Preface

The Brazilian Symposium on Bioinformatics (BSB, *Simpósio Brasileiro de Bioinformática*) is an international scientific conference with focus on Bioinformatics, Computational Biology, Systems Biology, Biomedical Informatics and related areas. It is organized every year by the Brazilian Computer Society (*Sociedade Brasileira de Computação* – SBC), under the steering of the Special Committee for Computational Biology (*Comissão Especial de Biologia Computacional* – CE-BioComp), which is presently coordinated by Sérgio Lifschitz (PUC-RJ) and co-coordinated by Kele Belloze (CEFET-RJ). BSB 2024 was the 17th edition of the conference, taking place in December 2–4, 2024, at Cidade da Inovação, Instituto Federal do Espírito Santo (IFES), Vitória, Brazil. Vitória is the capital of Espírito Santo, a state located in the Brazilian southeast.

BSB 2024 had as general chair Sérgio Nery Simões (IFES) and as vice-general chair Karin Satie Komati (IFES). The Local Organization Committee also had as members Francisco de Assis Boldt (IFES), Hilário Tomaz Alves de Oliveira (IFES), and Leandro Colombi Resendo (IFES). The Technical Program Committee (TPC) of this year was composed of 44 members from Brazil and also from Canada, France, Germany, Mexico, Portugal, and the USA. The symposium two tracks of this year, accepting contributions in the form of full and short papers, received a total of 33 full paper submissions, with 20 works being accepted, and a total of 7 short paper submissions, with 2 of them being accepted. The submitted articles were evaluated through a single-blind review process, with each paper having at least two independent reviews. The 22 accepted papers were presented at the conference by one of their authors at one of the four technical sessions held during BSB 2024, and now they have their archival versions in these annals.

Apart from the technical sessions, the conference had a mini-course ("Introduction to Machine Learning for Bioinformatics") and two poster sessions, alpha and beta, in which 59 posters were presented. BSB 2024 also counted four internationally renowned keynote speakers: Peter Stadler (Leipzig University, Germany), Giancarlo Guizzardi (University of Twente, the Netherlands), Yayoi Natsume-Kitatani (NIBIOHN, Japan), and Maria Emília Machado Telles Walter (University of Brasília, Brazil). Maria Emília Walter was the honored researcher of this year, "in recognition of her dedication and exceptional contribution to the advancement of Bioinformatics and the training of students throughout a brilliant career". Thank you so much, Maria Emília!

We also thank everyone who made BSB 2024 such a successful event, with around 100 (one hundred) delegates from Brazil and abroad: the TPC members, the local organization members, and volunteers, the keynote speakers, the session chairs, the several researchers that helped in the evaluation of the works presented during the poster sessions, SBC and CNPq for their support. Nominally, we also thank the best paper evaluation board (João Carlos Setubal, Alberto Paccanaro, and Maribel Hernandez) and the instructor of the mini-course offered during the conference (Ronaldo Nogueira). Last but not least, we thank all the authors who made contributions to this conference, either in the form of paper or poster abstract, and the conference delegates; thank you very much, and we hope to see you all again in BSB 2025!

December 2024

Marcelo S. Reis
Fabricio M. Lopes

Organization

General Chairs

Sérgio Nery Simões	Instituto Federal do Espírito Santo, Brazil
Karin Satie Komati	Instituto Federal do Espírito Santo, Brazil

Local Organization Committee

Sérgio Nery Simões	Instituto Federal do Espírito Santo, Brazil
Karin Satie Komati	Instituto Federal do Espírito Santo, Brazil
Francisco de Assis Boldt	Instituto Federal do Espírito Santo, Brazil
Hilário Tomaz Alves de Oliveira	Instituto Federal do Espírito Santo, Brazil
Leandro Colombi Resendo	Instituto Federal do Espírito Santo, Brazil

Technical Program Committee Chairs

Marcelo S. Reis	Universidade Estadual de Campinas, Brazil
Fabricio M. Lopes	Universidade Tecnológica Federal do Paraná, Brazil

Steering Committee (CE-BioComp)

Sérgio Lifschitz	Pontifícia Universidade Católica do Rio de Janeiro, Brazil
Kele Belloze	Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Brazil
Daniel de Oliveira	Universidade Federal Fluminense, Brazil
Diogo Tschoeke	Universidade Federal do Rio de Janeiro, Brazil
João Carlos Setubal	Universidade de São Paulo, Brazil
Marcelo S. Reis	Universidade Estadual de Campinas, Brazil
Raquel C. de Melo-Minardi	Universidade Federal de Minas Gerais, Brazil
Sérgio Nery Simões	Instituto Federal do Espírito Santo, Brazil

Technical Program Committee

Adriano Werhli	Universidade Federal do Rio Grande, Brazil
Alessandra Silva	Universidade Federal de Minas Gerais, Brazil
André Fujita	Universidade de São Paulo, Brazil
Andre Kashiwabara	Universidade Tecnológica Federal do Paraná, Brazil
Carlos da Silveira	Universidade Federal de Itajubá, Brazil
Christian H. zu Siederdissen	Friedrich-Schiller-Universität Jena, Germany
Daniel de Oliveira	Universidade Federal Fluminense, Brazil
David Martins-Jr	Universidade Federal do ABC, Brazil
Didier Vega-Oliveros	Universidade Federal de São Paulo, Brazil
Diego Mariano	Universidade Federal de Minas Gerais, Brazil
Dieval Guizelini	Universidade Federal do Paraná, Brazil
Diogo Tschoeke	Universidade Federal do Rio de Janeiro, Brazil
Emely Silva	Universidade Estadual de Campinas, Brazil
Eneas Carvalho	Instituto Butantan, Brazil
Felipe Louza	Universidade Federal de Uberlândia, Brazil
Giuseppe Leite	Universidade Federal de São Paulo, Brazil
Glaucia Bressan	Universidade Tecnológica Federal do Paraná, Brazil
Guilherme Telles	Universidade Estadual de Campinas, Brazil
João Carlos Setubal	Universidade de São Paulo, Brazil
João Meidanis	Universidade Estadual de Campinas, Brazil
José Patané	Instituto do Coração (InCor), Brazil
Karina dos Santos Machado	Universidade Federal do Rio Grande, Brazil
Kele Belloze	Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Brazil
Liliane Oliveira	Universidade Tecnológica Federal do Paraná, Brazil
Luís Cunha	Universidade Federal Fluminense, Brazil
Luiz Manoel R. Gadelha Júnior	Laboratório Nacional de Computação Científica, Brazil
Manuel Lafond	University of Sherbrooke, Canada
Marcelo Brigido	Universidade de Brasília, Brazil
Mariana Recamonde-Mendoza	Universidade Federal do Rio Grande do Sul, Brazil
Maribel Hernandez	CINVESTAV Irapuato, Mexico
Miguel Rocha	Universidade do Minho, Portugal
Milton Nishiyama-Jr	Instituto Butantan, Brazil
Mirela Darc	Universidade Federal do Rio de Janeiro, Brazil
Natasha Andressa N. Jorge	Leipzig University, Germany
Robson Bonidia	Universidade Tecnológica Federal do Paraná, Brazil
Ronaldo Hashimoto	Universidade de São Paulo, Brazil
Sérgio Lifschitz	Pontifícia Universidade Católica do Rio de Janeiro, Brazil
Sérgio Nery Simões	Instituto Federal do Espírito Santo, Brazil
Seyed Jamalaldin Haddadi	Universidade Estadual de Campinas, Brazil
Valerie Anda	University of Florida, USA
Víctor Martinez	Universidade Estadual de Campinas, Brazil
Zanoni Dias	Universidade Estadual de Campinas, Brazil

Realization

Sociedade Brasileira de Computação (SBC), Brazil

Financial Support

Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil

Institutional Organization

Instituto Federal do Espírito Santo (IFES), Brazil

Brazilian Symposium on Bioinformatics 2024

POSTER SESSION ALPHA ABSTRACTS

(Abstracts are sorted by title ascending order)

Cidade da Inovação (IFES), Vitória,
December 2, 2024

Summary

P29 – Group I introns in the mitochondrial genomes of Trichoderma spp.

P31 – Homology-based quantification of the fibrosis in the CT image: A proof of concept for CT image feature-assisted gene expression prediction

P33 – In silico validation of Linear B-Cell epitopes using Machine Learning: A proposed approach with organisms of genus Trypanosoma

P35 – Metabarcoding reveal Fusarium decemcellulare as the potential causal agent of the emergent disease in Coffea canephora

P37 – Molecules of the Amazon: Integration and Centralization of Data on the Amazon Flora and its Biomolecules

P39 – Network Pharmacology and UHPLC-ESI-Q-TOF-MS/MS Approaches to Explore Active Compounds and Mechanisms of Acerola Seed Hydroethanolic Extract in Obesity Treatment

P41 – Nuclear Segmentation of Oncology Microscopy Images through Convolutional Neural Networks: A Comparative Analysis

P43 – Predicting side effects of drug combinations in realistic experimental settings

P45 – Sialic Acid Enzymes Database

P47 – The potential role of the JAK/STAT pathways in the progression of depressive and anxiety disorders in Long COVID

P49 – The role of polyploid giant cells in cancer progression and their potential as therapeutic targets: a bioinformatic overview

P51 – Transcriptomic analysis of Crassostrea gigas oysters exposed to tamoxifen demonstrates alterations in cancer-associated metabolic pathways

P53 – Transcriptomic Analysis Reveals a Novel MicroRNA in Porcine Fetuses from Gilds Supplemented with L-Arginine

P55 – Unlocking The Anti-aging Potential: In silico Analysis of Astaxanthin, Curcumin, Quercetin, and Resveratrol in Modulating Skin Aging Pathways

