteins can be found attached to the membrane of the organelle or in its granular matrix.

Peroxisomes related research is in constant evolution, in accordance with the discovery of both their metabolic and non-metabolic roles and the association between their dysfunction and metabolic disorders in humans [1, 2].

Many roles of peroxisome are still unknown. As a starting point, it is relevant to detect membrane contact site (MCS) proteins [3] and peroxisomal transporters (PT) [4], to discover new peroxisomal proteins with relevant functions. Both the categories mentioned, are generally located on the peroxisomal membrane. That justifies our need to classify which proteins are located on the peroxisomal membrane and which ones are not.

Although many bioinformatics methods for sub-cellular and sub-organelle localization are easily findable and accessible [5–8], we do not have this possibility specifically for sub-peroxisomal localization. Moreover, the recent applications of deep-learning (DL) approaches to encode protein sequences, has shown promising results related to sub-cellular classification [9–12] but sub-peroxisomal classification is still not explored.

We here developed and compared different predictors for the sub-peroxisomal localization. Every predictor was trained on peroxisomal protein sequences and based on a

different method, namely Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Partial Least Square Discriminant Analysy (PLS-DA). In addition, we compared different sets of encoding methods, also know as features, including DL based ones. Among them, the combination of UniRep+SeqVec as feature and LR as method, was the most promising and therefor our choice for the final implementation of the predictor.

This approach could potentially be applied to other organelles, such as mitochondria. With mitochondria, we have to consider a multi-class classification problem. This organelle has two membranes, a space between them and an internal matrix, thus creating four compartments [13].

Considering the different structure of mitochondria, we finally applied our approach to sub-mitochondrial localization, realizing a predictor that outperformed most successful existing classifiers.

Materials and Methods

Datasets

The protein sequences (160) used for building our models were retrieved from UniprotKB/SwissProt (Dic 2019). The protein sequences proposed for the Uniprot annotation improvement were also retrieved from Uniprot/SwissProt (Jun 2020). The dataset is available for download at https://github.com/MarcoAnteghini.

Peroxisomal Membrane Proteins

First were selected reviewed, non-fragmented protein sequences, with peroxisomal membrane sub-cellular location (SL-0203), obtaining 327 entries. Clustering was performed with these sequences using Cd-hit [14], with sequence identity of 40%. The representative of each cluster was chosen, reducing the dataset to 162 sequences. An additional filtering process was applied selecting just those proteins with at least 1 publication specific for the sub-cellular localization (135). Three sequences were finally removed from the dataset, since not available for the UniRep encoding procedure. The final dataset contains 132 membrane proteins.

Peroxisomal Matrix Proteins

The same procedure was applied to select peroxisomal matrix proteins. In particular, were selected reviewed, non-fragmented protein sequences, with peroxisomal matrix sub-cellular location (SL-0202), obtaining 60 entries. The number of sequences was reduced to 22 after the same clustering procedure and to 19 after selecting just those proteins with at least 1 publication specific for the sub-cellular localization.
Due to the low number of peroxisomal matrix proteins we performed another advanced search in Uniprot selecting reviewed, non-fragmented protein sequences, with peroxisomal location (SL-0204), and not peroxisomal membrane location (SL-0203). 202 non membrane proteins were found. Applying the same filtering procedures as above described we reduced the dataset to 22 matrix proteins. Merging the two subsets that presented 13 common entries and clustering for 40% of sequence identity we finally obtained 28 sequences.

Datasets for sub-mitochondrial protein classification

Both datasets for sub-mitochondrial classification comparison, retrieved from UniProt/SwissProt,
were available from previous works [7, 15] and accessible at http://busca.biocomp.unibo.it/deepmito/datasets/

and http://proteininformatics.org/mkumar/submitopred/download.html. The two datasets show an overlap of 238 sequences.

- SM424-18: The dataset was used to build the DeepMito predictor [7]. It consists of 424 mitochondrial proteins for which the following filtering criteria were applied. Non-fragment protein sequences with evidence at protein level (1) and experimentally determined subcellular localization in one of the four sub-mitochondrial compartments: outer membrane, intermembrane space, inner membrane and matrix showing experimental evidence code ECO:0000269 (2). Exclusion of proteins also localized in compartments other than mitochondria (3). Clustering procedure using Cd-hit [14], with sequence identity of 40% where the longest sequence from each cluster were retained (4).

- SubMitoPred dataset: The dataset was used to build the SubMitoPred predictor [15]. It consists of 570 mitochondrial proteins for which the following filtering criteria were applied. Non-fragment protein sequences with evidence at protein level (1) and experimentally determined subcellular localization in just one of the four sub-mitochondrial compartments (2). Protein length grater than 50 residues (3). Clustering procedure using Cd-hit [14], with sequence identity of 40% (4).

Features selection

The following features were considered:

- Residue one-hot encoding (1HOT). The protein sequence is represented by a matrix $LX20$, where L is the length of the sequence and 20 is the one-hot encoding of the residue.

- Position Specific Scoring Matrices (PSSM). The protein sequence is represented by a matrix $LX20$, where L is the length of the sequence and each amino acid substitution scores are given separately for each position in a protein multiple sequence alignment (MSA) after running PSI-BLAST [16] against the Uniref90 dataset (release Oct 2019) for three iterations and e-value threshold set to 0.001. We used a sigmoid function to map the values extracted from the PSI-BLAST checkpoint file in the range [0-1].

- Residue physical–chemical properties (PROP). The protein sequence is represented by a matrix $LX10$, where L is the length of the sequence and each residue is encoded using 10 different numerical values representing its physical–chemical nature [17].

- Unified Representation (UniRep) [10]. The protein is represented by an embedding of length 1900 (average final hidden array). UniRep is based on a recurrent neural network architecture (1900-hidden unite) able to learn statistical representations of proteins from 24 million UniRef50 sequences.

- Sequence-to-Vector (SeqVec) [11]. The protein is represented by an embedding of length 1024. SeqVec is based on a transfer learning where ELMo [18] was trained on UniRef50.

We investigated the information content in each feature performing a Principal Component Analysis (PCA) followed by k-means clustering with k=2 clusters. The predicted clusters were compared with the original labels obtaining an overview of the intrinsic classification capability of the features. In parallel, we performed a forward feature selection using a Logistic Regression (LR) model. The method consists in iteratively adding one feature to the set of current best performing features and evaluating the

performance. The procedure is halted once the performance worsens and the best feature set from the previous round is retained. This evaluation is based on a 5-fold cross validation.

Classification methods

The best feature combination was used for training different machine learning (ML) methods: support vector machine (SVM), random forest (RF), partial least squares discriminant analysis (PLS-DA) and logistic regression (LR). We used the implementation available in scikit-learn python library (version 0.22.1) for all the methods except for the PLS-DA, for which we modified the partial least squares (PLS) regression code, available on the same scikit-learn version. We evaluated the performance of each model by performing a 5-fold double cross-validation (DCV) which contains two nested cross-validation (CV) loops. The inner loop is used to optimize model hyper-parameters through a grid search (GS), while the outer loop tests the optimized model performance on a held-out test set. In the inner loop, for each set of hyper-parameters, the average score across the folds was computed. The best score among these was used to select the best set of hyper-parameters. Among the hyper-parameters, the class weight (relevant to handle an unbalanced dataset) was analysed for all the tested methods. The optimized model was then tested in the outer loop.

Scoring metrics

We use F_1 score (macro average), accuracy (ACC), balanced accuracy (BACC), and Matthews correlation coefficient (MCC) throughout the paper to evaluate the performance of the machine learning models. These metrics are defined as follows:

$$F_1 = 2 * \frac{PPV * TPR}{PPV + TPR} \tag{1}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$BACC = \frac{TPR + TNR}{2} \tag{3}$$

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

where:
TP and FP are respectively true positive and false negative,
PPV (positive predictive value) $= TP/(TP + FP)$,
TPR (true positive rate) $= TP/(TP + FN)$,
TNR (true negative rate) $= TN/TN + FP$,
FPR (false positive rate) $= FP/(FP + TN)$

Results

Features Evaluation

We found that the deep learning embedding methods outperformed the more classical ones, when applied to sub-peroxisomal classification. Accessing the intrinsic classification capability of each individual feature as described in section 2.2, UniRep presented the

	F_1	PPV	TPR
PROP	0.57	0.62	0.57
PSSM	0.54	0.57	0.53
1HOT	0.41	0.50	0.45
UniRep	0.63	0.68	0.65
SeqVec	0.55	0.58	0.48

Table 1. Features intrinsic classification capability after PCA and k-means cluster. Metrics: F1, PPV, TPR (macro average)

	F_1
UniRep	0.81
SeqVec	0.79
1HOT	0.62
PSSM	0.60
PROP	0.52
UniRep + SeqVec	0.87
Unirep + PROP	0.66
Unirep + 1HOT	0.57
Unirep + PSSM	0.55
UniRep + SeqVec + PROP	0.73
UniRep + SeqVec + PSSM	0.65
UniRep + SeqVec + 1HOT	0.63

Table 2. Forward Feature Selection results form 5-fold cross validation. Metric: F_1 score (macro average)

best performances in terms of PPV (positive predicted value/precision), TPR (true positive rate/recall) and F_1 score as shown in Table 1. According to our LR model, the best feature combination was UniRep-SeqVec with an F_1 score of 0.87. The best individual feature was UniRep with an F_1 score of 0.81, close to SeqVec with 0.79. Groups of 3 features showed worse performances. A complete overview of the forward feature selection process (described in section 2.2) is visible in Table 2, where the results are reported in terms of F_1.

Methods Comparison

Among the inspected ML methods, LR and SVM showed similar metrics, superior to others. However, the performances obtained with LR are slightly better for all the considered scores, namely F1, BACC, MCC and ACC. The complete methods comparison is visible in Table 3 where the results refer to the DCV procedure explained in section 2.3. The F_1 (inner) score refers to the inner loop of the DCV while F_1 (outer) refers to the outer loop.

	F_1(inner)	F_1(outer)	BACC	MCC	ACC
LR	0.849	0.869	0.867	0.742	0.925
SVM	0.825	0.859	0.863	0.721	0.919
PLS-DA	0.834	0.805	0.802	0.620	0.888
RF	0.794	0.777	0.795	0.570	0.869

Table 3. Methods comparison. Where not indicated the scores refer to the outer CV loop in the DCV.

Extension to Sub-mithocondrial proteins prediction and comparison with other approaches

We also applied our framework for sub-mitochondrial classification, modifying the model for a multiclass classification problem. Mitochondria have four possible sub-

compartments, namely: matrix, internal membrane, intermembrane space and an external membrane.

Our method outperformed some of the available ones, especially the few designed to classify all four mitochondrial compartments. Moreover, it shows a balanced capability to predict the different compartments.

The compared predictors are SubMitoPred [15], DeepMito [7], and DeepPred-SubMito [19]. For the sake of comparisons, we trained our model with the SM424-18 and SubMitoPred datasets and performed a Random Split (RS) 5-fold CV as in SubMitoPred [15]. The results are visible in Table 4.

	CV	MCC(O)	MCC(I)	MCC(T)	MCC(M)
SubMitoPred	RS	0.42	0.34	0.19	0.51
DeepMito	RS	0.45	0.68	0.54	0.79
DeepMito	CL	0.42	0.60	0.46	0.76
DP-SM	RS	**0.92**	0.69	**0.97**	0.73
In-Mito	RS	0.69	**0.75**	0.62	**0.85**
In-Mito	DCV	0.67	0.75	0.62	0.85

Table 4. Performance comparison of different approaches. RS indicate a random split cross-validation, DCV a double cross-validation while CL is a cross-validation performed confining any local similarity into the same cross-validation set. (O),(I),(T),(M) are the 4 mitochondrial compartments, respectively: outer membrane, inner membrane, intermebrane space and matrix

Conclusion

Our predictor can help the research on peroxisomes and mitochondria accurately identifying sub-organelle protein localisation; an example case is shown in section 3.3, suggesting possible experimental validations. We also underlined the advantages of using deep learning encoding methods to represent the protein sequences for this task.

The approach initially designed for sub-peroxisomal protein classification (In-Pero), showed optimal performances, outperforming the state-of-the-art, also for sub-mitochondrial protein classification (In-Mito). That suggests the utility of an extension of this method to other sub-organelles in further studies. In particular, our method shows a balanced capapility in predicting all four sub-mitochondrial compartments.

As the application of DL encoding methods is increasing, we expect to be able to test new features in the next future and eventually test other methods for sub-organelle classification. In this work, we mainly focused on some ML strategies for sub-peroxisomal localisation since DL was relevant for the encoding procedure, but in principle, we can extend our analysis to DL methods, such as convolutional neural networks (CNN), recurrent neural networks (RNN) or combination of the two (e.g. CNN-RNN).

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska Curie grant agreement No. 812968.

Supporting Information

The sandalone versions of the predictors, the datasets and list of candidates are available at https://github.com/MarcoAnteghini

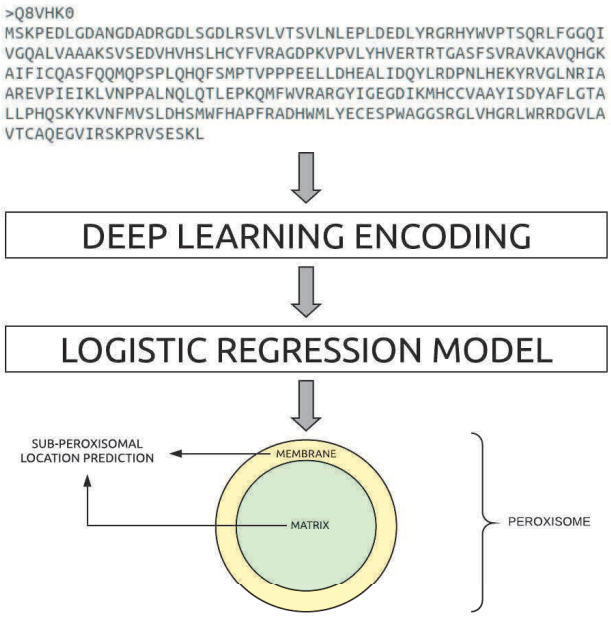


Figure 1. In-Pero workflow in three steps: (1) Protein sequence representation via DL encoding, (2) training the Sub-Pero model with Logistc Regression, (3) sub-peroxisomal location prediction.

	Fungi	Metazoa	Viridiplanta	Protozoa
Matrix	15	9	4	
Membrane	29	38	54	11

Table 5. In-Pero dataset composition in terms of different taxonomic kingdoms

	CV	MCC(O)	MCC(I)	MCC(T)	MCC(M)
DeepMito	10F-CL	0.46	0.47	0.53	0.65
DP-SM	10F	**0.85**	0.49	**0.99**	0.56
In-Mito	5F-DCV	0.68	**0.73**	0.69	**0.82**

Table 6. Comparison with DeepMito and DeepPred-SubMito (DP-SM) based on the SM424-18 dataset. (O),(I),(T),(M) are the 4 mitochondrial compartments, respectively: outer membrane, inner membrane, intermebrane space and matrix.

References

1. Wanders RJA, Waterham HR and Ferdinandusse S. Metabolic Interplay between Peroxisomes and Other Subcellular Organelles Including Mitochondria and the Endoplasmic Reticulum. Frontiers in Cell and Developmental Biology 2016;3:83.

2. Islinger M, Voelkl A, Fahimi H and Schrader M. The peroxisome: an update on mysteries 2.0. Histochemistry and Cell Biology 2018;150:1–29.

3. Farré JC, Mahalingam SS, Proietto M and Subramani S. Peroxisome biogenesis, membrane contact sites, and quality control. EMBO reports 2019;20:e46864.

4. Baker A, Carrier DJ, Schaedler T et al. Peroxisomal ABC transporters: functions and mechanism. Biochemical Society Transactions 2015;43:959–965.

5. Käll L, Krogh A and Sonnhammer EL. A Combined Transmembrane Topology and Signal Peptide Prediction Method. Journal of Molecular Biology 2004;338:1027 – 1036. ISSN 0022-2836.

6. Horton P, Park KJ, Obayashi T et al. WoLF PSORT: protein localization predictor. Nucleic Acids Research 2007;35:W585–W587.