P57 – Large scale analysis of sialic acid incorporation mechanisms of microorganisms from intestinal microbiota

Brazilian Symposium on Bioinformatics 2024

A bioinformatics approach to define host-element boundaries using *Microviridae* phages and casposons as study models

Giuliana L. Pola[1], Adam L.L. Ramalho[2], Julia S. Fiasca[1], Luciano A. Digiampietri[3], Arthur Gruber[4]

1. Biotechnology undergraduate course, EACH/USP, São Paulo, Brazil
2. Biological Sciences undergraduate course, Instituto de Biociências, IB/USP, São Paulo, Brazil
3. Escola de Artes, Ciências e Humanidades, EACH/USP, São Paulo, Brazil
4. Instituto de Ciências Biomédicas, ICB/USP, São Paulo, Brazil

Mobile genetic elements (MGEs), such as viruses and transposons, can integrate into the genomes of prokaryotic and eukaryotic hosts. The process involves several mechanisms that generate direct and/or inverted terminal repeats flanking the element. The precise identification of host-element boundaries may help to illuminate the possible insertion mechanisms. This work aimed to develop bioinformatics tools to identify inserted elements, terminal repeat signatures, and sequence patterns associated with the MGEs. We used two study models, the first consisting of *Microviridae*, a family of phages with icosahedral capsids and single-stranded DNA genomes that infect bacterial hosts [Roux et al. 2012]; and the second comprising casposons, self-synthesizing transposable elements found in Bacteria and Archaea [Krupovic et al. 2014].

We have previously constructed profile HMMs directed towards multiple sequence alignments of VP1, VP2 and VP4 proteins for *Microviridae* phages, and Cas1 and PolB for casposons, using TABAJARA program [Gruber et al. 2023]. The models were employed to perform a large scale survey of MGEs using e-Finder, a tool for the detection of multigene elements in syntenic context [Oliveira and Gruber 2021], against the PATRIC database, a collection of over 600,000 bacterial and archaeal genomes.

We present the development of two Python programs that employ complementary approaches to identifying prophage insertion sites. `Insertion_finder` compares prophages flanked by the host 5' and 3' sequences against homologous genomic sequences, presenting or not the corresponding elements. Based on similarity search results, the program determines the 5' and 3' boundaries and generates feature table annotation files compatible with graphical annotation tools like `Artemis`. The second program, `Select_repeats`, executes the `find-repeat` tool from the `UGENE` package to locate pairs of direct and/or inverted repeats. It then selects the most promising candidates according to defined criteria and generates a feature table annotation output. Using a previously characterized dataset of 400 *Microviridae* prophages, we combined these programs with manual curation, to accurately identify prophage-host boundaries in 140 instances. A similar approach was successfully applied to detect the boundaries of casposon elements. Finally, these results were manually curated and used as training sets for developing machine learning methods capable of detecting *Microviridae* and casposons.

For the machine learning model, we used 10-fold cross-validation to select both the algorithm and its parameters. We selected the Random Forest algorithm from the Python sklearn library, using its default parameters. As features, we extracted n-grams of characters (with *n* ranging from 1 to 3) from

the nucleotide sequences. The KBest algorithm was then applied to select the 500 most important features.

For the performance evaluation, the training set consisted of 80% of the positive sequences (*Microviridae* or casposons), and for each positive sequence, 100 sequences of the same length were randomly extracted from different genomes to form the negative set. The test set comprised the remaining 20% of the positive sequences, and for each positive sequence, 500 sequences of the same length were randomly extracted from different bacterial genomes as negatives.

The results for *Microviridae* showed 100% precision and recall for the positive class, and 100% precision with 67% recall for the negative class, leading to a macro F1 score of 0.83. In the case of casposons, we achieved over 99% precision and recall for the negative class, and 72% precision with 87% recall for the positive class, resulting in a macro F1 score of 0.93.

The comparison between host sequences containing or not the element, together with the identification of terminal repeats in a proper syntenic context, allowed us to successfully identify most *Microviridae* prophages and the respective host-element boundaries. In the case of casposons, we observed a much lower occurrence of host genomes lacking the respective element. Also, a much lower sequence conservation across repeats was found. Both results suggest that these elements have been inserted in ancient times and may have lost the ability to excise themselves from the respective host genomes.

Acknowledgements

GLP and JSF received fellowships from FAPESP; ALLR received a fellowship from the Programa Unificado de Bolsas (PUB) of the University of São Paulo. Correspondence: argruber@usp.br

Keywords: *Mobile genetic elements, casposons, Microviridae, insertion sites, machine learning*

References

Krupovic, M., Makarova, K.S., Forterre, P., Prangishvili, D. and Koonin, E.V. (2014). Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity, *BMC Biol*, vol.12, num.36, p.1-12.

Oliveira, L.S. and Gruber, A. (2021). Rational Design of Profile Hidden Markov Models for Viral Classification and Discovery, *Bioinformatics*, H. Nakaya, England, Academic Press Ltd., p. 449-460.

Oliveira, L.S., Reyes, A., Dutilh B.E. and Gruber, A. (2023). Rational Design of Profile HMMs for Sensitive and Specific Sequence Detection with Case Studies Applied to Viruses, Bacteriophages, and Casposons, *Viruses*, vol.15, num.2, p. 519.

Roux, S., Krupovic, M., Poulet, A., Debroas, D. and Enault, F. (2012). Evolution and Diversity of the Microviridae Viral Family through a Collection of 81 New Complete Genomes Assembled from Virome Reads, *PloS One*, vol.7, num.7, e40418.

Brazilian Symposium on Bioinformatics 2024

A genomic approach to the selection of growth-promoting bacteria in maize using reverse ecology

Mirelly Jady Fernandes e Silva [1], Luciano Nascimento de Almeida [1], João Paulo Lopes da Rocha [1], Sávio Ferreira Mendes [1], Guilherme de Castro Gonçalves [1], Gabriela Amaral Xavier [1], Blenda de Freitas Rodrigues Jesuino [1], Mateus Ferreira Santana [1]

1. Eco-evolutionary Microbial Genomics Group, Molecular Genetics of Microorganisms Laboratory, Department of Microbiology, Institute of Applied Biotechnology to Agriculture, Federal University of Viçosa, Minas Gerais, Brazil.

Plant growth-promoting rhizobacteria (PGPR) are beneficial bacteria that colonize the roots or nearby soil areas. They promote plant growth by assisting in nutrient assimilation and modulating signaling pathways. Typically, PGPR are selected through laboratory tests, which include cultivation and/or direct inoculation into the plant. Although screening techniques are efficient, they are limited by experimental design, space, and the number of isolates. With advances in genomics, new approaches for selecting PGPR have become possible, such as reverse ecology, which combines genomic analyses and metabolic network modeling to infer ecological interactions between organisms within their communities.
Five and four bacterial isolates belong to the collection of the Microbial Eco-Evolutionary Genomics Group (GGEM), located at the Institute of Biotechnology Applied to Agriculture at the Federal University of Viçosa (Bioagro/UFV). These isolates were previously classified by genus based on 16S rRNA sequencing. Protein-coding genes associated with KEGG orthology (KOs) were retrieved from all species of the genera present in the collection through the JGI Integrated Microbial Genomes and Microbiomes database. Additionally, genomic annotation profiles and FASTA protein files were obtained from the National Center for Biotechnology Information (NCBI) and analyzed using KofamKoala, with an e-value set to 0.0001. KO data were then employed in RevEcoR to obtain edge lists, which were subsequently used by the NetCooperate module to infer ecological interactions between the bacterial isolates and *Zea mays*. Bacterial genera with biosynthetic support indices between 0.5 and 1, with values close to *Azospirillum brasilense* (0.555455), were analyzed. Biosynthetic gene clusters (BGCs) were identified using antiSMASH. Protein sequences will be used to predict crop growth-promoting features through the PGPg_finder pipeline. The identified compounds and their metabolic functions will be recorded using the KEGG Compound databases. Differences in metabolic complementarity indices were not significant, but biosynthetic support and competition indices were considered. The genus *Klebsiella* scored the highest (0.618084), followed by *A. brasilense*, *Duganella*, and *Kosakonia* (0.545432), among others. The genus *Rugosimonospora* (0.206612), on the other hand, obtained the lowest score. 492 *seeds* from twelve selected genera were annotated; of these, 74 corresponded to compounds related to plant growth-promoting mechanisms. Among these, the observation of indole-3-acetic acid, organic acids, and siderophores stood out. Genes for seven different mechanisms, both direct and indirect, of plant growth promotion were identified, along with biosynthetic gene clusters to produce exopolysaccharides. Future perspectives should consider conducting in vivo tests to validate the effects of reverse ecology-selected rhizobacteria on plants. However, this study aims to explore new approaches

for the selection of plant growth-promoting rhizobacteria (PGPR), promoting the integration of genomics into agricultural techniques.

Keywords: Ecology, Bacteria, Compounds, Metabolic, Nutrient, Agricultural.

References

Blin, K., Shaw, S., Kloosterman, A. M., Charlop-Powers, Z., Wezel, G. P., Medema, M. H., Weber, T. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic acids research, v. 49, n. W1, p. W29-W35.

Borenstein, E., Kupiec, M., Feldman, M. W. and Ruppin, E. (23 sep 2008). Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proceedings of the National Academy of Sciences of the United States of America*, v. 105, n. 38, p. 14482–14487.

Borenstein, E. and Feldman, M. W. (feb 2009). Topological signatures of species interactions in metabolic networks. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, v. 16, n. 2, p. 191–200.

Cao, Y., Wang, Y., Zheng, X., Li, F. and Bo, X. (29 jul 2016). RevEcoR: an R package for the reverse ecology analysis of microbiomes. *BMC Bioinformatics*, v. 17, n. 1, p. 294.

Carr, R. and Borenstein, E. (1 mar 2012). NetSeed: a network-based reverse-ecology tool for calculating the metabolic interface of an organism with its environment. *Bioinformatics (Oxford, England)*, v. 28, n. 5, p. 734–735.

Chaudhari, N. M., Gupta, V. K., Dutta, C. (2016). BPGA-an ultra-fast pan-genome analysis pipeline. Scientific reports, v. 6, n. 1, p. 24373.

Chhetri, G., Kim, I., Kim, J., et al. (4 apr 2023). Paraburkholderia tagetis sp. nov., a novel species isolated from roots of Tagetes patula enhances the growth and yield of Solanum lycopersicum L. (tomato). *Frontiers in Microbiology*, v. 14.

Mahreen, N., Yasmin, S., Asif, M., et al. (23 jan 2023). Mitigation of water scarcity with sustained growth of Rice by plant growth promoting bacteria. *Frontiers in Plant Science*, v. 14.

Ming, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Haeseler, A., Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular biology and evolution, v. 37, n. 5, p. 1530-1534.

Parks, D., Imelfort, M., Skennerton, C. T., Hugenholtz, P., Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research, v. 25, n. 7, p. 1043-1055.

Zheng, J., Ge, Q., Yan, Y., Zhang, X., Huang, L., Yin, Y. (2023). dbCAN3: automated carbohydrate-active enzyme and substrate annotation. Nucleic Acids Research, v. 51, n. W1, p. W115-W121.

Brazilian Symposium on Bioinformatics 2024

A T_m-value prediction system and molecular dynamics analysis of AmNA-containing gapmer antisense oligonucleotides

Masataka Kuroda [1,2], Yuuya Kasahara [1,3], Masako Hirose [4], Harumi Yamaguma [1], Masayuki Oda [5], Chioko Nagao [6], Kenji Mizuguchi [6,1]

1. National Institutes of Biomedical Innovation, Health and Nutrition (NIBIOHN), Osaka, Japan
2. Mitsubishi Tanabe Pharma Corporation, Yokohama, Japan
3. Graduate School of Pharmaceutical Science, Osaka University, Osaka, Japan
4. Malvern Panalytical, Spectris, Tokyo, Japan
5. Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, Kyoto, Japan
6. Institute for Protein Research, Osaka University, Osaka, Japan

Nucleic acid medicines are a new class of drugs that contain RNase H-dependent antisense oligonucleotides (gapmer ASOs) that bind to target RNA to recruit RNase H for the RNA cleavage and regulate the expression of disease-associated proteins. Since gapmer ASOs, which consist only of natural nucleotides, have low stability in the human body, the introduction of artificial nucleic acids into them accelerates their therapeutic efficacy. In particular, amido-bridged nucleic acid (AmNA), an artificial nucleic acid with a bridge in the sugar ring to fix the three-dimensional conformation of the ring that stabilizes the duplex, has the potential to confer high affinity and stability through nuclease resistance. It also reduces the hepatotoxicity that is seen in other bridged-type artificial nucleic acids.

Melting temperature (T_m), defined as the temperature at which equal populations of the double-stranded oligonucleotides and monomer exist, is commonly used as an activity reference in the discovery research because higher T_m values indicate a stronger binding affinity of the ASO to its target RNA. As there were no prediction systems for oligonucleotides including AmNA, a T_m prediction model for them was newly developed. We chose a machine learning method, lightGBM, to build the prediction model, where the features were defined as the number of neighboring nucleotide pairs and guanine or cytosine nucleobases extracted from the gamper ASO sequence. In this study, we prepared the T_m values of 157 oligonucleotides by using differential scanning calorimetry (DSC) which allows the study of molecular denaturation or unfolding by heating a solution containing the molecule, and developed a highly accurate T_m prediction model. In the performance test performed 10 times, the average of mean squared error (MSE) was 15.1 and the coefficient of determination (R^2) was 0.840. A T_m prediction web service was created and is now used for drug discovery projects in our laboratory.

In addition, molecular dynamics (MD) simulations were performed to clarify how mutations to AmNA increase T_m. Three gamper ASOs with the identical GC ratios, similar T_m values and varied AmNA substitution patterns were selected. All-DNA oligonucleotides with identical sequences to these three ASOs were also prepared for comparison. The simulation started at 300 K and the system was heated to 500 K to accelerate duplex collapse. The MD trajectories show that the duplex dissociation pattern depends on the sequence and the position at which AmNA is introduced. DSC detects not only the T_m value but also the enthalpy change (ΔH) of a thermal transition. Although they are relatively correlated, the three oligomers have different ΔH values. To understand this, water molecule motions were analyzed from MD trajectories and the detailed results will be explained.

Keywords: T_m value prediction, machine learning, molecular dynamics (MD), amido-bridged nucleic acid (AmNA), RNase H-dependent antisense oligonucleotide (gapmer ASO)

References

Kuroda, M., Kasahara, Y., Hirose, M., Yamaguma, H., Oda, M., Nagao, C. and Mizuguchi, K. (2024). Construction of a T_m-value prediction model and molecular dynamics study of AmNA-containing gapmer antisense oligonucleotide., *Mol. Ther. Nucleic Acids.*

Brazilian Symposium on Bioinformatics 2024

Antimicrobial Resistance Genes in Oligotrophic Ecosystems: An Evolutionary Insight from Microbialites and Microbial Mats of Cuatro Ciénegas

Maribel Hernánez-Rosales[1], Manuel A. Barrios-Izás[2], Erika Cruz-Bonilla[1], Katia Aviña-Padilla[1], J. Norberto García-Miranda[3], Africa Islas-Robles[3] and Gabriela Olmedo-Alvarez[3]

1. Laboratorio de Bioinformática y Redes Complejas, CINVESTAV-Irapuato, Libramiento Norte Carretera Irapuato León Kilómetro 9.6, Irapuato, 36821, Guanajuato, Mexico
2. Instituto de Investigaciones, Centro Universitario de Zacapa, Universidad de San Carlos de Guatemala, Entrada a Pueblo Modelo, 19001, Zacapa, Zacapa, Guatemala.
3. Laboratorio de Biología Molecular y Ecología Microbiana CINVESTAV-Irapuato, Libramiento Norte Carretera Irapuato León Kilómetro 9.6, Irapuato, 36821, Guanajuato, Mexico

The Cuatro Ciénegas Basin, Mexico, in the Chihuahuan Desert is a xerophytic ecosystem with a unique mosaic of springs, streams, and pools—remnants of ancient marine environments—that host diverse endemic species (Souza et al., 2006; Alcaraz et al, 2008). Unlike clinical settings dominated by pathogenic antibiotic-resistant microbes, this environment offers a window into pre-antibiotic-era resistance mechanisms, including those in non-pathogenic bacteria.

These aquatic systems support diverse microbial communities that engage in intense competition for nutrients in oligotrophic conditions (Elser et al., 2006; Escalante et al., 2008; Pajarez et al. 2015). This microbial competition often involves the production of antibiotics by some bacteria to inhibit or kill competitors, prompting others to evolve defense mechanisms (Pérez Gutiérrez et al. 2013; Zapien et al 2015). To explore these natural bacterial defenses, we sequenced bacterial metagenomes from microbialites and microbial mats.

We hypothesized that understanding the evolution of antimicrobial resistance genes (ARGs) in these non-clinical settings could offer insights into mitigating resistance in clinical environments (Von Wintersdorff et al., 2016). Samples from distinct layers of microbial mats and microbialites were collected, processed, and analyzed to reveal their taxonomic composition and functional capabilities. ARGs were identified using the Comprehensive Antibiotic Resistance Database (CARD), shedding light on how these genes are distributed and function within different environmental contexts. Since microbes in these layered ecosystems experience varying conditions—such as light and oxygen gradients—decoding their taxonomy and gene functions offers insights into their specific adaptations and interactions (Tang & Roopnarine, 2003; García-Pichel et al., 2008).

Our analysis revealed a diversity of microbial taxa involved in key processes like nitrogen fixation, carbon cycling, photosynthesis, and organic matter degradation. At the genus level, we identified bacteria with specialized roles, including those adapted to the mineralized environments of microbialites and those participating in sulfur cycling). These findings highlight the intricate networks within these communities and help us understand how natural antibiotic resistance has evolved in response to environmental pressures rather than clinical antibiotic use. We identified various ARGs, including those involved in efflux pumps, target modification, and antibiotic degradation, reflecting different strategies to counteract antimicrobial threats. While microbialites and microbial mats shared several ARGs, each also contained unique gene families. Principal Component Analysis (PCA) showed distinct ARG profiles between the layers of microbialites and mats, consistent with their taxonomic differences. Genes linked to efflux pumps were widespread, indicating an evolutionary response to the presence of toxic compounds. Other resistance mechanisms, such as glycopeptide resistance and ribosomal RNA methylation, suggest strategies for maintaining cellular integrity under stress.

Our data also uncovered viral sequences, including giant Pandoravirus and phages like Kayvirus, underscoring the interkingdom interactions within these communities. These viruses play a role in shaping microbial dynamics, influencing gene transfer, and regulating microbial populations in these ancient ecosystems (Cisneros-Martínez et al., 2023).

Additionally, the presence of *Sorangium cellulosum*, a bacterium known for producing bioactive compounds, suggests significant potential for bioprospecting in Cuatro Ciénegas to discover new antibiotics. *S. cellulosum* has been extensively studied for its secondary metabolites, highlighting its promise for developing new antimicrobial agents (Gao et al. 2023).