7. Savojardo C, Bruciaferri N, Tartari G et al. DeepMito: accurate prediction of protein sub-mitochondrial localization using convolutional neural networks. Bioinformatics 2019;36:56–64.

8. Jiang Y, Wang D, Yao Y et al. MULocDeep: A deep-learning framework for protein subcellular and suborganellar localization prediction with residue-level interpretation 2020;.

9. ElAbd H, Bromberg Y, Hoarfrost A et al. Amino acid encoding for deep learning applications. BMC Bioinformatics 2020;21.

10. Alley E, Khimulya G, Biswas S et al. Unified rational protein engineering with sequence-based deep representation learning. Nature Methods 2019;16.

11. Heinzinger M, Elnaggar A, Wang Y et al. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics 2019;20.

12. Elnaggar A, Heinzinger M, Dallago C et al. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. bioRxiv 2020;.

13. Kühlbrandt W. Structure and function of mitochondrial membrane protein complexes. BMC Biology 2015;13.

14. Li W and Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–1659.

15. Kumar R, Kumari B and Kumar M. Proteome-wide prediction and annotation of mitochondrial and sub-mitochondrial proteins by incorporating domain information. Mitochondrion 2018;42:11–22.

16. Altschul S, Madden T, Shaffer A et al. Gapped blast and psi-blast:A new generation of protein database search programs. Nucl Acids Res 1996;25:3389–3402.

17. Kidera A, Konishi Y, Oka M et al. Statistical Analysis of the Physical Properties of the 20 Naturally Occurring Amino Acids. Journal of Protein Chemistry 1985; 4:23–55.

18. Peters M, Neumann M, Iyyer M et al. Deep contextualized word representations 2018;.

19. Wang X, Jin Y and Zhang Q. DeepPred-SubMito: A Novel Submitochondrial Localization Predictor Based on Multi-Channel Convolutional Neural Network and Dataset Balancing Treatment. International Journal of Molecular Sciences 2020;21:5710.

ResistoXplorer: a web-based tool for visual, statistical and exploratory data analysis of resistome data

Achal Dhariwal[1], Roger Junges[1], Tsute Chen[2,3] & Fernanda Cristina Petersen[1]

1 Institute of Oral Biology, Faculty of Dentistry, University of Oslo, Oslo, Norway

2 Department of Microbiology, The Forsyth Institute, Cambridge, MA, USA

3 Department of Oral Medicine, Infection, and Immunology, Harvard School of Dental Medicine, Boston, MA, USA.

Email of Corresponding author: f.c.petersen@odont.uio.no

Abstract

The study of resistomes using whole metagenomic sequencing enables high throughput identification of resistance genes in complex microbial communities, such as the human microbiome. Over recent years, sophisticated and diverse pipelines have been established to facilitate raw data processing and annotation. Despite the progress, there are no easy-to-use tools for comprehensive visual, statistical, and functional analysis of resistome data. Thus, exploration of the resulting large complex datasets remains a key bottleneck requiring robust computational resources and technical expertise, which creates a significant hurdle for advancements in the field. Here, we introduce ResistoXplorer, a user-friendly tool that integrates recent advancements in statistics and visualization, coupled with extensive functional annotations and phenotype collection, to enable high-throughput analysis of common outputs generated from metagenomic resistome studies. ResistoXplorer contains three modules- the 'Antimicrobial Resistance Gene Table' module offers various options for composition profiling, functional profiling and comparative analysis of resistome data; the 'Integration' module supports integrative exploratory analysis of resistome and microbiome abundance profiles in metagenomic samples; finally, the 'Antimicrobial Resistance Gene List' module enables users to intuitively explore the associations between antimicrobial resistance genes and the microbial hosts using network visual analytics to gain biological insights. ResistoXplorer is publicly available at http://www.resistoxplorer.no/ResistoXplorer.

Introduction

Antimicrobial resistance (AMR) has emerged as a major threat to global public health and the economy [1]. Although antimicrobial resistance is a natural phenomenon, the misuse and overuse of antimicrobial agents in humans, animals and agriculture have accelerated the development and spread of antimicrobial resistance. Consequently, antimicrobials are becoming less effective in the prevention and treatment of infections. In addition, the extensive use of antibiotics has resulted in large amounts being released into the environment. This is a matter of increasing concern, not only in relation to antimicrobial resistance, but also for the potential impact on the ecology of human, animal, and environmental ecosystems [2].

Microbial communities have traditionally been studied using culture techniques. Although valuable, culture-based strategies have a low-throughput, and can provide an incomplete microbial profile due to the fact that, in a majority of environments, only a limited proportion of microorganisms are cultivable. Such limitations have been largely circumvented with recent advancements in short-read based high-throughput DNA sequencing (HTS) technologies, using either amplicon targeted or whole metagenomic shotgun approaches. Amplicon approaches can be used to map both taxonomic and resistome profiles. However, one of the limitations is that only genes recognized by the specific primers can be identified. Shotgun metagenomics, on the other hand, enables unbiased taxonomic and antimicrobial resistance information. Such technology has been shown to provide valuable insights into the natural history of antimicrobial resistance genes (ARGs) in humans and nature [3-5], as well as how the resistome develops in early life, and how it can be affected by the use of antibiotics [6, 7]. It is also contributing to the identification of ARGs that may cross

environmental and host boundaries [8], and providing unprecedented knowledge into the large reservoir of ARGs in the human, animal and environmental microbial communities [9-14]. Currently, resistome profiles from complex and diverse microbial metagenomes are primarily generated using whole metagenome shotgun sequencing in which the total DNA extracted from a microbial community is sequenced. The resulting DNA fragments can be analyzed using read or assembly-based approaches to characterize their resistome composition [15]. These derived sequencing datasets are both large and complex, causing considerable 'big data' challenges in downstream data analysis.

The main computational effort in resistome analysis of metagenomic datasets so far has focused on processing, classification, assembly and annotation of sequenced reads. This has led to the development of a number of excellent bioinformatic pipelines and tools for detecting and quantifying antimicrobial resistance genes in metagenomes [15-17]. However, there is still no clear consensus with regards to standard analysis pipelines and workflows for high-throughput analysis of AMR metagenomic resistome data [17, 18]. Nonetheless, the outputs from most of these pipelines can be summarized as a data table consisting of feature (ARGs) abundance information across samples, i.e. resistome profiles, along with their functional annotations and sample metadata. For most researchers, the fundamental challenge in data analysis can often be centered on how to understand and interpret the information in the abundance tables especially within the context of different experimental factors and annotations.

Resistome data analysis can be separated into four main categories: (i) composition profiling- to visualize and characterize the resistome based on approaches developed in community ecology such as alpha diversity, rarefaction curves or ordination analysis; (ii) functional profiling- to assign antimicrobial resistance genes into different functional categories (e.g. Drug class, Mechanism) to gain better insights regarding their collective functional capabilities; (iii) comparative analysis- to identify features having a significant differential abundance between studied conditions and (iv) integrative analysis- to integrate the resistome and taxonomic data to understand the complex interplay and potential associations between microbial ecology and AMR. The computational methods and approaches to perform such analysis are fairly diverse. The first category of analysis can be more straightforward to perform, but the last three are challenging.

First of all, the metagenomic abundance data is often characterized by differences in library sizes, which requires correction. To address this issue, researchers often employ two common normalization approaches prior to analysis: subsampling the reads in each sample to the same number (rarefying) or rescaling the total number of reads in each sample to uniform sum (using proportions). The former may entail the loss of valuable information, while the latter could lead to issues related to data compositionality [19]. In addition to uneven library sizes, metagenomics data are also characterized by sparsity, over-dispersion, zero inflation and skewed distribution [20, 21]. Such unique features make standard parametric tests and most methods established in other omics fields unsuitable to apply directly for comparative analysis. To address these challenges, non-parametric permutation-based approaches have been adopted to identify significant features in metagenomic abundance data [22, 23]. Even though robust, these approaches are constrained by lack of statistical power, as well as inefficiency to model confounding factors, and inability to accommodate intricate experimental designs.

Overall, development of statistical models that account for features of metagenomic data or use of methods to transform data to have distributions that fit standard test assumptions is recommended [24]. A variety of strategies have propelled empirical development in these directions. For instance, metagenomeSeq algorithm incorporates cumulative sum scaling normalization and a zero-inflated Gaussian (ZIG) mixture model to reduce false positives and improve statistical power for differential abundance analysis [25, 26]. It has also been demonstrated that algorithms developed for RNA-seq data such as edgeR and DESeq2, along with their respective normalization methods, outperform other approaches used for metagenomic abundance data [26-29]. These standard strategies are widely employed, but does not explicitly account for compositional nature of whole metagenomic sequencing data [30, 31]. To address this issue, several Compositional Data Analysis (CoDA) approaches to analyze sequencing datasets have been recently proposed [32, 33]. ALDEx2 and ANCOM are two CoDA methods that integrate log-ratio transformation to explicitly deal with data compositionality while performing differential abundance testing [34, 35]. Naturally, the choice of

methodology depends on the research question, and one may be interested in comparing results obtained by different analytical methods. Both standard and CoDA statistical approaches have been implemented as R packages. Although flexible, learning R in order to use these methods can be challenging for most clinicians and researchers.

In addition to the challenges described above, analytical steps have another complexity layer represented by the wide variety of reference databases to identify and characterize the antimicrobial resistance genes from metagenomic data. These differ considerably in the quality of data presented, as well as the scope of resistance mechanisms and the type of information provided [15]. This can influence the accuracy of in silico characterization of resistome data, which depends substantially on the comprehensiveness and quality of reference databases [15, 17, 24]. Typically, resistome profiles are analyzed by mapping ARGs either to their respective class of drugs to which they confer resistance (Class-level) or to their underlying molecular mechanism of resistance (Mechanism-level). Analyzing resistomes at such high level categories enable researchers to gain more biological, actionable, and functional insights together with a better understanding of their data. However, these functional levels and categories, along with their classification scheme, vary largely between the databases [18]. Additionally, depending upon the database, users need to manually collect and curate such information and then generate separate abundance tables for each functional level. Hence, collecting appropriate functional annotation information for hundreds of ARGs in resistomes for functional profiling and further downstream analysis can be confusing, arduous, time-consuming and error-prone.

Finally, some AMR reference databases may also provide information regarding the microbial host that harbor or carry these reference ARGs. Information about such relationship can be complex as one microbe can carry multiple ARGs and single ARGs can in turn be present across multiple microbes. To explore such intricated ‘multiple-to-multiple’ relations, one option is to use a network-based visualization method. This approach, coupled with suitable functional annotations and enrichment analysis support, would have the potential to provide better interpretation of AMR resistance mechanisms and inform on possible dissemination routes of antimicrobial resistance genes. It is straightforward to identify key players from a network perspective, for instance, by looking for those ARGs that are found in multiple microbes or by identifying those microbes that simultaneously contain multiple ARGs of interest. Currently, there are no web-based tools that provide such functionality.

Methods

ResistoXplorer is implemented based on Java, R and JavaScript programing languages. The framework is developed based on the Java Server Faces technology using the PrimeFaces (https://www.primefaces.org/) and BootsFaces (https://www.bootsfaces.net) component library. The network visualization uses the sigma.js (http://sigmajs.org/) JavaScript library. Additionally, D3.js (https://d3js.org/) and CanvasXpress (https://canvasxpress.org/) JavaScript libraries are utilized for other interactive visualization. All the R packages for performing back-end analysis and visualization are mentioned in the ‘About’ section of the tool. At the start of the analysis, a temporary account is created with an associated home folder to store the uploaded data and analysis results. All the analysis results will be returned in real-time. Upon completing their analysis session, users should download all their results. The system is deployed on a dedicated server with 4 physical CPU cores (Intel Core i5 3.4GHz), 8GB RAM and Ubuntu 18.04 LTS was used as the operation system. ResistoXplorer has been tested with major modern browsers such as Google Chrome, Mozilla Firefox, Safari and Microsoft Internet Explorer.

Results

To address the above mentioned statistical, visual and functional gaps as well as to meet recent advances in resistome data analysis, we have developed ResistoXplorer, a user-friendly, web-based, visual analytics tool to assist clinicians, bench researchers, and interdisciplinary groups working in the AMR field to perform exploratory data analysis on abundance profiles and resistome signatures generated from AMR metagenomics studies. The key features of ResistoXplorer include:

- Support of a wide array of common as well as advanced methods for composition profiling, visualization and exploratory data analysis;
- Comprehensive support for various data normalization methods coupled with standard as well as more recent statistical and machine learning algorithms;
- Support of a variety of methods for performing vertical data integrative analysis on paired datasets (i.e. taxonomic and resistome abundance profiles);
- Comprehensive support for ARG functional annotations along with their microbe and phenotypes associations based on data collected from >10 reference databases;
- A powerful and fully featured network visualization for intuitive exploration of ARG-microbe associations, including functional annotation enrichment analysis support.

Collectively, these features consist of a comprehensive tool suite for statistical, visual and exploratory analysis of data generated from AMR metagenomics studies. ResistoXplorer is freely available at http://www.resistoxplorer.no/ResistoXplorer.

ResistoXplorer consists of three main analysis modules. The first is the **ARG List** module that is designed to explore the functional and microbial host associations for a given list of antimicrobial resistance genes (ARGs) of interest. The second is the **ARG Table** module, which contains functions for analyzing resistome abundance profiles generated from AMR metagenomics studies. Lastly, the **Integration** module enables users to perform integrative analysis on the paired taxonomic and resistome abundance profiles to further explore potential associations coupled with novel biological insights and hypotheses. Figure 1 represents the overall design and workflow of ResistoXplorer.

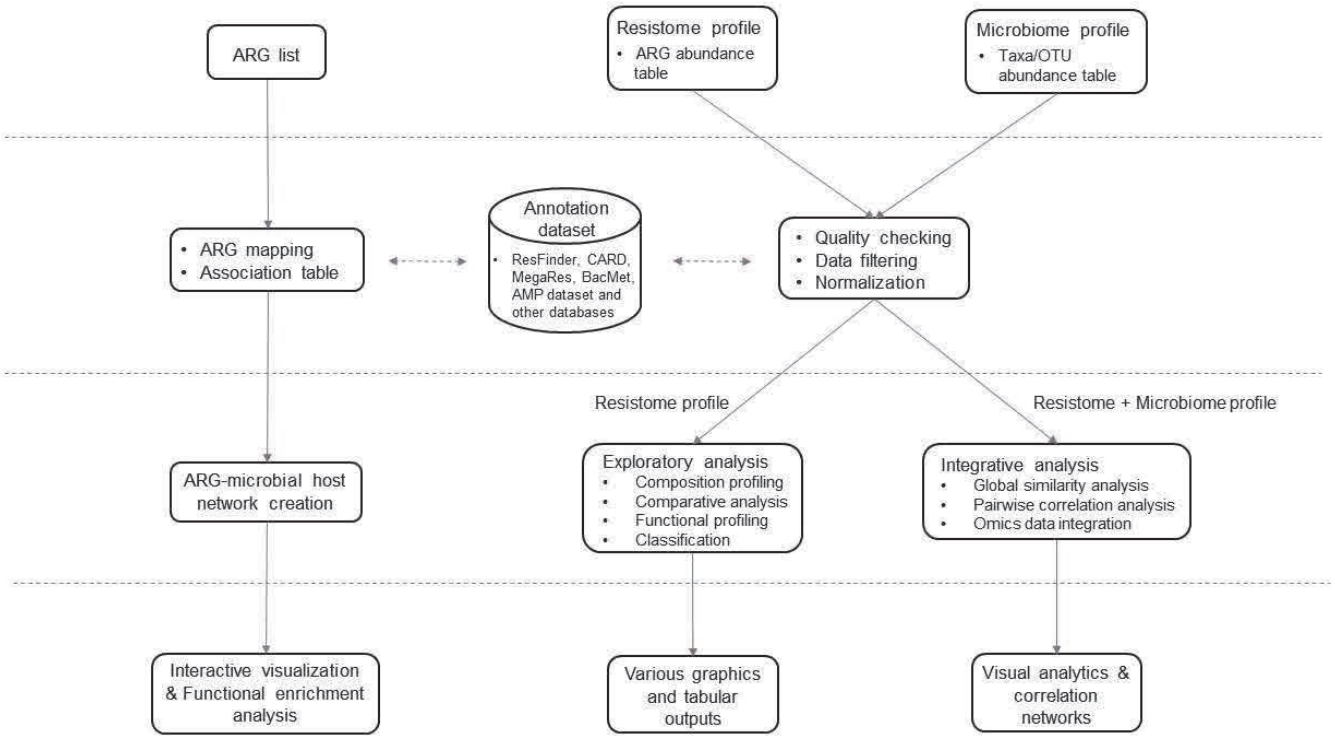


Figure 1: ResistoXplorer flow chart.