These findings enhance our understanding of microbial evolution and the natural development of resistance mechanisms, providing a valuable foundation for strategies to combat antibiotic resistance in modern clinical settings.

Keywords: *Ancient Ecosystems, Antimicrobial Resistance, Genomics, Microbial Mats, Microbialites.*

References

Alcaraz, L. D., Olmedo, G., Bonilla, G., Cerritos, R., Hernández, G., Cruz, A., Ramírez, E., Putonti, C., Jiménez, B., Martínez, E. V., López, V., Arvizu, J. L., Ayala, F. J., Razo, F., Caballero, J., Siefert, J. L., Eguiarte, L. E., Vielle, J.-P., Martínez, O., Souza, V., & Herrera-Estrella, L. (2008). The genome of *Bacillus coahuilensis* reveals adaptations essential for survival in the relic of an ancient marine environment. *Proceedings of the National Academy of Sciences*, *105*(36), 13442-13447. https://doi.org/10.1073/pnas.0800981105

Cisneros-Martínez, A. M., Eguiarte, L. E., & Souza, V. (2023). Metagenomic comparisons reveal a highly diverse and unique viral community in a seasonally fluctuating hypersaline microbial mat. *Microbial Genomics*, *9*(1), 001063. https://doi.org/10.1099/mgen.0.001063

Escalante, A. E., Eguiarte, L. E., Espinosa-Asuar, L., Forney, L. J., Noguez, A. M., & Souza, V. (2008). Diversity of aquatic prokaryotic communities in the Cuatro Cienegas Basin. *FEMS Microbiology Ecology*, *65*(1), 12-22. https://doi.org/10.1111/j.1574-6941.2008.00496.x

Elser, J. J., Watts, J. M., Schampel, J. H., & Farmer, J. D. (2005). Early Cambrian food webs on a trophic knife-edge? A hypothesis and preliminary data from a modern stromatolite-based ecosystem. *Ecology Letters*, *8*(8), 792-802. https://doi.org/10.1111/j.1461-0248.2005.00873.x

García-Pichel, F., Wade, B. D., & Farmer, J. D. (2002). Jet-suspended, calcite-ballasted cyanobacterial waterwarts in a desert spring. *Journal of Phycology*, *38*(6), 871-879. https://doi.org/10.1046/j.1529-8817.2002.01178.x

García-Ulloa, M., Escalante, A. E., Letelier, A. M., Eguiarte, L. E., & Souza, V. (2020). Fast eco-evolutionary changes in bacterial genomes after anthropogenic perturbation. *bioRxiv*. https://doi.org/10.1101/2020.03.13.990432

Gao, Y., Birkelbach, J., Fu, C., Herrmann, J., Irschik, H., Morgenstern, B., Hirschfelder, K., Li, R., Zhang, Y., Jansen, R. and Müller, R., 2023. The Disorazole Z Family of Highly Potent Anticancer Natural Products from Sorangium cellulosum: Structure, Bioactivity, Biosynthesis, and Heterologous Expression. *Microbiology Spectrum*, *11*(4), pp.e0073- 23. https://doi.org/10.1128/spectrum.00730-23

Pajares, S., Souza, V., & Eguiarte, L. E. (2015). Multivariate and phylogenetic analyses assessing the response of bacterial mat communities from an ancient oligotrophic aquatic ecosystem to different scenarios of long-term environmental disturbance. *PLOS ONE*, *10*(3), e0119741. https://doi.org/10.1371/journal.pone.0119741

Pérez-Gutiérrez, Rocío-Anaís, et al. (2015). Antagonism influences assembly of a Bacillus guild in a local community and is depicted as a food-chain network. *The ISME Journal* 7.3, 487-497. https://doi.org/10.1038/ismej.2012.119

Souza, V., Espinosa-Asuar, L., Escalante, A. E., Eguiarte, L. E., Farmer, J. D., Forney, L. J., Lloret, L., Rodríguez-Martínez, J. M., Soberón, X., Dirzo, R., & Elser, J. J. (2006). An endangered oasis of aquatic microbial biodiversity in the Chihuahuan Desert. *Proceedings of the National Academy of Sciences*, *103*(17), 6565-6570. https://doi.org/10.1073/pnas.0601434103

Tang, C. M., & Roopnarine, P. D. (2003). Complex morphological variability in complex evaporitic systems: Thermal spring snails from the Chihuahuan Desert, Mexico. *Astrobiology*, *3*(1), 283-292. https://doi.org/10.1089/153110703322610681

Von Wintersdorff, C. J., Penders, J., Van Niekerk, J. M., Mills, N. D., Majumder, S., Van Alphen, L. B., ... & Wolffs, P. F. (2016). Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. Frontiers in microbiology, 7, 173. https://doi.org/10.3389/fmicb.2016.00173

Zapién-Campos, R., Olmedo-Álvarez, G., & Santillán, M. (2015). Antagonistic interactions are sufficient to explain self-assemblage of bacterial communities in a homogeneous environment: A computational modeling approach. *Frontiers in Microbiology*, *6*, 489. https://doi.org/10.3389/fmicb.2015.00489

Brazilian Symposium on Bioinformatics 2024

Beyond mutations: Human Cytomegalovirus as a driver of heterogeneity in Glioblastoma.

Bianca Paulino Campanharo [1], Isabele Pagani Pavan1[1], Matheus Correia Casotti[1], Flávio dos Santos Alvarenga[1], Debora Dummer Meira[1], Iúri Drumond Louro[1]

1. Núcleo de Genética Humana e Molecular (NGHM), Universidade Federal do Espírito Santo (UFES)

Glioblastoma (GB) is one of the most aggressive tumors of the central nervous system, making up the so-called gliomas. GB is highly plastic and heterogeneous, providing mutant clones capable of resisting conventional treatments. At the same time, the human cytomegalovirus (HCMV), which is present in between 40% and 90% of the world's population, has been considered an oncovirus capable of crossing the blood-brain barrier. In addition, HCMV has the ability to induce severe mutations by deactivating apoptosis and increasing cell proliferation. In this sense, the relationship between HCMV targets and cell deregulation potentially correlated with GB initiation or progression was evaluated. Thus, we aimed to analyze the relationship between viruses and their influence on promoting or complementing the generation of heterogeneity in GB. To this end, holistic computational analyses were carried out with manual curation, using a bibliographic search in PUBMED, with the following complementary strategies: “Heterogeneity” AND “Glioblastoma” AND “Cytomegalovirus” & ("Glioblastoma" OR "Glioma") AND ("Histones" OR "Epigenetic" OR "Mutation") AND "HCMV", with the descriptors present in the title or abstract; and then the construction of the protein interaction network (PPIN) was carried out, using the protocol established in Casotti et al. (2022). The functional results were obtained using a pathway enrichment strategy, highlighting relationships with “Regulation of immune system process”, “Protein binding”, “External side of plasma membrane”, “Cytokine-cytokine receptor interaction” and “Immune System and Allograft rejection” (with FDR, for Biological process, Molecular function, Cellular component, KEGG, Reactome and WikiPathways respectively), for the first search strategy, in order to PPIN. In addition, the second strategy applied obtained relationships with “Chromatin organization”, “Structural constituent of chromatin”, “Chromosome”, “Alcoholism”, “Chromatin modifying enzymes” and “Histone modifications”, obtaining significant FDRs. Through these results, these findings corroborate the characteristic functional aspect of HCMV, being a virus associated with alterations in gene expression linked to pluripotency, maintenance of tumor stem cells and loss of apoptotic function. Together, these actions contribute to phenotypic diversification in tumors, promotion of an immunosuppressive environment and gene mutations linked to morphonuclear changes, given their genetic and epigenetic modulation. This study aims to deepen the relationships present between the virus and its influence on histones from the previous work. In short, new molecular targets can be highlighted as protagonists of a complex mechanistic process associated with the somewhat less elusive mechanism of action of HCMV on tumor heterogeneity in GB. However, more studies are needed to provide better molecular and cellular clarification. To this end, future *in vitro* and *in vivo* studies could make use of these preliminary *in silico* findings to provide further evidence of this onco-viral relationship.

Keywords: *Glioblastoma, Cytomegalovirus, Heterogeneity, Computational Biology, System Biology.*

References

CASOTTI, Matheus Correia et al. (2022). Construindo Redes de Interação Proteína-Proteína por Curadoria Manual. In: BASTOS, Luana Luiza et al. BIOINFO# 02-Revista Brasileira de Bioinformática e Biologia Computacional (Vol.2) Alfahelix Publicações.

DAEI SORKHABI, A., Sarkesh, A., Saeedi, H., Marofi, F., Ghaebi, M., Silvestris, N., ... & Brunetti, O. (2022). The basis and advances in clinical application of cytomegalovirus-specific cytotoxic T cell immunotherapy for glioblastoma multiforme. Frontiers in Oncology, 12, 818447.

EL BABA, R., Pasquereau, S., Haidar Ahmad, S., Monnien, F., Abad, M., Bibeau, F., & Herbein, G. (2023). EZH2-Myc driven glioblastoma elicited by cytomegalovirus infection of human astrocytes. Oncogene, 42(24), 2031-2045.

HERBEIN, G. (2018). The human cytomegalovirus, from oncomodulation to oncogenesis. Viruses, 10(8), 408.

HERBEIN, G. (2022). Tumors and cytomegalovirus: An intimate interplay. Viruses, 14(4), 812.

Brazilian Symposium on Bioinformatics 2024

Characterization of Psychiatric Disorders Cataloged in the DSM-5 Through a Network Approach

A. C. Nascimento[1], F. Ferreira[1,2], D. M. Gysi [3,*]

[1] *University of Maia, Maia, Portugal,*
[2] *Center for Psychology at University of Porto, Porto, Portugal*
[3] *Department of Statistics, Federal University of Paraná, Curitiba, Brazil.*
[*]*D.M.G. deisy.gysi@ufpr.br*

Mental disorders represent significant challenges for clinical diagnosis due to the extensive symptom overlap among disorders cataloged in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). The difficulties permeating the accurate identification of each disorder is aggravated by the similarity of symptoms across different conditions, often resulting in imprecise diagnoses. In this context, Network Science emerges as a promising approach for clinical psychology, offering tools for modeling complex interactions and promoting a more comprehensive understanding of symptom interrelations.

This study proposes a network-based approach to characterize psychiatric disorders as described in the DSM-5, utilizing a combination of network analysis, dynamic questionnaires, and clinical data. The literature indicates that disorders grouped within the same DSM-5 chapter exhibit significant genetic overlap, suggesting that they share not only clinical characteristics but also proximity within a network of biological interactions. Thus, network methodology arises as a valuable tool for exploring both symptom connectivity and underlying biological aspects.

The present work is organized around three main objectives: (i) the construction of symptom networks to characterize primary and secondary symptoms of disorders cataloged in the DSM-5; (ii) the detailed characterization of the main symptoms within each chapter and disorder; and (iii) the development of a dynamic, network-based inventory to facilitate personalized and rapid diagnostics.

To achieve the mapping of symptoms, psychiatric disorders are manually described based on the DSM-5, categorizing symptoms into specific networks. The properties of these networks will be analyzed to identify “bridge symptoms” — symptoms that link different disorders and aid in understanding comorbidities. Psychometric models are also applied to estimate symptom interactions, although questions remain about diagnostic accuracy and the resulting network structure of these models. Preliminary studies indicate that symptom networks assist in identifying key symptoms, such as in disorders like Post-Traumatic Stress Disorder (PTSD).

Additionally, the development of dynamic questionnaires aims to reduce the time required for diagnosis and the number of questions needed for precise evaluation, thereby making the diagnostic process more efficient and less exhaustive. By employing a network-based methodology, this study seeks to optimize clinical diagnosis and personalize treatment, promoting an integrated approach that combines clinical characteristics and biological data, ultimately contributing to diagnostic precision and individualized treatment in mental health.

Keywords: Precision Medicine; Network Medicine; Network Psychiatry

Brazilian Symposium on Bioinformatics 2024

ClusterONE Web: a web tool for detecting and visualizing overlapping protein complexes in protein-protein interaction networks

Marcelo Báez[1], Rubén Jiménez[1], María del Mar Sánchez Rojas[1], and Alberto Paccanaro[1,2]

1. Escola de Matemática Aplicada, Fundação Getúlio Vargas, Rio de Janeiro, Brazil
2. Department of Computer Science, Centre for Systems and Synthetic Biology, Royal Holloway University of London, Egham, UK

We present ClusterONE Web (https://paccanarolab.org/clusteroneweb/), a freely available, user-friendly web-based tool to identify, visualize, and analyze protein complexes. ClusterONE Web is based on the well-known ClusterONE algorithm, that was developed for detecting potentially overlapping protein complexes from protein-protein interaction (PPI) data [1]. Since its publication, ClusterONE has remained a powerful state-of-the-art method for clustering large-scale weighted and unweighted networks, allowing nodes to potentially belong to more than one cluster. We created an intuitive web interface that simplifies ClusterONE interaction and execution, making it accessible to a wider range of users. ClusterONE Web includes pre-loaded PPI networks from several biological databases: BIOGRID [2], INTACT [3], MINT [4], and DIP [5], making it easier for users to explore protein complexes across different organisms. It enables users to: visually explore the complexes identified by ClusterONE within the PPI network; navigate through overlapping complexes detected in the PPI network; and conduct an enrichment analysis of these complexes, offering functional insights based on Gene Ontology (GO) terms [6]. Proteins and GO terms are linked to their entries in UniProt [7] and EMBL-EBI QuickGO [8], providing a starting point for in-depth analysis and further understanding of the biological roles of the identified complexes. Additionally, ClusterONE Web offers the users the option to upload their own PPI data and GO annotation files, allowing for more tailored experiments and analysis. By combining the clustering capabilities of ClusterONE with functional enrichment, our web-based tool can provide additional insights into the organization and functionality of PPI networks and facilitate hypothesis generation. This accessible platform expands the utility of ClusterONE, making it more efficient and user-friendly for the broader scientific community.

Keywords: protein complexes prediction, proteomics, protein-protein interaction networks, functional enrichment analysis, ClusterONE, web-based tool

References

1. Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nature Methods, 9(5), 471–472.
2. Oughtred, R., et al. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Science, 30(1), 187–200.
3. Del Toro, N., et al. (2022). The IntAct database: efficient access to fine-grained molecular interaction data. Nucleic Acids Research, 50(D1), D648–D653.
4. Chatr-Aryamontri, A., et al. (2007). MINT: the Molecular INTeraction database. Nucleic Acids Research, 35(Suppl 1), D572–D574.
5. Salwinski, L., et al. (2004). The database of interacting proteins: 2004 update. Nucleic Acids Research, 32(Suppl 1), D449–D451.
6. Gene Ontology Consortium. "The Gene Ontology (GO) database and informatics resource." *Nucleic acids research* 32.suppl_1 (2004): D258-D261.
7. UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research, 47(D1), D506–D515.
8. Binns, D., et al. (2009). QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics, 25(22), 3045–3046.

Brazilian Symposium on Bioinformatics 2024

Comparative Genomic Analyses Highlight the Crucial Role of Mobile Genetic Elements in *Paenibacillus* strains

Blenda de Freitas Rodrigues Jesuino [1], Luciano Nascimento de Almeida [1], Sumaya Martins Tupy [1], Mirelly Jady Fernandes e Silva [1], Osiel Silva Gonçalves [1], João Paulo Lopes da Rocha [1], Gabriela Amaral Xavier [1], Mateus Ferreira Santana [1]

1. Eco-evolutionary Microbial Genomics Group, Molecular Genetics of Microorganisms Laboratory, Department of Microbiology, Institute of Applied Biotechnology to Agriculture, Federal University of Viçosa, Minas Gerais, Brazil.

The genus Paenibacillus encompasses widely distributed Gram-positive bacteria found in various ecological niches, including soil and rhizosphere. This broad distribution suggests that these bacteria possess diverse biological characteristics, reflecting a significant capacity for adaptation to different environments. This ecological variability may be related to the presence of Mobile Genetic Elements (MGEs), which facilitate the acquisition of new genes and enhance adaptability to various environmental conditions. In this study, we investigated the presence of MGEs, including integrative and conjugative elements (ICEs), mobilizable elements (IMEs), and plasmids, in 99 complete genomes of Paenibacillus. Additionally, we aimed to identify and characterize the accessory genes present in these elements through a set of comparative genomic analyses. To achieve this, we downloaded nucleotide sequences and protein files in FASTA format of complete genomes from RefSeq at the National Center for Biotechnology Information (NCBI) and used Geneious Prime as a viewer for these sequences. The detection of ICEs and IMEs was performed using ICEscreen v1.3.2, while transposons were identified through the databases TnCentral, Integrall, and ISFinder. Furthermore, we used previously identified and annotated plasmid sequences from NCBI for subsequent analyses. After identifying the elements, we utilized antiSMASH 6.0 to detect biosynthetic gene clusters of secondary metabolites and the PGPT-Pred module of PLaBAse to annotate genes related to plant growth promotion. A total of 81 MGEs were found, including 9 complete ICEs, 7 partial ICEs, 2 IMEs, 35 plasmids, 12 transposons, and 16 degenerated elements. The size of the elements varied from 834 bp to 213,282 bp and the GC content ranged from 36.0% to 54.4% in ICEs and IMEs; from 1,459 bp to 162,922 bp with GC content of 35.8% to 53.8% in transposons; and from 973 bp to 2,697,188 bp, with GC content of 37.1% to 52.0% in plasmids. The highest number of MGEs was found in genomes from soil samples (33). Genes related to plant growth promotion were more frequent in strains isolated from soil samples (43). Additionally, genes related to nitrogen fixation (nif), nodulation (nod), and phosphate solubilization (pho) had a higher number of copies in genetic elements from soil, particularly in plasmids, which had up to 20 copies of

these genes. In contrast, transposons and IMEs did not contain genes related to plant growth promotion. Moreover, plasmids had more regions with high identity to biosynthetic clusters of secondary metabolites (15) than other elements. These clusters are primarily associated with the production of antimicrobials, such as paenilipoheptin, as well as siderophores, which are essential for iron uptake in the environment. The predominance of MGEs in genomes from soil-isolated lineages may reflect the high abundance of these lineages in that environment. These findings are essential for understanding how MGEs facilitate the adaptation of the Paenibacillus genus, highlighting their potential for applications in agriculture.

Keywords: Adaptive genes; ICEs; Plant growth promotion; Transposons; Plasmids.

Blin, K., Shaw, S., Kloosterman, A. M., Charlop-Powers, Z., Wezel, G. P., Medema, M. H., Weber, T. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic acids research*, v. 49, n. W1, p. W29-W35.

Gardener, B. B. M. (2004). Ecology of *Bacillus* and *Paenibacillus* spp. in agricultural systems. *Phytopathology*. Anais, APS jornal, American Phytopathological Society.

Huang, W. C., Hu, Y., Zhang, G., Li, M. (2020*).* Comparative genomic analysis reveals metabolic diversity of different *Paenibacillus* groups. *Applied Microbiology and Biotechnology*, v. 104, p. 10133-10143.