Conclusion

Whole metagenomic sequencing studies are providing unparalleled knowledge on the diversity of resistomes in the environment, animals and humans, and on the impact of interventions, such as antibiotic use [6, 7, 9-14]. The analyses are usually exploratory in nature, and require bioinformatic skills, thus increasing costs, and preventing full data exploration. Therefore, it is critical to assist researchers and clinical scientists in the field to easily explore their own datasets using a variety of approaches, in real-time and through interactive visualization, to facilitate data understanding and hypothesis generation. ResistoXplorer meet these requirements by offering comprehensive support for composition profiling, statistical analysis, integrative

analysis and visual exploration. ResistoXplorer will continuously be updated to follow the advancements in approaches for resistome analysis. We believe ResistoXplorer has the potential to find large applicability as a useful resource for researchers in the field of AMR.

Acknoweldgements

This work has been financed by the INDNOR and INTPART programs funded by The Research Council of Norway, grant numbers 273833 and 274867, and the Olav Thon foundation. Additionally, this work has been supported and used the computing resources at the University of Oslo.

References

1. United IACG. No Time to Wait–Securing the Future from Drug-resistant Infections. Report to the Secretary General of the Nations. 2019.
2. Kraemer SA, Ramachandran A, Perron GG. Antibiotic pollution in the environment: from microbial ecology to public policy. Microorganisms. 2019; 7:180.
3. Hu Y, Yang X, Qin J, Lu N, Cheng G, Wu N, et al. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. Nature communications. 2013; 4:2151.
4. D'Costa VM, King CE, Kalan L, Morar M, Sung WW, Schwarz C, et al. Antibiotic resistance is ancient. Nature. 2011; 477:457.
5. Nesme J, Cécillon S, Delmont TO, Monier J-M, Vogel TM, Simonet P. Large-scale metagenomic-based study of antibiotic resistance in the environment. Current biology. 2014; 24:1096-1100.
6. Gasparrini AJ, Wang B, Sun X, Kennedy EA, Hernandez-Leyva A, Ndao IM, et al. Persistent metagenomic signatures of early-life hospitalization and antibiotic treatment in the infant gut microbiota and resistome. Nature microbiology. 2019; 1-13.
7. Gibson MK, Wang B, Ahmadi S, Burnham C-AD, Tarr PI, Warner BB, et al. Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. Nature microbiology. 2016; 1:16024.
8. Pehrsson EC, Tsukayama P, Patel S, Mejía-Bautista M, Sosa-Soto G, Navarrete KM, et al. Interconnected microbiomes and resistomes in low-income human habitats. Nature. 2016; 533:212-216.
9. Xia Y, Zhu Y, Li Q, Lu J. Human gut resistome can be country-specific. PeerJ. 2019; 7:e6389.
10. Forslund K, Sunagawa S, Kultima JR, Mende DR, Arumugam M, Typas A, Bork P. Country-specific antibiotic use practices impact the human gut resistome. Genome research. 2013; 23:1163-1169.
11. Munk P, Andersen VD, de Knegt L, Jensen MS, Knudsen BE, Lukjancenko O, et al. A sampling and metagenomic sequencing-based methodology for monitoring antimicrobial resistance in swine herds. Journal of Antimicrobial Chemotherapy. 2016; 72:385-392.
12. Pal C, Bengtsson-Palme J, Kristiansson E, Larsson DJ. The structure and diversity of human, animal and environmental resistomes. Microbiome. 2016; 4:54.
13. Forsberg KJ, Patel S, Gibson MK, Lauber CL, Knight R, Fierer N, et al. Bacterial phylogeny structures soil resistomes across habitats. Nature. 2014; 509:612-616.
14. Hendriksen RS, Munk P, Njage P, Van Bunnik B, McNally L, Lukjancenko O, et al. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nature communications. 2019; 10:1124.
15. Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. Nature Reviews Genetics. 2019:1.
16. McArthur AG, Wright GD. Bioinformatics of antimicrobial resistance in the age of molecular epidemiology. Current opinion in microbiology. 2015; 27:45-50.
17. Hendriksen RS, Bortolaia V, Tate H, Tyson G, Aarestrup FM, McDermott P. Using Genomics to Track Global Antimicrobial Resistance. Frontiers in public health. 2019; 7:242.
18. Lakin SM, Dean C, Noyes NR, Dettenwanger A, Ross AS, Doster E, et al. MEGARes: an antimicrobial resistance database for high throughput sequencing. Nucleic acids research. 2016; 45:D574-D580.
19. Filzmoser P, Hron K, Reimann C. Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Science of the Total Environment. 2009; 407:6100-6108.

20. Li H. Microbiome, metagenomics, and high-dimensional compositional data analysis. Annual Review of Statistics Its Application. 2015; 2:73-94.
21. Calle ML. Statistical analysis of metagenomics data. Genomics informatics. 2019:17.
22. Paulson JN, Pop M, Bravo HC. Metastats: an improved statistical method for analysis of metagenomic data. Genome Biology. 2011; 12:P17.
23. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome biology. 2011; 12:R60.
24. Bengtsson-Palme J, Larsson DJ, Kristiansson E. Using metagenomics to investigate human and environmental resistomes. Journal of Antimicrobial Chemotherapy. 2017; 72:2690-2703.
25. Paulson JN, Pop M, Bravo HC. metagenomeSeq: Statistical analysis for sparse high-throughput sequencing. Bioconductor package. 2013; 1:63.
26. Pereira MB, Wallroth M, Jonsson V, Kristiansson E. Comparison of normalization methods for the analysis of metagenomic gene abundance data. BMC genomics. 2018; 19:274.
27. Jonsson V, Österlund T, Nerman O, Kristiansson E. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. BMC genomics. 2016; 17:78.
28. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014; 15:550.
29. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139-140.
30. Quinn TP, Erb I, Richardson MF, Crowley TM. Understanding sequencing data as compositions: an outlook and review. Bioinformatics. 2018; 34:2870-2878.
31. Gloor GB, Reid G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. Canadian journal of microbiology. 2016; 62:692-703.
32. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Frontiers in microbiology. 2017; 8:2224.
33. Quinn TP, Erb I, Gloor G, Notredame C, Richardson MF, Crowley TM. A field guide for the compositional analysis of any-omics data. GigaScience. 2019; 8:giz107.
34. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome. 2014; 2:15.
35. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microbial ecology in health disease. 2015; 26(1):27663.

A novel algorithm for extracting common modular communities from Dual Networks.

Pietro Hiram Guzzi 1, Giuseppe Tradigo 2 , Pierangelo Veltri 1

1) Department of Medical and Surgical Sciences, University of Catanzaro.
2) eCampus Novedrate, Como, Italy

Email of Corresponding author: hguzzi@unicz.it

Abstract

Networks-based models are continuously used in computational biology to model and analyse. More recently, many different works have shown that standard models may fail to capture some aspects of the investigated scenarios. Therefore, more enriched models, such as dual networks may be used to better analyse data. A dual network uses a pair of network sharing the same nodes. One network, said physical, represents binary associations among nodes using unweighted edges, while the other is an edge-weighted one where weights represent the strength of the associations among nodes. In a previous job, we focused on to find the Densest Connected Subgraph (DCS), while we here extend such work to find common communities having maximum modularity following the Louvain approach and we propose a novel heuristics to solve it. We tested the proposed algorithm on some biological networks. Results demonstrate that our approach is efficient and can extract meaningful information from dual networks.

Introduction

In biology and medicine, there exist many approaches based on the use of graphs for modelling and analysis data produced by high-throughput platforms [1,2]. Many approaches are based on the use of a single graph to model data. Usually, a node of such a graph represents the modelled entities, while edges represent associations among them. After that, graphs have been obtained from raw data, many different algorithms are used to study networks properties or to compare many networks to identify biologically relevant information related to topological properties [3,4,5,**6]**. For instance, considering networks of proteins (aka protein-protein interaction networks), community extraction algorithms are used to find protein complexes, while network comparison algorithms are used to analyse evolutionarily conserved mechanisms.

Recently it has been shown [7] that the use of a pair of related graphs representing two different views of the same scenario may help to detect hidden properties which are not detected by single network-based models. In this work, we focus on the use of a pair of graphs, also known as dual networks, i.e. two graphs with the same vertices set and two different edges sets. One graph, called a physical graph, has unweighted edges, while the second one, called a conceptual graph, has weighted edges. Dual networks have natural applications whenever it is needed to model two kinds of relations among the same set of nodes. The two networks represent physical and conceptual interactions. The literature contains many examples of the use of dual networks for mining biological data spanning different scenarios, such as the analysis of genetic variants or the correlation of genetic interactions and gene-coexpressions [8,9].

In Dual networks, there is the interest for extracting common subgraphs of the two networks that have some given properties. For instance, common subgraphs that are maximally dense in the conceptual network and connected in the physical one, i.e. densest common subgraphs, have been studied in [10] A variant of this problem, based on the detection of the Top-k Densest Connected subgraphs, i.e. a set of k subgraphs having the largest density in the conceptual network which are also connected in the physical network has been presented in [11]

We here focus on the extraction of common subgraphs having maximum modularity on the conceptual network that is also connected in the physical one. Finding such subgraphs in dual networks is still a challenging problem [12]; thus we propose a heuristic based on a modification of the local network alignment problem and we propose a novel algorithm to solve it. The proposed heuristic is based on two main steps: (i) initially we integrate both networks onto a single graph (Weighted Alignment Graph – WAG), (ii) then we apply the Louvain algorithm [13] to extract modular communities from this graph. The nodes of each subgraph of the WAG induces a connected graph on the physical graph. Therefore the subgraphs having maximum modularity are connected. Therefore they are solutions to the initial problem. To the best of our knowledge, literature do not contain similar approaches, since Wu et al., focus on densest connected subgraphs. We provide an implementation of our approach showing a case study on biological networks confirming the effectiveness of our approach

Methods

The proposed algorithm receives as input two networks of the same dual network. In the first step, the two networks are merged together into a single Weighted Alignment Graph (WAG). Each node of the WAG represents a pair of related nodes of the input network. Correspondences among nodes are given as input, and in the simplest case, the mapping merges together identical nodes of the physical and conceptual network. Then the algorithm completes the WAG by inserting weighted edges are inserted considering the two input networks. Then the Louvain algorithm is used for extracting modular communities of the alignment graph. Each sub-graph of the alignment graph represents a connected sub-graph of the unweighted networks. Therefore, it is a densest connected sub-graph for the dual network.

Algorithm 1:

Input: A conceptual Network $G_c=(W,E)$, and a physical network $G_p=(V,E)$

Input: a mapping among nodes of both network M, a distance threshold **delta, Q parameter for Louvain algorithm**

Output: a list of communities C_{com}

Begin:

1 : WAG <- BuildAlignmentGraph (G_c,G_p,M,delta, Q)

2: C_{com} <- Louvain(WAG)

3: return C_{com}

End

During the first step of the algorithm, the two networks are merged together into a single weighted alignment graph. The algorithm uses three parameters: (i) a mapping representing the correspondences among nodes, (ii) a distance threshold delta that represents the maximum distance threshold that two nodes should have in the physical networks, and (iii) the Q parameter representing the modularity value of the Louvain algorithm. In the first step, the algorithm merges the input networks into a single alignment graph following the approach described in [10].

In this way, each connected sub-graph of this graph represents a connected subgraph into the physical network.

In the second step, we use the Louvain algorithm [13] for extracting modular communities from the WAG. The Louvain algorithm is a community detection algorithm based on a greedy strategy. It aims to optimise the **modularity** of the network, i.e. a measure of the density of the edges inside communities with respect to edges

outside communities. The Louvain algorithm first finds small communities by looking at the optimisation of the modularity on a local sale considering all the nodes. Then small local communities are merged into a single node, and the first step is repeated considering as a single node each community. The cycle is repeated until the modularity increases.

Results and Discussion

We tested the performances of our algorithm on some dual networks representing biological scenarios to extract modular communities. We downloaded data from the STRING database [14] that contains data about protein and their interactions. Each node represents a protein, and each edge takes into account the reliability of the interaction between two proteins with a value in the interval 0-1. Therefore, we obtained two networks: a conceptual network, which represents the strength of associations among proteins; a physical network, which stores the binary interactions among proteins. We obtained two networks having 19.354 nodes and 5.879.727 edges. We ran our approach by obtaining a set of communities. Finally we evaluated the biological significance of our approach. Preliminary results confirm that communities contains some enriched function with respect to ground distribution.

Conclusion

Dual networks are a pair of graphs composed by one unweighted graph (physical network) and a second edge-weighted (conceptual network). We presented a novel algorithm for obtaining modular connected subgraphs that induce a maximum modularity subgraph in the conceptual network being also connected in the physical network. We formalised the problem, and we proposed a solution based on graph alignment theory. Finally, we proposed a possible solution and presented a set of experiments, which demonstrate the effectiveness of our approach.

Acknoweldgements

Authors thank Emanuel Salerno for working in this project by implementing some software module.

References

[1] Di Martino, M. T., Guzzi, P. H., Caracciolo, D., Agnelli, L., Neri, A., Walker, B. A., et al., Integrated analysis of microRNAs, transcription factors and target genes expression discloses a specific molecular architecture of hyperdiploid multiple myeloma. Oncotarget, 6(22), 19132. (2015)

[2] Cannataro, M., Guzzi, P. H., & Veltri, P. (2010). Protein-to-protein interactions: Technologies, databases, and algorithms. ACM Computing Surveys (CSUR), 43(1), 1-36.

[3] C. Clark and J. Kalita, “A comparison of algorithms for the pairwise alignment of biological networks,” Bioinformatics, vol. 30, no. 16, pp. 2351–2359, 2014.

[4] Guzzi, P. H., & Milenković, T. (2018). Survey of local and global biological network alignment: the need to reconcile the two sides of the same coin. Briefings in Bioinformatics, 19(3), 472-481.

[5] Cannataro, Mario, Pietro H. Guzzi, and Pierangelo Veltri. "IMPRECO: Distributed prediction of protein complexes." Future Generation Computer Systems 26.3 (2010): 434-440.

[6] Nelson, W., Zitnik, M., Wang, B., Leskovec, J., Goldenberg, A., & Sharan, R. (2019). To embed or not: network embedding as a paradigm in computational biology. Frontiers in genetics, 10, 381.

[7] Wu, Y., Zhu, X., Li, L., Fan, W., Jin, R., & Zhang, X. (2016). Mining dual networks: models, algorithms, and applications. ACM Transactions on Knowledge Discovery from Data (TKDD), 10(4), 1-37.

[8] S. Tornow and H. Mewes, “Functional modules by relating protein interaction networks and gene expression,” Nucleic Acids Research, vol. 31, no. 21, pp. 6283–6289, 2003.

[9] P. C. Phillips, “Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems,” Nature reviews. vol. 9 no 11 pp 885-867 (2008)

[10] Guzzi PH, Tradigo G, Salerno E, Veltri P. (2020) Extracting Dense and Connected Communities in Dual Networks: an Alignment Based Algorithm. IEEE Access DOI: 10.1109/ACCESS.2017.DOI

[11] Dondi, R., Guzzi, P. H., & Hosseinzadeh, M. M. (2020). Top-k Connected Overlapping Densest Subgraphs in Dual Networks. arXiv preprint arXiv:2008.01573.

[12] Marcus, D. A. (2020). Graph theory (Vol. 53). American Mathematical Soc..

[13] Seifikar, Mahsa, Saeed Farzi, and Masoud Barati. "C-Blondel: An Efficient Louvain-Based Dynamic Community Detection Algorithm." IEEE Transactions on Computational Social Systems 7.2 (2020): 308-318.

[14] Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al., . (2016). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids research, gkw937.

Zero-imputation in 16S rRNA gene studies: do we need it?

Giacomo Baruzzo[*,1], Ilaria Patuzzi[*,2], Barbara Di Camillo[1,3]

[1] Department of Information Engineering, University of Padua, Padua, Italy

[2] Microbial Ecology Unit, Istituto Zooprofilattico Sperimentale delle Venezie, Padua, Italy

[3] CRIBI Biotechnology Centre, University of Padua, Padua, Italy

* These authors contributed equally to this work.

Email of Corresponding author: barbara.dicamillo@unipd.it

Abstract

16S rRNA-gene sequencing allows characterizing bacterial communities, achieving the taxonomic profiling of the bacterial population and provides a valuable tool to study bacteria and their role in different health and environmental scenarios. The analysis of such sequencing data, however, brings several methodological issues that need to be addressed to obtain reliable biological conclusions. Among these, 16S count data are very sparse, with many null values reflecting species that are present but got unobserved during the experimental process. However, current data workflows do not include a step to recover that lost information.

In this work, we evaluate, for the first time, the effect of introducing in the 16S data workflow a new pre-processing step -zero-imputation- to recover this lost information. Due to the lack of zero-imputation methods specifically designed for 16S count data, we considered a set of zero-imputation strategies available for other frameworks and benchmarked them using several in-silico 16S count data. Additionally, we assess the role of zero-imputation compared with count normalization, and their combined effect.

The results show that properly performing zero-imputation can improve the quality of 16S data analyses workflow ultimately leading to more robust and accurate results.

Introduction

Today, microbial community profiling is almost uniquely performed by sequencing the DNA content of the community by means of Next-Generation Sequencing technologies, mainly through shotgun sequencing and 16S rRNA gene sequencing (16S rDNA-seq). The latter is less cost- and resource-demanding, thus achieving an increasing growth in election rate as preferred methodology to perform microbiome studies. After sequencing, 16S microbial community data are typically summarized into large matrices, where the columns represent samples and the rows contain operational taxonomic unit (OTU) [1] or amplicon sequence variant (ASV) [2] count values, that represent (broadly speaking) bacteria types.

As pointed out in Brooks et al. [3], the experimental procedure introduces many biases that affect the reliability of the values reported in the OTU/AVS matrices. First, amplification introduces several biases due to the unequal primers efficiency within and across genomes. Second, different prokaryotic species have different number of 16S replicons with varying degree of sequence variability. Third, after the sequencing step the total microbial community in each sample is represented by very different amount of sequences (i.e. library sizes), sometimes differing by several orders of magnitude. All these elements affect the measured composition of the bacteria population in the OTU/ASV matrix, resulting in altered abundances and undetected species [3].

To mitigate the effect of some of the above biases and to avoid misleading results, data should be treated prior to perform downstream analysis. Usual analysis workflow starts with a pre-processing step called normalization. Normalization is the process of eliminating artifactual systematic biases between samples, making possible a direct comparison of species abundance between them or between groups of them. However, normalization cannot solve or even diminish data biases linked to undetected species and the high sparsity (70-95% of zero values) of sequencing count data.

The present work had the main objective to test and measure the effects of preserving low abundance OTU/ASV information, performing an additional pre-processing step for lost-information recovery (zero-imputation). To best of our knowledge, only benchmarks considering the normalization step are nowadays available in the literature, whereas no effort was done so far to test the potential benefits of introducing the

zero-imputation step. In the present work, a collection of six normalization and six zero-imputation approaches was tested and combined in 48 pre-processing pipelines, providing the first results about the effects of introducing the zero-imputation step in 16S data analysis workflow.

Methods

Datasets

In this study, synthetic data were simulated using metaSPARSim [4] for their ability to provide a known ground truth in terms of sample composition and biological (missing species) vs. technical (undetected species) zero, thus enabling the assessment of the methods performance. Datasets were simulated using three simulation settings available in the metaSPARSim simulator, describing a range of biological and technical scenarios (Table 1). ***Scenario1*** mimics a "medium difficulty" scenario characterized by a limited level of sparsity and low variability among replicates. ***Scenario2*** describes low sparsity but high biological variability scenario, having characteristics similar to data from the Human Microbiome Project [5,6]. ***Scenario3*** mimics a high budget experiment with low multiplexing level and thus with little sequencing loss of information. This latter setting is meant to test the effects of possible over-imputation.

	Scenario1	Scenario2	Scenario 3
Type	Animal gut	Human	Food
Groups	14	8	12
Samples	140	80	120
Replicates	10	10	10
Features	3326	758	1140
Sequencing depth (range)	16347-995050	2763-97612	30165-293285
Sparsity level	75.56%	67.91%	94.34%

Table 1. Simulated datasets used in this study

Bioinformatics tools

Normalization tools were selected among the most widely used or the most recent and promising now available. Due to the lack of zero-imputation methods specific for 16S data, tools from other fields were considered, including single cell RNA-seq (scRNA-seq) data, microarray data and matrix completion. A total of six normalization methods (***Total Sum Scaling (TSS)*** [7], ***Cumulative Sum Scaling (CSS)*** [8], ***edgeR*** [9], ***DESeq2*** [10], ***scran*** [11] and ***GMPR*** [12]) and six zero-imputation tools (***DrImpute*** [13], ***scImpute*** [14], ***LLSimpute*** [15], ***Low-Rank*** [16] and ***zCompositions*** [17], the latter one used in both "CZM" and "SQ" modes) were included in this study. The tools were combined to form **48 different pre-processing pipelines** (6 normalization-only, 6 imputation-only, 36 normalization+imputation).

Evaluation criteria

The adopted benchmarking framework, represented in Figure 1, involved ground truth data jointly with raw and pre-processed data. metaSPARSim was used to simulate taxa abundances sampled from the biological niche of interest before sequencing (ground truth) and after sequencing (raw data). Then, raw matrices were pre-processed with all the 48 pipelines. Finally, ground truth and pre-processed data were used to assess pipelines performance as explained in the following.

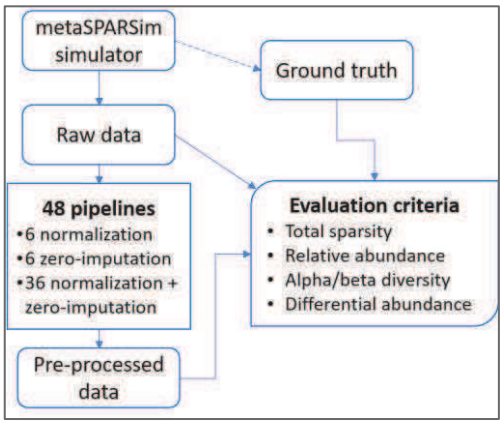


Figure 1. Benchmarking framework

Total sparsity

As a first assessment metric, the pipelines were evaluated for their ability to reproduce original data sparsity. Each pipeline results were compared to the ground truth in terms of percentage of zero counts, i.e. the ratio between the number of zero counts and the total number of count matrix entries.

Relative abundance profile

To assess the ability of recovering "true" (ground truth) proportional abundances, two different metrics were considered: ***Symmetric Mean Absolute Percentage Error*** (***SMAPE***) and ***Aitchison's distance*** [18]. The first one is a quantitative metric based on percentage (or relative) errors and it was chosen as an alternative to the classical relative error because of its suitability to heavily sparse data. The second measure, i.e. Aitchison's

distance, is a distance measure that accounts for the compositional nature of sequencing data [18]. One side Mann-Whitney paired U-test (p-value < 0.05 after Benjamini-Hochberg FDR correction [19]) was used to identify pipelines that achieved SMAPE and Aitchison's distance significantly lower compared to the ones between ground truth and raw data. In addition, effect size calculation was used to measure the magnitude of possible significant differences between distributions.

Impact on bacterial diversity

One of the most important aspects to look at when performing a microbiome analysis is the population diversity, usually measured with the so-called diversity indices, measuring the species diversity in sites belonging to a niche (alpha diversity) and the differentiation among those sites (beta diversity). In this work, **five alpha and two beta diversity measures were considered** to assess the effect of each pre-processing pipeline on microbial community composition, with the aim of identifying the pipelines that would preserve the most the real structure of the golden standard data.

In terms of alpha diversity, richness indices (***observed richness***), evenness indices (***Pielou index***) and diversity indices (***Shannon entropy*** [20], ***inverse Simpson index*** [21] and ***Tail statistic*** [22]) were used. The impact of different pipelines were measured by looking at the consequences in terms of detected differences in alpha diversity values distributions across group of samples (one-sided Mann-Whitney U-test, p-value <0.05), as done in practice when using alpha diversity indices. Then, we performed the same procedure on ground truth data, thus measuring the true differences in terms of alpha diversity present in the data. Last, we computed the percentage of detected differences that were wrong (i.e. not detected in ground truth data) and used such percentage as a measure of error.

In terms of beta diversity, ***Whittaker beta diversity*** [23] and ***Bray-Curtis dissimilarity*** [24] were used. The first dissimilarity was then used to build a distance matrix on which Non-metric Multidimensional Scaling (NMDS) dimensionality reduction was performed to assess spatial distribution of samples, whereas Whittaker dissimilarity values were graphically represented using heatmaps.

Differential abundance analysis

Differential abundance (DA) analysis is a fundamental step in each microbiome study. In this work, we chose to perform this analysis using a non-parametric ***Mann-Whitney U-test***. The Mann-Whitney test was preferred in this evaluation framework in order not to add further potential biases to the results, since each existing DA tool has its own assumptions and underlying model.

In particular, Mann-Whitney U-test was performed to identify DA features (p-value <0.05 after Benjamini-Hochberg FDR correction) across different groups in each scenario, running the analysis on the ground truth, the raw and the pre-processed datasets. The consistency between results on the ground truth and on each of the other datasets was measured using ***Jaccard index*** [25]:

$$I^{ab}_{Jaccard} = \frac{GT_{ab} \cap D_{ab}}{GT_{ab} \cup D_{ab}}$$

where GT_{ab} is the set of DA features for conditions a and b of ground truth data and D_{ab} is the correspondent set of features identified as DA in the generic raw or pre-processed dataset D.

For each dataset, we obtained different Jaccard index values (one for each pairwise group-group comparison). To test for (possible) improvement in DA consistency obtained using raw data or pre-processed data with respect to ground truth, a one-sided, paired Mann-Whitney test was performed between Jaccard index values obtained using raw and pre-processed datasets.

Results

Total sparsity

In terms of sparsity, LLSimpute, LowRank and zCompositions (both SQ and CZM), tended to heavily underestimate data sparsity in all the simulated datasets, recovering the majority (LLSimpute and LowRank) or also the totality (zCompositions) of zero counts in combination with all normalization approaches. On the contrary, scImpute and DrImpute pipelines recreated true sparsity very well, slightly overestimating or underestimating the true zero counts, depending on the dataset. As an example, Table 2 shows the results in terms of sparsity for Scenario 2. Please note that normalization-only pipelines do not alter the sparsity of the data.

Relative abundance profile

To evaluate the ability of different pipelines in recovering data information, SMAPE and Aitchison's distance between the ground truth and the different pipelines outputs were calculated on sample proportional abundances. As for total sparsity metric, normalization-only pre-processing pipelines inherently did not act on the present metrics, thus obtaining values equal to the raw matrix ones. To summarize the results in term of imputation choice, first the median of SMAPE and Aitchison's distance were calculated for each pipeline as an aggregate measure, and then pipelines based on the same imputation method were combined, computing mean and standard deviation of such aggregated measure. As an example, Table 3 shows the results for Scenario 2. scImpute and DrImpute pipelines obtained the best results in terms of SMAPE on Scenario 1, with scImpute performing well even in Scenario 2. Regarding normalization-only pipelines, they always performed better than LLSimpute, LowRank, and zCompositions in terms of SMAPE. In terms of Aitchison's distance, some pipelines containing DrImpute and zCompositions achieved a better performance than raw/normalized-only data on Scenario 1 and 2. Normalization-only pipelines achieved also a lower Aitchison's distance compared to LLSimpute and LowRank, while they resulted in a higher Aitchison's distance compared to some zCompositions pipelines in Scenario 1 and 2. A special case was observed for Scenario 3, where all the pipelines involving zero-imputation performed worse than raw/normalized-only data in terms of both SMAPE and Aitchison's distance.

	Sparsity (%)	
Data/Pipeline	**Mean**	**SD**
Ground truth	56.61	0
Raw & Normalization-only	67.91	0
scImpute	55.84	0.004
DrImpute	42.32	0
LLSimpute	23.31	1.63
LowRank	2.09	1.82
zCompositions_SQ	0	0
zCompositions_CZM	0	0

Table 2. Count matrix sparsity for Scenario 2. Pre-processed datasets results were aggregated according to the zero-imputation method included in the pipeline; for each, the mean and standard deviation over different normalizations are shown.

Imputation	**SMAPE** Mean (SD)	**Aitch distance** Mean (SD)
Raw - None	15.49 (0)	23.75 (0)
DrImpute	28.35 (2.35)	24.28 (2.27)*
scImpute	12.06 (0.01)**	25.11 (0)
LLSimpute	79.95 (1.84)	81.89 (0.68)
LowRank	70.10 (0.44)	44.26 (12.24)
zCompositions_CZM	70.16 (1.46)	28.68 (4.65)*
zCompositions_SQ	71.68 (0.55)	23.36 (0.18)*

Table 3. SMAPE and Aitchison's distance between the ground truth and different pipelines for Scenario 2. Imputation strategies that achieve a statistically significant improvement with any normalization method are indicated with "**". Imputation strategies that achieve a statistically significant improvement only with some normalization methods are indicated with "*".

Impact on bacterial diversity

Alpha diversity indices were used to evaluate the impact of the different pipelines in terms of consequences on a statistical testing procedure (see Methods). Figure 2 shows a subset of the results on Scenario 2.

In terms of richness, scImpute pipelines achieved comparable or better results than normalization-only pipelines. The remaining zero-imputation pipelines always performed worse than normalized-only data, except for the good performance of DrImpute pipelines in Scenario1.

In terms of evenness (Pielou index), scImpute pipelines showed good results in Scenario1 and 2, while they performed worse than normalized-only data in Scenario 3. Again, the remaining imputation pipelines always performed worse than normalized-only data, except for DrImpute pipelines in Scenario1.

In terms of Tail diversity index, scImpute, DrImpute and zCompositions_SQ pipelines were the only zero-imputation pipelines performing comparable or better than normalization-only pipelines in Scenario 1 and 2. None of the zero-imputation pipelines improved the results in Scenario 3 compared with normalization-only pipelines.

In terms of Shannon and iSimpson indices, zCompositions pipelines performed comparable to normalization-only pipelines in all the tested datasets. scImpute and DrImpute achieved variable performance, being comparable to normalization-only pipeline in some dataset and slightly worst in others.

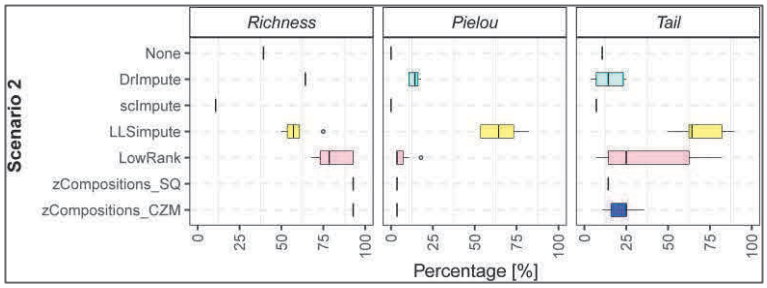


Figure 2. Results on alpha diversity indices (Richness, Pielou and Tail) for Scenario 2 in terms of percentage of group-group comparisons disagreeing with the ground truth (the lower the better). Results are aggregated according to the imputation method used in the pipelines.

As previously introduced, beta diversity indices are used to measure dissimilarity between samples, in order to collect the ones that resemble to each other and divide the whole into different groups. With this aim, Bray-Curtis dissimilarity and Whittaker index were calculated to show different aspects of the considered matrix. Figure 3 and 4 show the results of these two metrics of some pipelines on Scenario 2.