Lao, J., Lacroix, T., Gérard Guédon, Coluzzi, C., Payot, S., Leblond-Bourget, N., Hélène Chiapello. (2022). ICEscreen: a tool to detect Firmicute ICEs and IMEs, isolated or enclosed in composite structures. *NAR Genomics and Bioinformatics*, 4(4).

Moura, A., Soares, M., Pereira, C., Leitao, N., Henriques, I., Correia, A. (2009). INTEGRALL: a database and search engine for integrons, integrases and gene cassettes. *Bioinformatics*, 25(8), 1096–1098.

Patz, S., Rauh, M., Gautam, A., Huson, D. H. (2024). mgPGPT: Metagenomic analysis of plant growth-promoting traits. bioRxiv.

Ross, K., Varani, A. M., Snesrud, E., Huang, H., Alvarenga, D. O., Zhang, J., Wu, C., McGann, P., Chandler, M. (2021). TnCentral: a Prokaryotic Transposable Element Database and Web Portal for Transposon Analysis. *mBio*, 12(5), e0206021.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., Chandler M. (2006). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Research*, 34(90001), D32–D36.

Xie, J., Shi, H., Du, Z., Wang, T., Liu, X., Chen, S. (2016*).* Comparative genomic and functional analysis reveal conservation of plant growth promoting traits in *Paenibacillus polymyxa* and its closely related species. *Nature - Scientific Reports*, v. 6, 9 fev.

Brazilian Symposium on Bioinformatics 2024

De novo transcriptome assembly and annotation for gene discovery in *Euterpe edulis*

Francine Alves Nogueira de Almeida[1], Layra Medeiros Cardozo[1], Miquéias Fernandes[1], Adésio Ferreira[1], Marcia Flores da Silva Ferreira[1].

1. Department of Agronomy, Federal University of Espírito Santo, Alegre, Espírito Santo, Brazil.

Bioeconomy is an emerging field that seeks to integrate the principles of biology and economics to promote sustainable development by using biological resources efficiently and responsibly. In the context of Brazilian biodiversity, *Euterpe edulis*, known as Juçara, plays an important role. This species is endemic to the Atlantic Forest and has gained attention for its nutritional characteristics and bioeconomic importance. The bioeconomy related to *Euterpe edulis* ranges from biodiversity conservation and sustainable management of its resources to the commercial exploitation of its products. This palm tree shows high genetic diversity among its populations and individuals with relevant phenotypic differences. In this context, obtaining a transcriptome for this species is essential to facilitate research on the different phenotypes it presents. Here, we present the first draft transcriptome of *Euterpe edulis*, derived from leaf and root tissues of two different phenotypes, using the Illumina Novaseq 6000 platform. Using normalized cDNA libraries, we generated comprehensive RNA-Seq datasets, resulting in a total of 1,334,593 contigs with an average length of 585 bp and an N50 value of 974 bp, through de novo transcriptome assembly. These unigenes were functionally annotated using the Basic Local Alignment Search Tool (BLAST) to query the Universal Protein Resource Knowledgebase (UniProtKB). A workflow covering RNA extraction, library preparation, transcriptome assembly, redundancy reduction, assembly validation, and annotation is provided. This study offers transcriptome and annotation data for *Euterpe edulis*, which are useful for gene discovery experiments, gene expression profiling, as well as for current and future applications in genome annotation and marker development.

Keywords: bioeconomy; Juçara; palm; RNA-Seq; transcriptomic.

References

Canal, G.B., Oliveira, G.F., de Almeida, F.A.N. *et al.* (2023). Genomic studies of the additive and dominant genetic control on production traits of *Euterpe edulis* fruits. *Sci Rep* **13**, 9795. https://doi.org/10.1038/s41598-023-36970-z

de Almeida, F.A.N., Santos, J.G., Pereira, A.G. *et al.* (2024). Genetic diversity analysis of *Euterpe edulis* based on different molecular markers. *Tree Genetics & Genomes* **20**, 31. https://doi.org/10.1007/s11295-024-01663-9

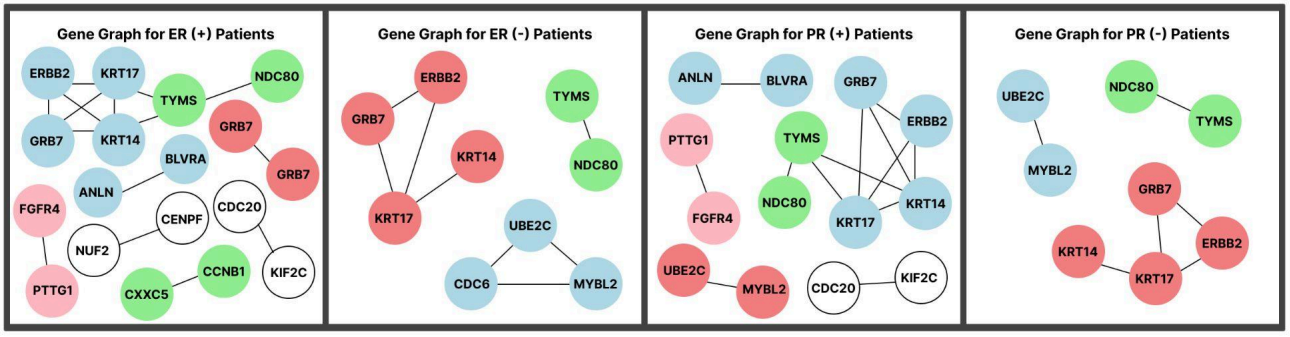
BSB 2024 Brazilian Symposium on Bioinformatics 2024

Discovery of Conditionally Independent Networks Among Gene Expressions in Breast Cancer Using Fast Step Graph

Grecia C. G. Rivera[1], Yunevda E. L. Rojas[1], Juan G. Colonna[1], Marcelo Ruiz[2]

1. Universidade Federal do Amazonas (UFAM), Instituto de Computação (IComp), Brasil
2. Universidad Nacional de Río Cuarto (UNRC), Departamento de Matemática, Argentina

The heterogeneity of the causes of breast cancer and the complex genetic interactions that characterize this neoplasm present significant challenges for understanding and treating the disease [Polyak 2011]. This study addresses the gap in understanding the conditional independence relationships among genes in breast cancer. Discovering networks of independent genes in genomic datasets is not a trivial task, since each patient is represented by a high-dimensional vector (p), and the number of patients (n) is small, i.e., $n \ll p$. When the number of variables (p) is greater than or close to the number of samples (n), the empirical covariance matrix Σ becomes singular or non-invertible, making it impossible to estimate the precision matrix Ω. This occurs because Σ is constructed from the available samples, and if the number of samples is not sufficiently large relative to the dimensionality of the data, there is not enough variation to adequately estimate the covariance between the variables [Lauritzen 1996, Liang and Jia 2023]. As a result, the matrix exhibits degeneracy issues, meaning that it is not possible to compute its inverse. To overcome this challenge, we propose the use of the Fast Step Graph (FSG) algorithm, an optimized version of StepGraph published on R-Cran, that executes in a fraction of the time compared to the original version [Zamar et al. 2021]. In this study, we used FSG to efficiently estimate the matrix Ω in proteogenomic samples from the Clinical Proteomic Tumor Analysis Consortium (CPTAC), specifically using the gene expression dataset, which includes 122 breast cancer patients and 23,691 gene expressions. In this dataset, the variables were standardized to zero mean and unit standard deviation. The CPTAC includes a pre-selected version with the 50 most significant genes called PAM50 [Parker et al. 2009, Liu et al. 2016, Okimoto et al. 2024], which we will use with the ultimate goal of discovering networks of relationships among these important genes. As part of our methodology, we proposed to stratify the original database according to the presence of estrogen and/or progesterone receptors, crucial elements for prognosis and personalized therapy. In this way, patients who share the same values for these markers were grouped together, resulting in four new databases where the n to p ratio becomes critical. To identify the best hyperparameter settings for FSG, we used the cross-validation method in the CV algorithm with 5 folds. This process allowed us to estimate the optimal values of the hyperparameters α_f and α_b, which serve as thresholds that enable the identification of independence relationships among the genes. The experiments were repeated over 100 iterations with different random seeds to obtain the average values with their confidence intervals. The application of the algorithm resulted in four graphs that highlight the conditional independence relationships among the genes involved in breast cancer (Figure 1). Our findings support the hypothesis that there are specific gene subnetworks that interact among breast cancer genes. The results may contribute to a deeper understanding of gene interactions in breast cancer, potentially offering new insights for future research and new therapeutic strategies.



Keywords: *Breast cancer, Gene networks, Conditional independence, Proteogenomic analysis, Fast Step Graph, Graphical Gaussian Models*

References

Polyak, K. (2011). Heterogeneity in breast cancer. *The Journal of Clinical Investigation*, 121(10):3786–3788.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160.

Liu, M. C., Pitcher, B. N., Mardis, E. R., Davies, S. R., Friedman, P. N., Snider, J. E., Vickery, T. L., Reed, J. P., DeSchryver, K., Singh, B., et al. (2016). Pam50 gene signatures and breast cancer prognosis with adjuvant anthracycline-and taxane-based chemotherapy: correlative analysis of c9741 (alliance). *NPJ breast* cancer, 2(1):1–8.

Okimoto, L. Y. S., Mendonca-Neto, R., Nakamura, F. G., Nakamura, E. F., Feny¨o, D., and Silva, C. T. (2024). Few-shot genes selection: subset of pam50 genes for breast cancer subtypes classification. *BMC Bioinformatics*, 25:92.

Liang, F. and Jia, B. (2023). Sparse Graphical Modeling for High Dimensional Data: A Paradigm of Conditional Independence Tests. *CRC Press*.

Zamar, R., Ruiz, M., Lafit, G., and Nogales, J. (2021). A stepwise approach for high-dimensional gaussian graphical models. *Journal of Data Science, Statistics, and Visualisation*, 1(2).

Colonna, J. G. (2023). FastStepGraph: First version of Fast Step Graph in R, https://cran.r-project.org/web/packages/FastStepGraph/readme/README.html, October.

Brazilian Symposium on Bioinformatics 2024

Enhancing Enzyme Generation with Fine-Tuned Conditional Transformers

Marco Nicolini[1], Emanuele Saitto[1], Rubén Jiménez[4], Emanuele Cavalleri[1], Marco Mesiti[1], Aldo Galeano[4], Dario Malchiodi[1], Alberto Paccanaro[4,5], Peter N. Robinson[2,3], Elena Casiraghi[1,2], and Giorgio Valentini[1,2]

1. AnacletoLab, Dipartimento di Informatica, Universita degli Studi di Milano, Italy
2. ELLIS – European Laboratory for Learning and Intelligent Systems
3. Berlin Institute of Health at Charité (BIH), Berlin, Germany
4. Escola de Matemática Aplicada, Fundação Getúlio Vargas, Rio de Janeiro, Brazil
5. Department of Computer Science, Centre for Systems and Synthetic Biology, Royal Holloway University of London, Egham, UK

We introduce *Finenzyme*, a Protein Language Model (PLM) that models specific Enzyme Commission (EC) categories by integrating transfer learning from a decoder-based Transformer, conditional learning using specific functional keywords, and fine-tuning. Using *Finenzyme*, we analyze how fine-tuning improves the prediction and generation of EC categories. Our findings reveal a two-fold reduction in perplexity for EC-specific categories compared to a general model, showing that fine-tuning helps capture specialized enzymatic functions that are not well represented in general models. We evaluated the generated enzymes using state-of-the-art tools such as ESMFold [1] for structure prediction and Foldseek [2] for structural similarity assessment. Despite low sequence identity, the generated proteins exhibit high structural resemblance to natural enzymes. Functional characterization using the CLEAN [3] tool confirms that the generated enzymes maintain the same EC functions as natural enzymes. Additionally, we demonstrate that the embedded representations of the generated enzymes closely resemble those of natural ones, making them suitable for downstream tasks such as enzyme classification and functional annotation. Clustering analysis reveals that the generated enzymes form clusters that largely overlap with those of natural enzymes, indicating that *Finenzyme* effectively captures the structural and functional properties of target enzymes. Finally, we showcase a practical application of *Finenzyme* in generating enzymes with specific functions using in-silico directed evolution – a computationally efficient fine-tuning methodology that significantly enhances and can assist targeted enzyme engineering tasks.

Keywords: *Large Language Models, Protein Language Models, Conditional Transformers, Enzyme design and modelling, Protein engineering*

References

1. Lin, Z., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637), 1123–1130.
2. Van Kempen, M., et al. (2024). Fast and accurate protein structure search with Foldseek. Nature Biotechnology, 42(2), 243–246.
3. Yu, T., et al. (2023). Enzyme function prediction using contrastive learning. Science, 379(6639), 1358–1363.

Brazilian Symposium on Bioinformatics 2024

Exploring Genetic Determinants of Post-COVID-19 Dyspnea: An Exome Sequencing Approach

Ventorim VP[1], Silva DRC[1], Basílio LA[1], Barbosa KRM[1], Morais LC[1], Morelatto SA[1], Meira DD[1], Louro ID[1].

1 Universidade Federal do Espírito Santo (UFES)

Objective: To investigate inflammation network genes and potential genetic variants of dyspnea progression after SARS-CoV-2 infection. **Methods**: 277 volunteers with different clinical spectra of COVID-19 were recruited between 2020 and 2023. The case-control study inclusion criteria were: 18 to 65 years, SARS-CoV-2+ confirmed and no convalescence in the last 30 days. After consent and questionnaire application, blood samples were collected for whole exome sequencing (CAAE: 37094020.6.0000.5060). Sixty genes of interest were pre-select for exome analysis. Variant selection was performed through the Partial Least Squares (PLS), which were then used as independent variables in logistic regression (LR) models to evaluate their association with dyspnea. For each significant genetic variant ($p < 0.05$), odds ratio (OR) and their respective 95% confidence intervals were calculated. **Results**: Participants were distributed in 171 (cases) and 106 (controls). From the case group, 44.76% reported progression of post-COVID-19 dyspnea. Exome sequencing provided 2038 variants, from which 30 variants were selected for LR, and 6 were significant with OR > 1. The variants found are: rs747917576 (gene: *CD8A*; OR: 5.89; 95% CI: 1.29 – 36.95; p-value: 0.03), rs3800788 (gene: *NOS3*; OR: 2.10; 95% CI: 1.15 – 3.91; p-value: 0.02), rs148125791 (gene: *PBX2*; OR: 5.60; 95% CI: 1.31 – 30.83; p-value: 0.03), rs6128 (gene: *SELP*; OR: 2.37; 95% CI: 1.31 – 4.38; p-value: <0.01), rs3753999 (gene: *TNNT2*; OR: 2.78; 95% CI: 1.30 – 6.13; <0,01) and rs1800380 (gene: *VWF*; OR: 2.44; 95% CI: 1.32 – 4.59; p-value: <0.01). **Discussion**: According to **Wada *et al*. (2018)** and **Ramachadran *et al*. (2021)**, intronic insertion variants, such as the rs747917576 variant and PBX2 variants, may be clinically relevant due to immunological control, which favor chronic inflammation. These two genes may be involved in the development of post-COVID-19 sequelae. Furthermore, synonymous variants, such as rs6128 and rs1800380, may be associated with viral resistance and cytokine release syndrome, linked to microvascular lesions and local inflammation **(Ruf *et al*., 2024; Wang *et al*., 2024)**. In this pathophysiology, shortness of breath in dyspnea may be due to an imbalance in vascular tone caused by microlesions in the endothelium. In the context of endothelial damage, vascular tone imbalance can be responsible for shortness of breath in dyspnea. Studies by **Tervi *et al*. (2024)** and **Voinescu *et al*. (2024)**, indicate that changes in NOS3 and TNNT2 are involved in heart failure, through vascular rigidity and heart hypertrophy lesions, respectively, generating hypoxia, cardiac output failure and possibly dyspnea. **Conclusion**: This research supports the hypothesis that genetic variants may be correlated with the progression of dyspnea. However, large cohort studies and connection disequilibrium analysis should be included to better predict post-COVID-19 dyspnea development.

Keywords: *SARS-CoV-2; Sequelae; Inflammation; Gene; Variant*

References

Ramachandran, D. *et al.* (2021). Association of genomic variants at PAX8 and PBX2 with cervical cancer risk. *International journal of cancer*, p. 893 - 900, March.

Ruf, W. (2024). Immune damage in Long Covid. *Science*, v. 383, n. 6680, p. 262–263, January.

Tervi, A. *et al.* (2024). Genetic and functional analysis of Raynaud's syndrome implicates loci in vasculature and immunity. *Cell Genomics*, p. 100630, August.

Voinescu, O. R. *et al.* (2024). Genotype-Phenotype Insights of Inherited Cardiomyopathies—A Review. *Medicina*, p. 543–543, March.

Wada, H. *et al*. (2018). Requirement for intron structures in activating the Cd8a locus. *Proceedings of the National Academy of Sciences of the United States of American,* p. 3440 - 3445, February.

Wang, C. *et al*. (2024). P-selectin Facilitates SARS-CoV-2 Spike 1 Subunit Attachment to Vesicular Endothelium and Platelets. *ACS infectious diseases*, p.2656 - 2667, June.

Gene Expression Patterns and Their Impact on Muscle pH in Four Swine Genetic Groups

Natanieli S. Máximo[1], Francelly G. Campos[1], Giarlã Cunha da Silva[2], Paula da Fonseca Pereira[1], Simone E. F. Guimarães[1]

[1]Departamento de Zootecnia – Universidade Federal de Viçosa (UFV)
Caixa Postal 36.570-900 – Viçosa – MG – Brazil

[2]Departamento de Microbiologia – Universidade Federal de Viçosa (UFV)
Caixa Postal 36.570-900 – Viçosa – MG – Brazil

natanieli.maximo@ufv.br, francellycampos@gmail.com,
giarla.silva@ufv.br, paula.fonseca@ufv.br, sfacioni@ufv.br

Abstract

Muscle pH significantly impacts pork quality and is influenced by factors related to its decline, glycolytic enzymes, and muscle fiber composition. Aiming to understand the mechanisms associated with energy metabolism and the conversion of muscle into meat, this study investigated the expression of genes involved in glycolytic metabolism and muscle regulation, as well as their correlation with muscle pH in different pig breeds. Twenty-four animals from four genetic groups were used: Piau (PP), Large White (LW), and the Piau-Large White (PL) and Duroc-Large White (DL) crosses. pH was measured at three intervals (0 h, 45 min and 24 h), followed by gene expression analysis in the *Longissimus dorsi.* Gene selection was based on the Pig QTL database for meat and carcass traits, with an emphasis on pH. Expression analysis was performed via qPCR, and fold change was used to determine gene expression variation relative to the control (PP). Results were processed using R software with the DESeq2 package for normalization and statistical analysis, and visualized in GraphPad Prism. It was observed that higher expression of the genes HK2, PFKM, and AMPKs in the commercial breeds (LW and DL) is associated with accelerated glycolytic metabolism, resulting in a rapid pH decline and lower final pH values. This phenomenon can be explained by use of energy for muscle growth and development, resulting in increased expression of MHC IIB and MHC IIX. In contrast, the PL cross demonstrated an additive genetic effect, preserving favorable characteristics, with a more gradual pH decline and a more oxidative metabolism, related to higher expression of genes involved in lipid oxidation (PPARGC1A), as well as the presence of oxidative fibers (MHC I). This is reflected in better final meat quality due to higher intramuscular fat content. These findings suggest that variations in gene expression influence metabolic traits and that the incorporation of local breeds can modulate pH levels, optimizing the final quality of pork.