Both in terms of Bray-Curtis dissimilarity and Whittaker index, scImpute and DrImpute pipelines performed comparable or better than normalization-only pipelines in re-join sample belonging to the same group and dividing different groups. Pipelines including the remaining imputation methods performed worse than normalization-only pipelines.

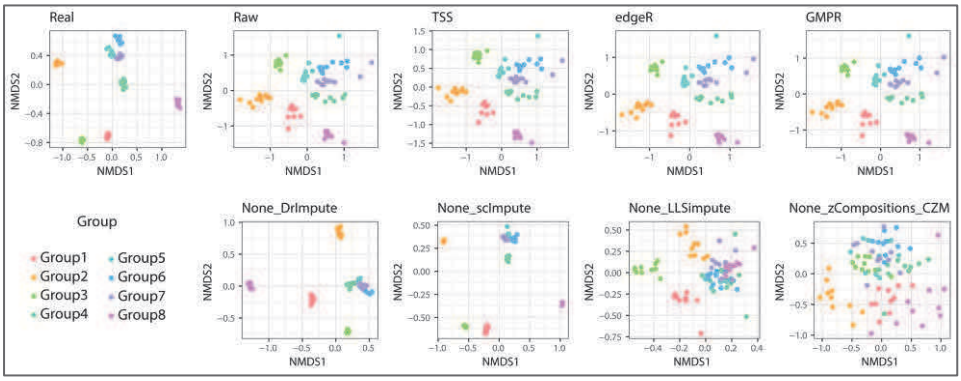


Figure 3. Bray-Curtis dissimilarity on Scenario 2. For some example pipelines, the plots show the Non-metric Multidimensional Scaling (NMDS) dimensionality reduction on beta diversity values. The top-left plot shows the true structure of the data in terms of beta diversity. The colours encode the different groups within the datasets.

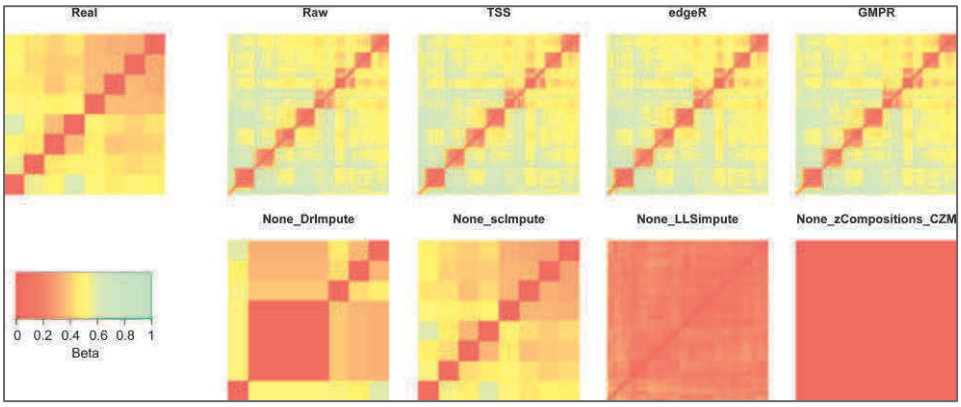


Figure 4. Whittaker dissimilarity on Scenario 2. For some example pipelines, the heatmaps show the beta diversity values computed from each pair of samples within the dataset. The top-left heatmap shows the true structure of the data in terms of beta diversity.

Differential abundance analysis

A test for change in differential abundance analysis results was performed as a measure on each pipeline ability to recover original data structure. A subset of the results for Scenario 2 is shown in Figure 5.

About normalization-only pipelines, they always improved the results compared to use raw data on Scenario 1 and 2 except for scran normalization, while on Scenario 3 scran was the only normalization method that improved the results of DA analysis compared with using raw data. scImpute pipelines always improved the results compared to raw/normalized-only data on all the datasets. DrImpute pipelines performed better than normalization-only pipelines in datasets 1 and 3. In the large majority of pipelines and datasets, the other imputation methods did not improve the results of DA analysis compared with the use of raw data.

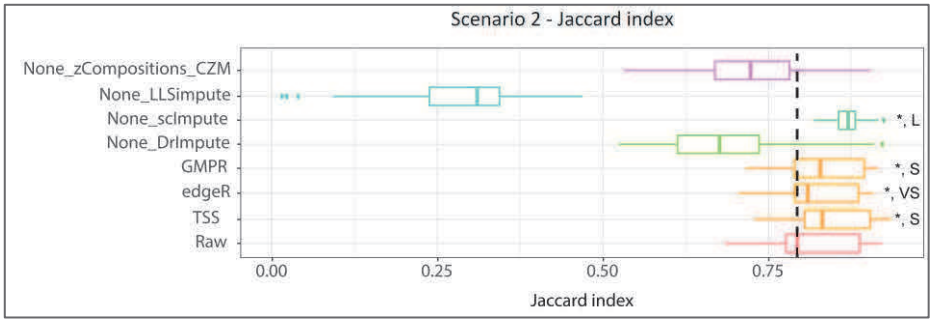


Figure 5. Boxplot of Jaccard indices on Scenario2 for some example pipelines. Distributions of Jaccard index values that result statistically lower than Jaccard index values calculated on raw data are indicated with the symbol "*", followed by the interpretation of Cohen's d effect size (VS: very small, S: small, L: large). The vertical dashed line indicates the median Jaccard index value of raw data.

Discussion and conclusion

Our analysis indicates that a properly performed zero-imputation can improve the results of a 16S data analysis workflow in terms of sparsity, relative abundance profile, bacteria diversity (alpha/beta) analysis and differential abundance analysis. For all the benchmark datasets and assessment scores used in this work, it always exists at least one zero-imputation pipeline that performed comparably or better than using raw/normalized-only data. The only exception is represented by Scenario 3, were no zero-imputation pipeline improved the results in terms of relative abundance profile compared with normalized-only data.

Zero-imputation showed very often higher impact than normalization in improving the quality of the results. The tested normalization methods showed comparable results among each other, with no normalization method that clearly outperformed the others. This fact is confirmed even when normalization was applied prior to zero-imputation, resulting in negligible differences in the final quality of the pre-processed data.

Moreover, the results highlighted that the choice of the imputation tool has a major role in the quality of pre-processed data. Indeed, some zero-imputation tools performed even worse than using raw or normalized-only data, and the performance were variable across the 3 test datasets. The variable and suboptimal performance of some tools are not surprising if we consider that none of the tools used in this study was designed specifically for imputation of 16S data. In particular, zCompositions, LLSimpute and Low-Rank recover the totality/majority of zeros, thus not differentiating biological zeros (real zeros) from technical zeros (missing data), generally producing worse results compared to using raw/normalized-only data. Not surprisingly, the best performing imputation tools were the ones developed for single cell RNA-seq data (i.e. scImpute and DrImpute), suggesting that some characteristic/biases of 16S can be modelled/tackled as the ones present in scRNA-seq data. This also suggests that zero-imputation tool specifically designed for 16S data would probably achieve even better performance, thus encouraging the development of such tools and further studies about zero-imputation of 16S count data.

References

1. Sneath PHA, Sokal RR, others. Numerical taxonomy. The principles and practice of numerical classification. 1973;
2. Bolyen E, Rideout JR, Dillon MR, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 2019; 37:852–857
3. Brooks JP, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC microbiol. 2015; 15.1:1-14.
4. Patuzzi I, et al. metaSPARSim: a 16S rRNA gene sequencing count data simulator. BMC Bioinformatics 2019; 20:1–13
5. Consortium THMP. Structure, Function and Diversity of the Healthy Human Microbiome. Nature 2013; 486:207–214
6. Consortium THMP. A framework for human microbiome research. Nature 2012; 486:215–221
7. Marioni JC, Mason CE, Mane SM, et al. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008; 18:1509–1517
8. Paulson JN, et al. Differential abundance analysis for microbial marker-gene surveys. Nat. Methods 2013; 10:1200–1202
9. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010; 26:139–140t. Methods 2013; 10:1200–1202
10. Love MI, et al. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:1-21
11. Lun ATL, et al. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016;17:1–14
12. Chen L, Reeve J, Zhang L, et al. GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. PeerJ 2018; 6:e4600
13. Gong W, et al. DrImpute: imputing dropout events in single cell RNA sequencing data. BMC Bioinform 2018; 19:220
14. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat. Commun. 2018; 9:997
15. Kim H, Golub GH, Park H. Missing value estimation for DNA microarray gene expression data: Local least squares imputation. Bioinformatics 2005; 21:187–198
16. Chen C, He B, Yuan X. Matrix completion via an alternating direction method. IMA J. Numer. Anal. 2012; 32:227–245
17. Palarea-Albaladejo J, Martín-Fernández JA. ZCompositions - R package for multivariate imputation of left-censored data under a compositional approach. Chemom. Intell. Lab. Syst. 2015; 143:85–96
18. Aitchison J, Barceló-Vidal C, et al. Logratio analysis and compositional distance. Math. Geol. 2000; 32:271–275
19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 1995; 289–300
20. Shannon CE. A mathematical theory of communication. Bell Syst. Tech. J. 1948; 27:379–423
21. Simpon EH. Measurement of Diversity. Nature 1949; 163:688
22. Li K, Bihan M, Yooseph S, et al. Analyses of the microbial diversity across the human microbiome. PLoS One 2012; 7:e32118
23. Whittaker RH. Vegetation of the Siskiyou Mountains, Oregon and California. Ecol. Monogr. 1960; 30:279–338
24. Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. Ecol. Monogr. 1957; 27:325–349
25. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull. del la Société Vaudoise des Sci. Nat. 1901;37:547–579

Novel multi- input logic gates for Synthetic Biology: analysis of the interaction between transcription factors and CRISPR interference

Massimo Bellato[1,2], Lorenzo Pasotti[2], Giuseppe Serio[2], Michela Casanova[2], Barbara Di Camillo[1], Paolo Magni[2]

[1]Department of Information Engineering, University of Padova
[2]Department of Electrical, Computer and Biomedical Engineering, University of Pavia

Corresponding author: massimo.bellato@unipd.it

Abstract

Synthetic biology aims to engineer sophisticated biological functions in live cells by designing increasingly complex genetic circuits. Therefore, researches are constantly looking for innovative architectures and regulatory logics mechanisms to be adopted in rational design of such synthetic devices. In this work, a novel approach to design and implement multi-input logic gates had been introduced. Specifically, it has been studied the interaction between two different transcriptional regulation mechanisms (i.e. transcription factors and CRISPR interference) which compete for the modulation of the same promoter. As a test bed, a NOR logic gate based on this architecture had been implemented and characterized in vivo; two alternative mathematical models for the co-regulation had been compared, unveiling the actual mechanism of interaction between the two regulators. We believe that the approach and general mathematical structure used is generalizable for the description of different types of multi-modal genetic circuits. As an example, a NIMPLY logic gate exploiting the same multi-input regulatory system had been finally designed and simulated in silico.

1 Introduction

One of the aims of Synthetic biologists is to build complex genetic programs through the interconnection of pre-characterized biological parts, as it happens for other engineering fields such as electrical components in electronics or scripts and functions in computer engineering (1). Among others, transcription factor (TF)-based logic gates are one of the simplest and most common biological devices used to design genetic circuits (2); however, the use of such devices is hampered by the lack of available orthogonal and reliable TF/regulated promoter pairs.

In such framework, the introduction of the increasingly popular CRISPR technology, in the catalytically inactive version called CRISPR interference (CRISPRi), have enabled the design of a virtually infinite amount of orthogonal NOT gates (3), due to its ease of design and extreme efficiency. Indeed, while a traditional TF-based regulation system is typically implemented through a constitutive expression of a regulator protein – activated or inhibited upon binding of a specific molecule –, CRISPRi systems require the expression of the protein dCas9 along with a guide RNA (gRNA) which can be designed ad hoc. The latter molecule, once bound to dCas9 via a specific binding region, form a complex that can in turn bind a DNA tract correspondent to a second sequence encoded by the gRNA; this leads to the impossibility for the RNA polymerase to read the downstream nucleotide sequence, hence it acts as a universal programmable transcriptional repressor. While both the transcription regulator mechanisms have been deeply studied and exploited for myriads of synthetic biological systems (4), for the best of our knowledge, the possible interaction between CRISPR interference and transcription factor competing for the same binding region has not been studied yet. This possibility of usage would enable to upgrade known genetic architectures through the introduction of CRISPRi in circuits already bearing TF-based regulation; moreover, through a proper modeling approach, such novel mixed architectures could be exploited to design multi-modal gene regulation mechanisms for synthetic circuits in a predictable manner. In this work, the functioning of such mixed interactions has been investigated through new experimental data and modelling approaches, and this architecture has been used to build new synthetic circuits implementing multi-input logic functions.

2 Methods

2.1 Circuit design

As a proof of concept, a NOR gate has been conceived exploiting the TF and CRISPRi mixed interactions. Most of the parts adopted to assemble the following circuit have been previously characterized in other works of our groups, and further details can be found in (5), (6), (7). The system, as shown in Figure 1, is composed by a set of four expression devices. The first one is the dCas9 constitutive expression cassette, which is a device expressing an amount of dCas9 sufficient to fully repress the transcription from an high copy plasmid – in presence of a proper amount of gRNA – without toxic effects nor providing metabolic burden to the cell (6). The second one is an inducible expression cassette for a gRNA designed to bind the promoter in

the Target plasmid, which expression is driven by the promoter *Plac*[1] tuned through the molecule IPTG; the latter molecule is the first input of the system. These two first cassettes – borne in a medium and a low copy plasmid, respectively – implement the CRISPRi system, designed and tuned to repress a synthetic promoter developed in (7), called J119H, driving the expression of the reporter gene of a red fluorescent protein (RFP) encoded in the high copy Target plasmid. The last expression cassette encodes the *LuxR* protein, constitutively expressed, which can bind the HSL molecule and the complex can in turn bind the core of the J119H promoter, which includes a binding site for the complex, and thus repress its transcriptional activity.

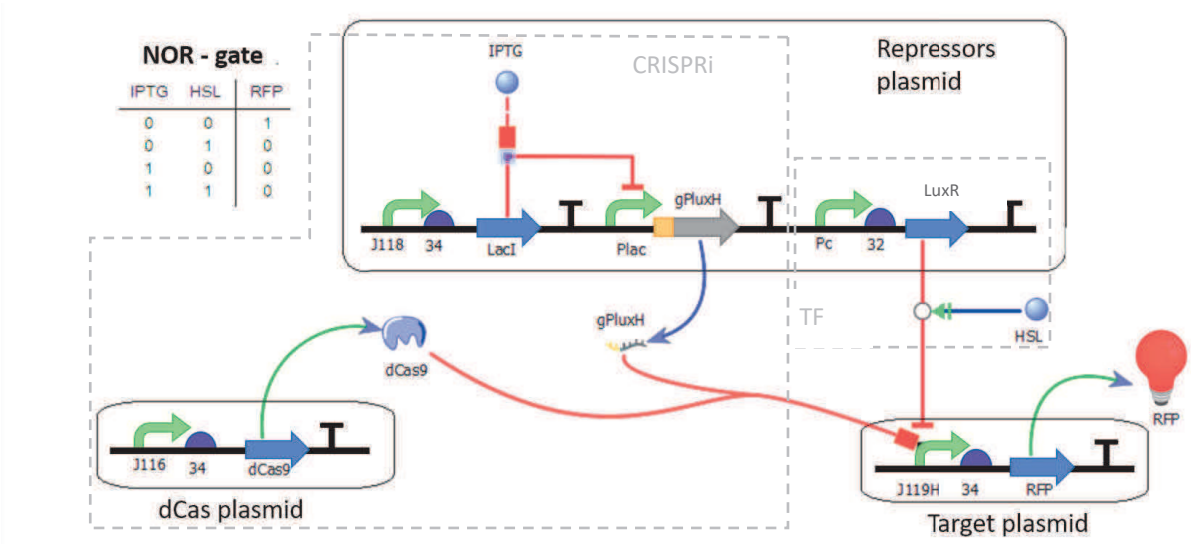


Figure 1 ***Circuit schema.*** *The circuit is composed by three plasmids; the upper one is a low copy plasmid carrying the two repressor components which are an IPTG-inducible expression cassette for the gRNA and a constitutive expression cassette for LuxR. The dCas plasmid bears a constitutive expression cassette for the dCas9 protein which binds the gRNA to repress the J119H promoter encoded in the Target plasmid. The latter plasmid can also be repressed by the complex formed by the binding between LuxR and the inducer molecule HSL, implementing a NOR gate, as shown in the table on the upper left.*

Considering the two molecules IPTG and HSL as the tunable input of the system, the circuit implements a NOR gate, providing the synthesis of the reporter gene only in absence of both the inducers. Control circuits (not shown) were also built to facilitate the parameter estimation to decouple the single contribution of each repression mechanism exerted on the Target plasmids.

2.2 Experimental setup

The system was designed and implemented in E. coli Top10 strain and genetic parts were assembled through BioBrick Standard Assembly (8) using genetic parts already available in our laboratories, re-adapted from the genetic circuits used in a previous work (6), from which the same experimental characterization setup was adopted.
For quantitative experiments, cells were first streaked on selective LB agar plates from glycerol stocks and grown overnight (~16 h); a second overnight incubation in selective M9 minimal medium supplemented with glycerol followed. IPTG was provided to this liquid culture at the indicated concentrations due to the slow dynamics of its induction. Lastly, cell cultures were 100-fold diluted in 96 well plates, in selective M9, adding the proper IPTG and HSL amount when needed.
To monitor the growth and the reporter gene synthesis, measurements of absorbance at 600nm and red fluorescence signal were performed using a Tecan Infinite F200 plate reader on the growing liquid cultures, taking measurements every 5 minutes and analyzing data with Matlab (MathWorks) and Microsoft Excel. The acquired time series were background-subtracted and the synthesis rate per cell (i.e. the average time derivative of the fluorescence signal normalized on the optical density during the exponential growth phase of the cell culture) was adopted as final outcome (9).

2.3 Modeling Framework

A key aspect of the study has been the identification of a modeling setup that properly described molecular interactions in the two mechanisms (TF and CRISPRi) repressing the same promoter. To this aim, two alternative binding mechanisms were considered: an exclusive binding in which only one repressor can bind, and an independent binding under the hypothesis that more than one repressor can bind the target promoter.

2.3.1 Exclusive binding

In exclusive binding, the basic assumption is that only one of the two mechanisms exerts its repression at once, meaning that the promoter can be repressed either from the dCas9:gRNA complex or the LuxR:HSL one but only one repressor per time. This assumption is based on the possible competition between TF and CRISPRi complex due to their footprint.
The associated kinetic model can be expressed as follows:

[1] In absence of IPTG induction, the promoter is fully repressed by the protein LacI which is constitutively expressed; the latter can be bound by IPTG, which acts as in indirect activator of *Plac*, since the higher is the molecule induction, the lower is the amount of unbound LacI able to bind and repress the promoter.

$$dCas + sgRNA \underset{k_1}{\overset{k_{-1}}{\rightleftarrows}} C_1$$

$$LuxR + HSL \underset{k_2}{\overset{k_{-2}}{\rightleftarrows}} C_2$$

$$C_1 + D_L \underset{q^-}{\overset{q^+}{\rightleftarrows}} C_3$$

$$C_2 + D_L \underset{h^-}{\overset{h^+}{\rightleftarrows}} C_4$$

$$D_L + RNAp \underset{k_{-T}}{\overset{k_T}{\rightleftarrows}} [D_L RNAp] = C_T \xrightarrow{k_R} D_L + RNAp + RFP$$

$$C_3 + RNAp \underset{k_{-T3}}{\overset{k_{T3}}{\rightleftarrows}} [D_L RNApC_1] = C_{T3} \xrightarrow{k_R} D_L + RNAp + RFP + C_3$$

$$C_4 + RNAp \underset{k_{-T4}}{\overset{k_{T4}}{\rightleftarrows}} [D_L RNApC_2] = C_{T4} \xrightarrow{k_R} D_L + RNAp + RFP + C_4$$

Where C_1 and C_2 are the dCas:gRNA and LuxR:HSL repressor complexes, respectively; C_3 and C_4 are formed when C_1 and C_2 complexes bind the target DNA, respectively.
By considering the system at the steady state for the complexes due to their fast binding dynamics compared to the *RFP* maturation, the following mass conservation laws are considered on *DNA, LuxR and dCas*:

$$D_{tot} = D_L + C_3 + C_4 + C_T + C_{T3} + C_{T4}$$
$$LuxR_{tot} = LuxR_L + C_2$$
$$dCas_{tot} = dCas_L + C_1$$

The following lumped parameters can be derived:

$$\tilde{Q} = dCas_{tot} Q \frac{1 + \beta_{RFP}}{1 + \alpha_{RFP}} [n.a.]$$
$$\tilde{H} = LuxR_{tot} H \frac{1 + \gamma_{RFP}}{1 + \alpha_{RFP}} [n.a.]$$
$$\hat{D_{tot}} = \frac{D_{tot}}{1 + \alpha_{RFP}} [\mu M]$$

With:

$$\alpha_{RFP} = \frac{k_T RNAp}{k_{-T} + k_R} [n.a.]$$
$$\beta_{RFP} = \frac{k_{T3} RNAp}{k_{-T3} + k_R} [n.a.]$$
$$\gamma_{RFP} = \frac{k_{T4} RNAp}{k_{-T4} + k_R} [n.a.]$$

And:

$$K_1 = \frac{k_1}{k_{-1}} [\mu M]$$
$$K_2 = \frac{k_2}{k_{-2}} [nM]$$
$$Q = \frac{q^+}{q^-} [\mu M]^{-1}$$
$$H = \frac{h^+}{h^-} [\mu M]^{-1}$$

Assuming the absence of basic promoter activity in the repressed state, so that k_{T3}=0 and k_{T4}=0 and therefore β_{RFP} and γ_{RFP} null, the synthesis of the measurable outcome *RFP* can finally be written as:

$$\dot{RFP} = \frac{\hat{D_{tot}} k_R \alpha_{RFP}}{1 + \frac{\tilde{Q}}{1 + \frac{K_1}{sgRNA}} + \frac{\tilde{H}}{1 + \frac{K_2}{HSL}}} = \frac{\tilde{\mathbf{D_{tot}}}}{1 + \frac{\tilde{\mathbf{Q}}}{1 + \frac{\mathbf{K_1}}{sgRNA}} + \frac{\tilde{\mathbf{H}}}{1 + \frac{\mathbf{K_2}}{HSL}}} [\mu M/s]$$

Which is a model with 5 parameters.

2.3.2 Independent binding

Assuming the independence of the binding for the two repressor complexes, a new species (C_5) is considered, deriving from the binding of C_3 and C_4 with LuxR:HSL and dCas:gRNA, respectively.
The associated kinetic model can be therefore expressed as follows:

$$dCas + sgRNA \underset{k_1}{\overset{k_{-1}}{\rightleftarrows}} C_1$$

$$LuxR + HSL \underset{k_2}{\overset{k_{-2}}{\rightleftarrows}} C_2$$

$$C_1 + D_L \underset{q^-}{\overset{q^+}{\rightleftarrows}} C_3$$

$$C_2 + D_L \underset{h^-}{\overset{h^+}{\rightleftarrows}} C_4$$

$$C_1 + C_4 \underset{q^-}{\overset{q^+}{\rightleftarrows}} C_5$$

$$C_2 + C_3 \underset{h^-}{\overset{h^+}{\rightleftarrows}} C_5$$

$$D_L + RNAp \underset{k_{-T}}{\overset{k_T}{\rightleftarrows}} [D_L RNAp] = C_T \xrightarrow{k_R} D_L + RNAp + RFP$$

$$C_3 + RNAp \underset{k_{-T3}}{\overset{k_{T3}}{\rightleftarrows}} [D_L RNApC_1] = C_{T3} \xrightarrow{k_R} D_L + RNAp + RFP + C_3$$

$$C_4 + RNAp \underset{k_{-T4}}{\overset{k_{T4}}{\rightleftarrows}} [D_L RNApC_2] = C_{T4} \xrightarrow{k_R} D_L + RNAp + RFP + C_4$$

$$C_5 + RNAp \underset{k_{-T5}}{\overset{k_{T5}}{\rightleftarrows}} [D_L RNApC_1C_2] = C_{T5} \xrightarrow{k_R} D_L + RNAp + RFP + C_5$$

Where the parameters have the same meaning as in the exclusive binding model with the addition of C_5 that represents the complex corresponding to the binding of both the dCas:gRNA and the LuxR:HSL complexes with the target promoter, at once. By considering the system at the steady state for the complexes, the mass conservation equations on *DNA, LuxR* **and** *dCas* are:

$$D_{tot} = D_L + C_3 + C_4 + C_T + C_{T3} + C_{T4} + C_{T5}$$
$$LuxR_{tot} = LuxR_L + C_2$$
$$dCas_{tot} = dCas_L + C_1$$

The following lumped parameters can be derived:

$$\tilde{Q} = dCas_{tot} Q \frac{1 + \beta_{RFP}}{1 + \alpha_{RFP}} [n.a.]$$
$$\tilde{H} = LuxR_{tot} H \frac{1 + \gamma_{RFP}}{1 + \alpha_{RFP}} [n.a.]$$

With:

$$\alpha_{RFP} = \frac{k_T RNAp}{k_{-T} + k_R} [n.a.]$$
$$\beta_{RFP} = \frac{k_{T3} RNAp}{k_{-T3} + k_R} [n.a.]$$
$$\gamma_{RFP} = \frac{k_{T4} RNAp}{k_{-T4} + k_R} [n.a.]$$
$$\xi_{RFP} = \frac{k_{T5} RNAp}{k_{-T5} + k_R} [n.a.]$$

And:

$$K_1 = \frac{k_1}{k_{-1}} [\mu M]$$
$$K_2 = \frac{k_2}{k_{-2}} [nM]$$
$$Q = \frac{q^+}{q^-} [\mu M]^{-1}$$
$$H = \frac{h^+}{h^-} [\mu M]^{-1}$$

Assuming the absence of basic promoter activity in the repressed state, such as k_{T3}=0 and k_{T4}=0 and therefore β_{RFP} and γ_{RFP} null, the synthesis of the measurable outcome *RFP* can be written as:

$$\dot{RFP} = \frac{D_{tot} k_R \alpha_{RFP}}{(1 + \alpha_{RFP}) + \frac{dCas_{tot}}{1 + \frac{K_1}{sgRNA}} Q + \frac{LuxR_{tot}}{1 + \frac{K_2}{HSL}} H + \frac{dCas_{tot}}{1 + \frac{K_1}{sgRNA}} \frac{LuxR_{tot}}{1 + \frac{K_2}{HSL}} QH}$$

$$\dot{RFP} = \frac{\tilde{D_{tot}}}{1 + \frac{\tilde{Q}}{1 + \frac{K_1}{sgRNA}} + \frac{\tilde{H}}{1 + \frac{K_2}{HSL}} + \frac{\tilde{Q}}{1 + \frac{K_1}{sgRNA}} \frac{\tilde{H}}{1 + \frac{K_2}{HSL}}} \alpha_{\hat{RFP}} \quad [\mu M/s]$$

Which is a model with 6 parameters with an additional term at the denominator which describes the effect of a simultaneous co-repression exerted by the two mechanisms.

This can be explained by the fact that during the parameter estimation procedure, the lower bound of the $\hat{\alpha}_{RFP}$ in the independent model had been set to 1, being $\hat{\alpha}_{RFP} = \alpha_{RFP} + 1$; setting $\alpha_{RFP} = 0$ the models results nested and therefore the bigger model does not add enough information to outperform the smaller one. Lastly, it is worth to notice that both the models can be reconducted to Hill-like structures (procedure not shown) and, by including in the analysis a proper set of control experiments to decouple the contributes of each single repressor, they are a-priori identifiable.

3 RESULTS

As shown in

Figure 2, the circuit behaves as expected, implementing a NOR logic with respect to the two input molecules IPTG (reported on the x-axes) and HSL (color coded). Indeed, for both the molecules, an increment in the induction is associated with a decrease in the RFP, which is the output signal. It is worth to notice that the system can be completely repressed even in presence of only the LuxR regulation system fully induced with HSL (i.e. control circuit with a non-targeting gRNA), despite data are not reported for the sake of clarity of the figure.

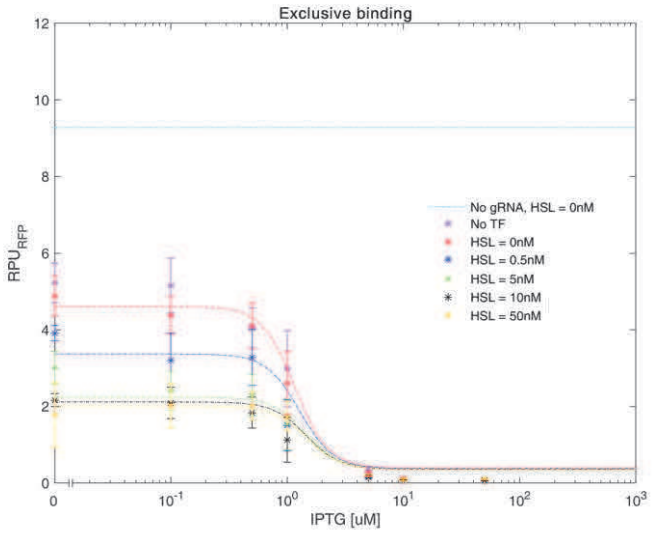


Figure 2 ***Fitting with Exclusive binding model.*** *Experimental data are represented as average on at least 3 biological replicates with stars while model predictions are reported with dashed lines. The purple line, correspondent to the control circuit without TF is overlapped by the red one, for which the TF is completely uninduced. The cyan line represents the average on several samples of the control circuit bearing a non-targeting gRNA, for several IPTG values. Error bars represents the 95% confidence interval of the mean.*

3.1 SYSTEM CHARACTERIZATION AND PARAMETER ESTIMATION

The study focused on the evaluation of the hypotheses of exclusive or independent binding. An additional hypothesis was the adoption of a Hill-like function to describe the gRNA production, being adopted in other previous works. Indeed, gRNA expression is driven by the well characterized IPTG-inducible promoter, which transfer function has been reported to fit with a Hill equation (5). The parameters of this latter equation had been estimated from previously acquired data (not reported). Then, model parameters were estimated via Matlab *lsqnonlin* function for both the model, as reported in Table 1.

Table 1 ***Parameter estimation***

Model	$\bar{D}_{tot}$	$\bar{Q}$	K_1	$\bar{H}$	K_2	$\hat{\alpha}_{RFP}$	MSE
Exclusive	9.277	34.801	1.546	2.680	1.303	-	0.443
Independent	9.002	36.280	1.823	1.995	3.801	1.418	0.678

As a metrics to evaluate which model better described the system, the mean square error (MSE) was adopted. It is worth to notice that, despite this index does not enable a fair comparison between models with different number of parameters, in this case the less complex model (Exclusive binding) provides an better MSE. The results of the model fitting using the chosen Exclusive binding model are reported in Figure 2.

3.2 PREDICTION OF ALTERNATIVE LOGICS

Lastly, the implemented model has been used to study the theoretical implementation of another logic gate that is a NIMPLY gate. The hypothetic circuit would differ from the one in Figure 1 in only two components that are the promoter driving the RFP and the associated gRNA. Indeed, by using the *Plux* promoter [REF] which is activated by the LuxR:HSL complex, the logic of the TF gate is inverted. It is possible to demonstrate that, by using the Exclusive binding model, the RFP synthesis can be in this case written as:

$$R\dot{F}P = \frac{\frac{\check{\mathbf{H}}}{1+\frac{\mathbf{K_2}}{HSL}}}{1+\frac{\tilde{\mathbf{Q}}}{1+\frac{\mathbf{K_1}}{sgRNA}}+\frac{\tilde{\mathbf{H}}}{1+\frac{\mathbf{K_2}}{HSL}}} \ [\mu M/s]$$

With $\check{\mathrm{H}} = k_R Lux R_{tot} H \gamma_{RFP} \frac{D_{tot}}{1+\alpha_{RFP}}$.

Simulation of the expected behavior are reported in Figure 3, with all the parameters set to previously estimated ones, except for $\tilde{Q}$ that represents the affinity between the gRNA and the promoter, which is a-priori unknown since it depends on the gRNA sequence.

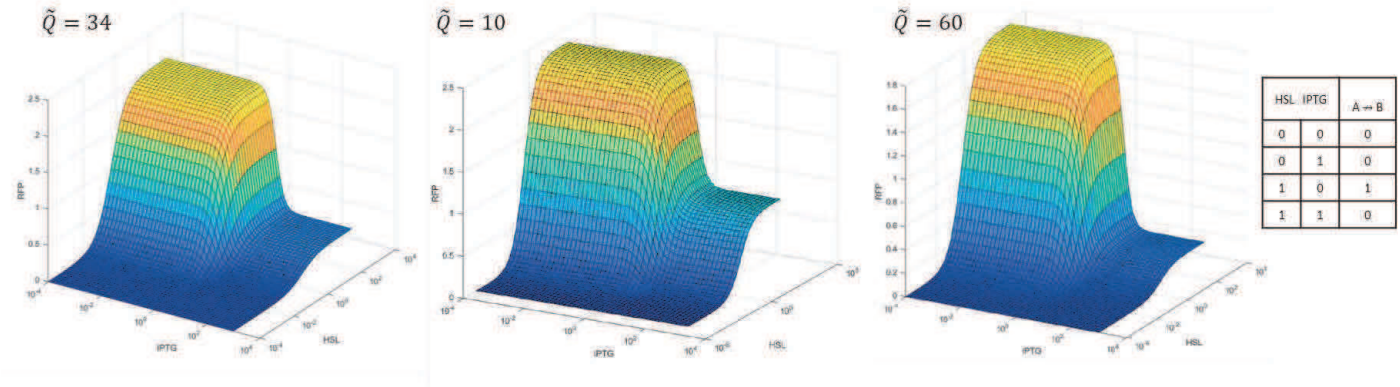


Figure 3 ***Prediction of a NIMPLY logic gate.*** *The system differs from the previous one in the activation of the target promoter via LuxR:HSL complex. Simulations of the system output are provided for different values of* $\tilde{Q}$.

4 DISCUSSION

Synthetic biologists are in a constant research of new tool to develop genetic circuits implementing ever more complex functions in living cells, with eventually a predictable behavior. This research focused on unveiling the interaction between CRISPRi and transcriptional factors in simultaneously regulating the same promoter, to enable rational design of multimodal NOR gate in synthetic genetic circuit. The case study of LuxR-HSL system has been exploited as a model. First, it has been demonstrated that such architecture is effective, and the intended Boolean logic can be fully achieved by using the desired combined mechanism. The data collected from the characterization of the genetic constructs implemented ad-hoc to realize the NOR gate were hence used to develop a the mathematical model that best represented the underlying biological interactions; in particular, two models had been implemented starting from the several biological assumption adopted in the circuit design, which were an exclusive and an independent binding models for the two transcriptional regulators to the shared target promoter (i.e. dCas9:gRNA complex and LuxR:HSL transcription factor). From the model's parameter estimation it was possible to demonstrate that the interaction between the two regulators is more likely to be an exclusive one, in which the target DNA can be bound only to one of the two complexes at the same time. This result is in accordance with a biological hypothesis of physical mutual obstruction of the repressors once bound, due to the proximity of the two binding sites and the considerable size of the complexes.

The derived model has been finally used to simulate a NIMPLY logic gate, implementable through a LuxR:HSL inducible promoter instead of the previous J119H; the results shows that the response of the system could qualitatively match the required Boolean logic, despite a proper tuning of the expression levels might be required to achieve a more ideal on/off response to the inducers to reach a fully dichotomy in the system response.

With respect of the state of art, a novel architecture for genetic logic gates had been proposed, which exploit the interaction between two transcriptional repressors; a multi-input logic gate has been implemented in vivo and a mathematical model has been developed. The approach and general mathematical structure used is generalizable for the description of different types of multi-modal genetic circuits.

References

1. **Kitney R., Calvert J., Challis R., Cooper J., Elck A., Freemont P.S., Haselo J., Kelly M., Paterson L.** *Synthetic biology: scope and applications and implications.* 2009, The Royal Academy of Engineering.

2. **Brophy J.A.N, Voigt C.A.** *Principles of Genetic Circuit Design.* 5, 2013, Nature Methods, Vol. 11, p. 508–520.

3. **Qi L.S., Larson M.H., Gilbert L.A., Doudna J.A., Weissman J.S., Arkin A.P., Lim W.A.** *Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression.* 5, 2013, Cell, Vol. 152, p. 1173-1183.

4. **Engstrom M.D., Pfleger B.F.** *Transcription control engineering and applications in synthetic biology.* 3, 2017, Synthetic and Systems Biotechnology, Vol. 2, p. 176-191.

5. **Pasotti L., Bellato M., Casanova M., Zucca S.,Cusella De Angelis M.G., Magni P.** *Re-using biological devices: a model-aided analysis of interconnected transcriptional cascades designed from the bottom-up.* 2017, Journal of Biological Engineering, p. 11-50.

6. **Bellato M., Fristeri Chiacchiera A., Salibi E., Casanova M., De Marchi D., Cusella De Angelis M.G., Pasotti L., Magni P.** *CRISPR interference as low burden logic inverters in synthetic circuits: characterization and tuning.* 2020, Biorxiv.

7. **Zucca S., Pasotti L., Politi N., Casanova M., Mazzini G., Cusella De Angelis M.G., Magni P.** *Multi-Faceted Characterization of a Novel LuxR-Repressible Promoter Library for Escherichia coli.* 2015, PLoS One.

8. **Knight T.F.** *Idempotent vector design for standard assembly of biobricks.* 2003.

9. **Endy D.** *Foundations for engineering biology.* 7067, 2006, Nature, Vol. 438.

Abstracts of Posters

Investigating structural and functional properties of menin protein

Carmen Biancaniello [1], Antonia D'Argenio [2,3], Serena Dotolo [1],
Deborah Giordano [3], Bernardina Scafuri [2], Antonio d'Acierno [3], Anna Marabotti [2],
Roberto Tagliaferri [1], Angelo Facchiano [3]

[1] NeuRone Lab, Dep. of Management & Innovation Systems (DISA-MIS), University of Salerno, Italy

[2] Dept. Chemistry and Biology "A. Zambelli", University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy

[3] National Research Council, Institute of Food Science (CNR-ISA), via Roma 64, Avellino, Italy

Corresponding author: angelo.facchiano@isa.cnr.it

Abstract

Menin protein sequence consists of 615 amino acids, coded by the MEN1 gene, located on the chromosome in position 11q13 and made up of 10 exons. Mainly present at the nuclear level, menin is an extremely versatile scaffold protein from a functional point of view, such as to be involved in transcription regulation, genome stability, DNA repair, signaling and cell division.

Mutations in the MEN1 gene are responsible for the onset of a rare autosomal dominant disease, multiple endocrine neoplasia type 1 (MEN1), characterized by endocrine alterations that must be present in a combined manner for at least two of the following conditions: parathyroid glands, anterior pituitary gland and neuroendocrine tumors of the gastro-entero-pancreatic tract (GEP-NET). These conditions may occur in a non-hereditary form with no family history of MEN1 (sporadic MEN1) or in several members of a family (familial MEN1). Adrenal cortical tumors, carcinoid tumors and skin lesions such as facial angiofibromas, collagenomas and lipomas can also be associated.

We investigated the structural properties of menin and the potential effects of known mutations. The results are deposited in dedicated data base and can be explored by a web interface. A comprehensive classification of the effects of single mutations is presented in the poster.

Identification of novel potential gene involved in programmed cell death by integrated and comparative analyses

Francesco Monticolo [1], Emanuela Palumbo [2], Maria Luisa Chiusano [1]

[1] Department of Agricultural Sciences, Università degli studi di Napoli Federico II, Portici, Italy

[2] Department of RIMAR, Stazione Zoologica "Anton Dohrn", Naples, Italy

Email of Corresponding author: chiusano@unina.it

Abstract

The explosion of omics technologies offers challenging opportunities to identify molecular agents and processes that may play relevant role in programmed cell death. They can support comparative investigations, in one/multiple experiments, exploiting evidence from one/multiple species. We here propose a pipeline to considered gene expression data from induction of programmed cell death and stress response in Homo sapiens and compared the results with Saccharomyces cerevisiae gene expression during the response to cell death. The aim was to identify conserved candidate genes associated to human or yeast cell death, favored by crosslinks based on orthology relationships between the two species. We identified differentially expressed genes, pathways that are significantly dysregulated across the treatments and characterized genes among those involved in induced cell death. We investigated on co-expression patterns and identified novel genes that were not expected to be associated to death pathways that have a conserved pattern of expression between the two species. The pipeline that we designed can be further exploited expanding the number of experiments and/or the reference species to consider and, as a consequence, it can result in additional conserved genes involved in cell death. These efforts can contribute to the knowledge on cell death molecular pathways in distantly related species, and paves the way to novel discovery in the field, also contributing with new key targets for cancer therapy.

Microalgal RNA-seq analyses to identify enzymes involved in the synthesis of bioactive compounds

Giorgio Maria Vingiani [1], Pasquale De Luca [2], Daniele De Luca [3] and **Chiara Lauritano** [1]

[1]Marine Biotechnology Department, Stazione Zoologica Anton Dohrn, Villa Comunale, CAP80121 Napoli, Italy;

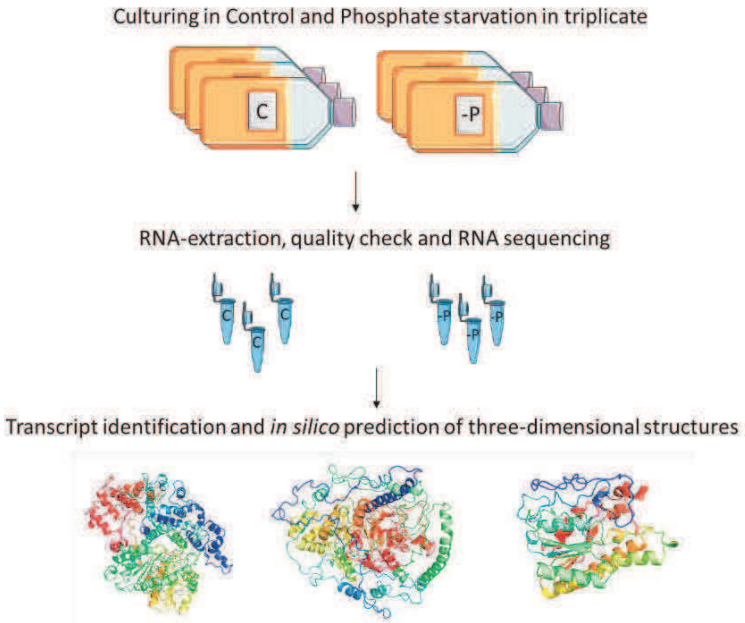
[2]Research Infrastructure for Marine Biological Resources Department, Stazione Zoologica Anton Dohrn, Villa Comunale, CAP80121 Napoli, Italy;

[3]Department of Biology, Università degli Studi di Napoli Federico II, CAP80139 Naples, Italy;

Email of Corresponding author: chiara.lauritano@szn.it

Abstract

Microalgae have shown to be excellent producers of bioactive compounds, such as lipids, vitamins, as well as defence metabolites which have also shown possible applications for managing human pathologies. Many microalgae produce toxic compounds with negative impacts on human and environmental health. The dinoflagellate *Alexandrium tamutum* was discovered for the first time in the Gulf of Naples, and it is not known to produce saxitoxins. However, a clone of *A. tamutum* from the same Gulf showed toxicity on predators and anti-proliferative activity on human cells. *A. tamutum* RNA-seq approach was used for *in silico* identification of transcripts that can be involved in the synthesis of toxic compounds (Both in control and phosphate starvation condition, to induce toxin production). Results showed the presence of three transcripts related to saxitoxin synthesis (sxtA, sxtG and sxtU), and others potentially related to the synthesis of additional toxic compounds (e.g., 44 transcripts annotated as "polyketide synthase"). These data suggest that even if this *A. tamutum* clone does not produce saxitoxins, it has the potential to produce toxic metabolites, in line with the previously observed activity. These data give new insights into toxic microalgae, toxin production and their potential applications for the treatment of human pathologies.



A comprehensive evaluation of differential alternative splicing tools for RNA-seq data

Jamal Elhasnaoui, Giulio Ferrero, Michele De Bortoli

Department of Clinical and Biological Sciences, University of Turin, Orbassano, 10043 Turin, Italy

Email of Corresponding author: jamal.elhasnaoui@unito.it

Abstract

Alternative splicing (AS) is an important molecular mechanism regulating gene expression and is involved in a plethora of cellular process like proliferation, differentiation and development. This fine-tuned molecular mechanism enables a tightly regulated generation of multiple mRNA and protein products from the same gene, thus allowing an increase in the complexity and diversity of the proteome content of the cell. To date, a number of computational approaches have been developed to identify and quantify differentially spliced genes from RNA-seq data, but a comprehensive comparison or benchmarking of these approaches is lacking. In this study, 6 different tools were used to identifying differentially spliced genes and were evaluated for consistency and reproducibility, precision, recall, false discovery rate and functional enrichment analysis. The selected tools represent three different methodological categories: isoform-based (IsoformSwitchAnalyzeR, DEXSeq), event-based methods (rMATS, SUPPA2) and junction-based (PSIsigma, Whippet). Overall, all the junction-based methods (PSIsigma, Whippet) and the event-based method (rMATS) scored well on the selected measures. Using a golden standard dataset, of the 6 tools tested, the junction-based methods performed generally better than the isoform-based and event-based methods. However, overall, the assessment of different data analysis tool performance was dependent on the sequencing depth, number of samples and the types of the analysed dataset. We foresee that our study will help in selecting the best approach to analyse differential AS from RNA-seq data.

The AgroLD project: A Knowledge Graph Database for rice functional genomics

Pierre Larmande[1]
[1]UMR DIADE, IRD, Univ. Montpellier, France.

Keyword:Graph Database, Ontologies, Semantic web, multi-omics Data Integration

Recent advances in high-throughput technologies have resulted in tremendous increase in the amount of data in the agronomic domain. There is an urgent need to effectively integrate and assimilate complementary information to understand the biological system in its entirety. We have developed AgroLD, a knowledge graph system that exploits the Semantic Web technology and some of the relevant standard domain ontologies, to integrate information on rice species and in this way facilitating the formulation of new scientific hypotheses. We present some integration results of the project, which initially focused on genomics, proteomics and phenomics. AgroLD is now an RDF knowledge base of 100M triples created by annotating and integrating more than 50 datasets coming from 10 data sources –such as Gramene.org and TropGeneDB– with 10 ontologies –such as the Gene Ontology and Plant Trait Ontology. Our objective is to offer a domain specific knowledge platform to solve complex biological and agronomical questions related to the implication of genes in, for instances, plant disease resistance or high yield traits. We expect the resolution of these questions to facilitate the formulation of new scientific hypotheses to be validated with a knowledge-oriented approach.

Integrated bioinformatics to investigate novel biological processes in model species

Emanuela Palomba[1], Francesco Monticolo[2], Stefano Mazzoleni[2], Maria Luisa Chiusano[1,2]

[1]Stazione Zoologica "Anton Dohrn", Napoli, Italy

[2]Department of Agricultural Sciences, University of Naples Federico II, Via Università 100, Portici 80055 (NA), Italy

Email of Corresponding author: emanuela.palomba@szn.it

ABSTRACT

We here propose an integrated bioinformatics approach to investigate on the molecular mechanisms of a biological process that had no previous characterization: the response to self (homologous, i.e. DNA from the same or closely related species) and nonself (heterologous, i.e. DNA from phylogenetically unrelated species) extracellular DNA.In particular, in this study we analysed the early response after exposure to extracellular self- and nonself-DNA in the plant model *Arabidopsis thaliana* by performing a whole-plant trascriptome profiling by RNA-seq. Our aim was to shed light on the cellular molecular mechanisms activated following the treatment with exDNA.

The resulted highlighted a different response to self and nonself DNA.

INTRODUCTION

In 2015, Mazzoleni and coworkers (1) demonstrated that fragmented extracellular DNA (exDNA) triggers a a concentration dependent and species-specific inhibitory effect on root growth and seed germination in plants, and it was proposed this could contribute to the phenomenon of plant–soil negative feedback (2). This discovery was also extended to organisms of other taxonomic groups including microbes, fungi, protozoa, and insects (3). Despite some hypoteses (4,5), still little is known about the cellular sensing and molecular mechanisms underlying plant growth inhibitory effect of extracellular self-DNA, as well as plant response to extracellular nonself-DNA.

METHODS

The experimental design included the following treatments after three different stages within 16 hours: exposure to sterile distilled water, to self-DNA, to nonself-DNA (*Clupea harengus)*. The RNA extracted was sequenced by the Illumina Hiseq2500. The cleaned reads were mapped to the *Arabidopsis* nuclear and cytoplasmic genomes (version TAIR 10) using the STAR software (version 2.4.2a) (6). The mapped reads were counted by featureCounts (version 1.4.6-p5) (7). Differentially expressed genes (DEGs) call, comparing DNA treatments at each stage with the respective control, were made performing three different statistical approaches (FDR < 0.05): i) DESeq2 (Love et al. 2014); ii) edgeR and iii) edgeR GLM (8) . The union of the three approaches was considered for subsequent analyses.

A K-means cluster analysis on DEGs ($|\log2(FC)| \geq 1$) was performed with MeV (9), using the Pearson Correlation as distance metric. DEGs and samples were also submitted to Principal Component Analysis (PCA), plotting vector loadings for treatment combinations (timing and type of exposure to DNA) and factorial scores of cluster centroids in the multivariate space defined by the first three ordination axis. Gene ontology (GO) enrichment analyses on DEGs were performed using the GOseq package (10) (FDR $\leq$ 0.05), and the reference GO annotation for Arabidopsis (http://plants.ensembl.org/index.html).

Finally, lists of genes annotated with the most enriched GOs, showing different expression pattern between self-DNA and nonself-DNA treatments at each observation stage (i.e. 1, 8 and 16 hours.) were collected in order to quantitatively assess between-treatment differences and discuss them in detail at single-gene level.

RESULTS AND DISCUSSION

The sensing of exDNA has been considered in the framework of DAMP (damage-associated molecular pattern) sensing (11)

Therefore, we considered all possible current annotation to filter out these evidences. However, no clear pattern arose, especially for self-DNA response.

The integrate bioinformatics approaches allowed the identification of a clear differential pattern of response to extracellular self- and nonself-DNA helping the understanding of the molecular mechanisms of a novel and

uncharacterized biological process. In particular, from differentially expressed gene and GO enrichment analysis, a primary response to self-DNA sensing is represented by an altered chloroplast functioning and ROS production, eventually leading to damages and cell cycle arrest. Differently, nonself-DNA analysis highlighted the upregulation of genes related to the hypersensitive response, possibly evolving into systemic acquired resistance.

REFERENCES

1. Mazzoleni S, Bonanomi G, Incerti G, Chiusano ML, Termolino P, Mingo A, et al. Inhibitory and toxic effects of extracellular self-DNA in litter: a mechanism for negative plant-soil feedbacks? New Phytol. 2015 Feb;205(3):1195–210.
2. van der Putten WH, Bradford MA, Pernilla Brinkman E, van de Voorde TFJ, Veen GF. Where, when and how plant–soil feedback matters in a changing world. Funct Ecol [Internet]. 2016;30(7):1109–21. Available from: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2435.12657
3. Mazzoleni S, Cartenì F, Bonanomi G, Senatore M, Termolino P, Giannino F, et al. Inhibitory effects of extracellular self-DNA: a general biological process? New Phytol. 2015 Apr;206(1):127–32.
4. Duran-Flores D, Heil M. Extracellular self-DNA as a damage-associated molecular pattern (DAMP) that triggers self-specific immunity induction in plants. Brain Behav Immun. 2018 Aug;72:78–88.
5. Veresoglou SD, Aguilar-Trigueros CA, Mansour I, Rillig MC. Self-DNA: a blessing in disguise? New Phytol [Internet]. 2015;207(3):488–90. Available from: https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/nph.13425
6. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan;29(1):15–21.
7. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014 Apr;30(7):923–30.
8. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan;26(1):139–40.
9. Howe EA, Sinha R, Schlauch D, Quackenbush J. RNA-Seq analysis in MeV. Bioinformatics. 2011 Nov;27(22):3209–10.
10. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010;11(2):R14.
11. Albert I, Hua C, Nürnberger T, Pruitt RN, Zhang L. Surface Sensor Systems in Plant Immunity. Plant Physiol [Internet]. 2020;182(4):1582–96. Available from: http://www.plantphysiol.org/content/182/4/1582

jewel: a novel method for data integration

Claudia Angelini, Daniela De Canditiis, Anna Plaksienko

Istituto per le Applicazioni del Calcolo, 'Mauro Picone', CNR-Napoli, Italy
Istituto per le Applicazioni del Calcolo, 'Mauro Picone', CNR-Roma, Italy
Gran Sasso Science Institute, L'Aquila, Italy

anna.plaksienko@gssi.it

Abstract

Modern high-throughput technologies allow the collection of large omics data-sets with decreasing costs and times, and researchers can use such data to understand both complex cellular mechanisms and the molecular basis of disease onset and progression. For such purpose, scientists developed several computational methods for the analysis of omics data-sets. Among the most exciting approaches, network inference methods allow inferring the relationships among the system's unit. The units/variables (for example, the genes) represent a network's vertices, with the edges representing some form of relation. In this context, graphical models can describe the conditional dependence among two variables, given the remaining ones, as a graph.

Nowadays, it is common to collect and analyze more than a single data-set for the same question of interest. The collected data-sets can be of different omics types, arise from various studies, collected in different laboratories, or with different technologies. The joint analysis of such data-sets can lead to a more accurate characterization of the exam's biological system. However, to fully exploit the advantages of having multiple data-set, it is necessary to develop novel data integration methods.

We propose jewel, a novel method for the joint analysis of multiple data-sets under the assumption that each of them follows the Gaussian distribution and is, in fact, a graphical model. In this context, the conditional independence relationships between variables (genes) are encoded in the inverse covariance matrix. We assume that the conditional independence structure is shared among the different data-sets, but the covariance matrices can differ. In this setting, combining the individual data-sets into a single one and estimating a unique graphical model would mask the covariance matrices' underlying heterogeneity while estimating separate models for each case would not take advantage of the common underlying structure.

This work describes the novel R package jewel. jewel is a novel method based on a group penalized regression approach to estimate a common dependency graph from multiple matrices and guarantees the sparsity and symmetry of the estimated graph. We illustrate the performance through simulated and real data examples describing transcriptional regulatory networks based on gene expression data. We compare the proposed approach with other available alternatives.

Identifying biological functions underlying phenotypes using PhenPath

Giulia Babbi1, Pier Luigi Martelli1*, and Rita Casadio1

[1] Biocompiuting Group, University of Bologna, Italy.

Email of Corresponding author: pierluigi.martelli@unibo.it

Abstract

Co-occurrence of different phenotypes hampers the understanding of the molecular mechanisms characterizing diseases. While many resources focus on the relationship among phenotypes, diseases, and genes, little is known about the relevance of molecular functions and functional processes underlying the occurrence of phenotypes.

To this aim, here we describe a new resource called PhenPath (phenpath.biocomp.unibo.it), recently published. PhenPath allows phenotype functional annotation, after an enrichment procedure of the functional annotation of the different phenotype/disease-associated genes. Functional annotations consider Gene Ontology (Molecular Function, Biological Process and Cellular Component), KEGG and Reactome pathways.

PhenPath can be adopted to endow a disease (described with a set of phenotypes) with novel links to genes and functional terms, retrieved by intersecting the sets of genes and functional terms associated with the single phenotypes in PhenPath, as we proved in many study cases (e.g.: Rett syndrome, Tourette syndrome). We participated in the Critical Assessment of protein Function Annotation algorithms (CAFA) 4th edition using a method based on PhenPath, and the preliminary results of the experiment show that PhenPath is among the top-scoring methods.

We propose our resource for directing scientific efforts, helping the diagnosis and retrieving new possible associations among biological processes and diseases.

Designing and selection of lineage-specific baits for the genomic study of non-model organisms

T. R. Galise[1], D. Cafasso[1], P.M. Schlüter[2], P.M. Pinheiro[3], M. Ayasse[3], S. Cozzolino[1]

[1] Department of Biology, University of Naples Federico II

[2] Institute of Biology, University of Hohenheim

[3] Institute of Evolutionary Ecology and Conservation Genomics, University of Ulm

Email of Corresponding author: teresargalise@gmail.com

Abstract

Hybrid capture-based target enrichment is often used for enrichment specific genes, exons and/or other genomic regions of interest in non-model organisms without a reference genome. Here we describe a workflow for designing baits suitable for orchid genomic hybridization. Starting by available *Ophrys sphegodes* and *Phalaenopsis equestris* genomes and transcriptomes we firstly selected 1,000bp and 500bp sliding windows respectively and 1,680,457 baits were designed (769,809 *Ophrys* baits; 910,648 *Phalaenopsis* baits) with a length 100bp, CG content between 20-50%. For excluding baits too variable (<90 %) or too conserved (>95%) *Ophrys* baits within 90-95% identity interval with *Phalaenopsis* genome and transcriptome, and viceversa, were filtered by BLASTN retaining only those mapping once. The resulting baits (13894 from *Ophrys*, 11907 from *Phalaenopsis*) were further tested for mapping on alternative genome and transcriptome through simulations with Bowtie2/samtools and 23.200 baits were finally selected for synthesis.

As a testing ground, CAPSIM was used to simulate the dynamics of the capture process between the selected baits and some available orchid genomes/transcriptomes (*Dendrobium catenatum, Apostasia shenzhenica, Orchis italica* and *Cymbidium faberi*), by defining an Illumina sequencing of 150bp single-end reads. The simulation generates approx. 17,000 raw reads resulting from cross-hybridization with selected baits.

Screening procedure for selection of putative ligands of SARS-COV-2 proteins

Deborah Giordano [1], Maria Antonia Argenio [1], Bernardina Scafuri [2], Virginia Carbone [1], Anna Marabotti [2], Angelo Facchiano [1,*]

[1] National Research Council, Institute of Food Science (CNR-ISA), via Roma 64, Avellino, Italy

[2] Dept. Chemistry and Biology "A. Zambelli", University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy

Corresponding author: angelo.facchiano@isa.cnr.it

Abstract

This study has been aimed to the screening of putative ligands of SARS-CoV2 proteins. The investigation has been performed by means of a semi-automatic procedure that includes molecular docking simulations with a large number of ligands against the PDB structures available of proteins from the SARS-Cov2. The semi-automatic procedure is configured to work on a linux server and, depending on the server computational power, it may produce thousands of protein-ligand simulations in about 24 hours. Results of the screening evidenced a number of ligands with potential ability to bind the viral proteins, with a perspective interest in the finding of molecules useful against the pandemic urgency.

Modeling DNA methylation profiles through a dynamic equilibrium between methylation and demethylation

Giulia De Riso [1]*+, Damiano Francesco Giuseppe Fiorillo [2,3] +, Annalisa Fierro [4], Mariella Cuomo [1,5], Lorenzo Chiariotti [1,5], Gennaro Miele [2,3] and Sergio Cocozza [1]

+ These authors contributed equally to the work

1 Dipartimento di Medicina Molecolare e Biotecnologie Mediche, Università degli Studi di Napoli "Federico II", Via S. Pansini 5, 80131 Naples, Italy

2 Dipartimento di Fisica "E. Pancini", Università degli Studi di Napoli "Federico II", Naples, Italy

3 Istituto Nazionale di Fisica Nucleare, Sezione di Napoli, Naples, Italy

4 CNR-SPIN, c/o Complesso di Monte S. Angelo, via Cinthia, 80126 Naples, Italy

5 CEINGE Biotecnologie Avanzate, via Gaetano Salvatore 482, 80145 Naples, Italy

Email of Corresponding author: giulia.deriso@unina.it

Abstract

DNA methylation is a heritable epigenetic mark that plays a key role in regulating gene expression.

In this study, we performed a high-depth analysis of DNA methylation at the Transcription Start Site surrounding region of the mouse D-Aspartate Oxidase gene and of the human D-Serine Oxidase gene. DNA methylation was analyzed in different tissues and/or different developmental stages, for a total of 17 different conditions. For each condition, three individuals were analyzed, for a total of 51 analyzed samples. For each sample, we grouped the epialleles in methylation classes (MCs), based on the number of methylated cytosines they bore.

When analysing the MC distributions, we noted that in all the conditions almost all the possible MCs were represented, although with different frequencies. Furthermore we found that the DNA methylation status at these loci differed among cell. However, these cell-to-cell differences were maintained between different individuals, which indeed showed very similar DNA methylation profiles.

We therefore hypothesized that the observed cellular heterogeneity of DNA methylation profiles reflected a dynamic balance between DNA methylation and demethylation. Considering the low inter-individual variability, we also hypothesized a steady-state equilibrium in the cell population. We hence developed a simple mathematical model to test this hypothesis. It is worth noting that in the present analysis we modeled both DNA methylation and demethylation as cooperative processes, which probability rates were linearly dependent on the number of methylated cytosines in the epiallele.

When comparing the predicted MC distributions with the experimentally determined ones, we observed a very satisfactory agreement, thus suggesting that the data were compatible with the co-occurrence of DNA methylation and demethylation at the same genomic locus. Furthermore, our model suggested that the methylation status of neighboring cytosines contributes to this balance.

Structural model for recruitment of RIT1 to the LZTR1 E3 ligase: evidences from an integrated computational approach

Antonella Paladino[1], Michele Ceccarelli [1,2]

[1] BIOGEM Istituto di Ricerche Genetiche G. Salvatore, via Camporeale 83031, Ariano Irpino AV

[2] Dipartimento di Ingegneria Elettrica e delle Tecnologie dell' Informazione DIETI ,Università degli Studi di Napoli "Federico II"

Email of Corresponding author: antonella.paladino@biogem.it

Abstract

LZTR1 (leucine-zipper like transcriptional regulator 1) was first identified as a tumor suppressor gene mutated in glioblastoma multiforme.[1] Inactivating mutations have been reported as somatic events in cancer, while rare LZTR1 variants have been linked to schwannomatosis and Noonan syndrome, a RASopathy with a wide spectrum of developmental disorders and predisposition to certain cancers.[2-4] Important efforts in the characterization of ubiquitin pathway across many cancer types also found that LZTR1 is among the frequently mutated genes.[5]

LZTR1 encodes a multidomain protein of the BTB-Kelch superfamily; it is involved in apoptosis and ubiquitination, as a substrate adaptor in cullin 3 RING E3 ubiquitin ligase (CUL3) complexes.
Recent mass spectrometry studies have detected the physical interaction between LZTR1 and RIT1, a RAS-related small GTPase, confirming that either pathogenic mutations in LZTR1 or RIT1 fails to promote RIT1 CUL3-mediated proteosomal degradation. [6] Yet, very little is known about its active state and how it triggers the recruitment of substrates to ubiquitinate for degradation.
Here we address the structural characterization of LZTR1-RIT1 binding using an integrated computational approach[7]: 1) homology modeling to obtain the full-length LZTR1 3D structure; 2) molecular docking experiments for the prediction of the protein-protein complex; 3) mutational scanning to identify hotspots of the interaction and 4) all-atom Molecular Dynamics studies. Our findings yield important insights into the stability and conformational behavior of LZTR1-RIT1 complexes. We clarify the key role of specific pathogenic mutations on the recognition patterns and in the elicitation of E3 ubiquitin activity, thus contributing to elucidate E3-substrates relationships.

1. Frattini et al., *Nat Genet 2013*; 45(10):1141-9*;* ***2.*** Piotrowski et al., *Nat Genet 2014*; 46(2):182-7*;* ***3.*** Steklov et al., *Science 2018*; 362(6419):1177-1182*;* ***4.*** Yamamoto et al., *J Med Genet 2015*; 52(6):413-21*;* ***5.*** Ge et al., *Cell Rep 2018*; 23(1):213-226.e3*;* ***6.*** Castel et al., *Science 2019*; 363(6432):1226-1230*;* ***7.*** Rinaldi et al., *JCIM 2020*; 60(3):1469-1480.