Keywords*:* genomics, meat quality, metabolic stress, mRNA, skeletal muscle.

References

Albuquerque, A., Óvilo, C., Núñez, Y., Benítez, R., López-Garcia, A., García, F., ... and Martins, J. M. (2021). "Transcriptomic profiling of skeletal muscle reveals candidate genes influencing muscle growth and associated lipid composition in Portuguese local pig breeds", In: *Animals*, 11(5), 1423.

Lin, J., Wu, H., Tarr, P. T., Zhang, C. Y., Wu, Z., Boss, O., ... and Spiegelman, B. M. (2002). "Transcriptional co-activator PGC-1α drives the formation of slow-twitch muscle fibres", In: *Nature*, 418(6899), 797-801.

Listrat, A., Lebret, B., Louveau, I., Astruc, T., Bonnet, M., Lefaucheur, L., ... and Bugeon, J. (2016). "How muscle structure and composition influence meat and flesh quality", In: *The Scientific World Journal*, *2016*(1), 3182746.

Matarneh, S. K., Scheffler, T. L., and Gerrard, D. E. (2023). "The conversion of muscle to meat", In: *Lawrie's meat science* (pp. 159-194), Woodhead Publishing.

Ren, C., Li, X., Bai, Y., Schroyen, M., and Zhang, D. (2022). "Phosphorylation and acetylation of glycolytic enzymes cooperatively regulate their activity and lamb meat quality", In: *Food Chemistry*, 397, 133739.

Ryu, Y. C., and Kim, B. C. (2006). "Comparison of histochemical characteristics in various pork groups categorized by postmortem metabolic rate and pork quality", In: *Journal of animal science*, 84(4), 894-901.

Tan, X., He, Y., He, Y., Yan, Z., Chen, J., Zhao, R., ... and Li, B. (2023). "Comparative proteomic analysis of glycolytic and oxidative muscle in pigs", In: *Genes*, 14(2), 361.

Yao, C., Pang, D., Lu, C., Xu, A., Huang, P., Ouyang, H., and Yu, H. (2019). "Data mining and validation of AMPK pathway as a novel candidate role affecting intramuscular fat content in pigs", In: *Animals*, 9(4), 137.

Zhang, J., Wang, J., Ma, C., Wang, W., Wang, H., and Jiang, Y. (2022). "Comparative transcriptomic analysis of mrnas, mirnas and lncrnas in the longissimus dorsi muscles between fat-type and lean-type pigs", In: *Biomolecules*, 12(9), 1294.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Genome mining unveils the *Algibacter* genus as a treasure trove of biologically-active compounds

Erica de Souza Monteiro[1], Eduardo Costalonga Alves[1], Bruno Francesco Rodrigues de Oliveira[1]

1. Laboratory of Marine Bacteriomes, Department of Microbiology and Parasitology, Biomedical Institute, Fluminense Federal University, Niterói, Brazil.

The *Algibacter* genus, member of the family *Flavobacteriaceae*, encompasses rod-shaped and Gram-negative bacteria, characterized by its facultative anaerobic metabolism, gliding motility, production of non-diffusible orange pigments and agarolytic activity. Most importantly, all representatives in the fifteen validly-accepted species in this genus have been isolated and/or detected from marine environments, with the first isolates recovered from green algae. To date, *Algibacter* has been mostly explored as a source of novel enzymes, especially carbohydrate-active ones, due to its intrinsic role in algae degradation. However, few studies have investigated its potential for other biotechnological uses, such as antibiotic and antineoplastic activities, which is counterintuitive considering that macroalgae-associated microorganisms have been more extensively harnessed as promising producers of antimicrobial and cytotoxic substances. In the last decade, genome mining has emerged as a valuable approach to the discovery of novel drugs with wide-ranging therapeutic applications against infectious and chronic diseases. In this framework, this study aimed to apply genome mining to assess the genetic repertoire associated with the production of bioactive metabolites in currently-available *Algibacter* genomes. First, a total of 68 genome sequences (.fasta assembly files) were recovered from the GenBank (NCBI) and Genome Taxonomy Database (GTDB). Then, they were submitted to antiSMASH 7.0 for the identification of secondary metabolite biosynthetic gene clusters (BGCs) and BAGEL4 for the specific detection of gene clusters for bacteriocins and ribosomally synthesized and post-translationally modified peptides (RiPPs). BGCs for fifteen different classes of secondary metabolites were predicted with antiSMASH 7.0, with terpene being the most abundant class, found in 60 (88.8%) genomes, distantly followed by non-ribosomal peptides, present in 23 (33.82%) genomes. Seven strains (10.3%) had no BGCs found. Following BAGEL4 analyses, sactipeptides was the top RiPPs class, for which gene clusters were present in 29 strains (42.64%); however, no results were found for 36 genomes with this genome mining tool. Both antiSMASH 7.0 and BAGEL4 particularly detected the same BGCs encoding for class I lanthipeptides in metagenome-assembled genomes (MAGs) of kelp-derived *Algibacter* sp. representatives and lassopeptides in two *Algibacter mikhailovii* strains, compounds for which antimicrobial activity against Gram-positive bacteria and cytotoxic properties have already been reported. Thus, our preliminary assessment reveals that this marine bacterial genus potentially represents a paramount reservoir of hydrocarbon- and peptide-derived natural products with antimicrobial and anticancer activities deserving further attention in future bioprospecting endeavours.

Keywords: blue biotechnology, marine bacteria, microbial genomics, new drugs.

Brazilian Symposium on Bioinformatics 2024

Group I introns in the mitochondrial genomes of *Trichoderma* spp.

Xavier, L. F. S.[1], Silva, M. L.[1], M., Vidigal, P. M.[2], Queiroz, M. V.[1]

1. Departamento de Microbiologia, Instituto de Biotecnologia Aplicada à Agropecuária (BIOAGRO), Laboratório de Genética Molecular de Microrganismos (LGMM), Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brasil
2. Núcleo de Análise de Biomoléculas (NuBioMol), Centro de Ciências Biológicas, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brasil

Trichoderma is a globally distributed genus of filamentous fungi. The mitochondrial genome of *Trichoderma* spp. ranges from 27 to 94 kb. The chromosome is circular (GC ~27%) and contains 14 conserved genes involved in respiratory metabolism. Furthermore, the presence of introns is ubiquitous in these genomes and directly affects genomic plasticity. Organelle introns are autocatalytic sequences that can be divided into two main groups: I and II. Group I introns are common in mitochondrial genomes, while group II introns are common retroelements in plastid genomes [1, 2]. Group I introns can encode homing endonucleases, and group II introns can encode reverse transcriptases [3]. This work aims to describe the group I introns found in the mitochondrial genome of *T. brasiliensis*, a recently characterized endophytic fungus from the Amazon rubber tree *Hevea brasiliensis* [4] and compare them with group I introns present in 16 mitochondrial genomes from other *Trichoderma* species. The genomic DNA of *T. brasiliensis* (strain T17F5R-AM) was extracted and sequenced. After quality analysis using FastQC software, the genome was assembled using SPAdes software and circularized using CLC Main Workbench software. The annotation of mitochondrial genes was done using MITOS2, MFAnnot, and RNAweasel (available online), configured with genetic code 4 (fungi, protozoa, and cnidarians). The reference genomes were obtained from GenBank and re-annotated to standardize the analyses. In total, 85 introns were identified in the 17 mitochondrial genomes analyzed. The *T. brasiliensis* mitogenome has 4 introns: 1 intron from subgroup IA (2,657 bp) in the *rrnL* gene and 3 introns from subgroup IB (1,271, 1,130, and 977 bp) in the *cox1*, *cox2*, and *cob* genes, respectively. This number is below the average observed in the mitogenomes of other *Trichoderma* species. In *T. brasiliensis*, the largest intron (subgroup IA/*rrnL* gene) is 303 bp shorter than the largest intron, which was observed in *T. cornu-damae* KA19-0412C (subgroup IB/*cob* gene) of 2,960 bp. In contrast, the smallest intron of *T. brasiliensis* (subgroup IB/*cob* gene) is 3 times larger than the smallest intron (274 bp), which was observed in *T. koningiopsis* POS7 (subgroup IB/*cox3* gene). Forty-three introns carry homing endonuclease sequences (LAGLIDADG and GIY-YIG), 3 of which are in *T. brasiliensis*. In *T. brasiliensis*, introns represent 17.45% of the genome size, but this value can reach 40% in *T. cornu-damae* KA19-0412C, for example. The results show that in *Trichoderma*, the size of mitochondrial genomes is closely related to the presence of group I introns. Although they are not essential to the functioning of mitochondria, introns can participate in both the regulation of gene expression and the horizontal acquisition of new genes, which may ultimately represent an evolutionary advantage to the host.

Keywords: *Trichoderma*, mitochondria, splicing, homing endonuclease

References

1. Lambowitz, A. M. and Zimmerly, S. (2004). Mobile group II introns. *Annual Review of Genetics*, 38, pp. 1–35.

2. Lang, B. F., Laforest, M. J. and Burger, G. (2007). Mitochondrial introns: a critical view. *Trends in Genetics*, 23(3), pp. 119–125.

3. Stoddard, B. L. (2014). Homing endonucleases from mobile group I introns: discovery to genome engineering. *Mobile DNA*, 5, pp. 1–16.

4. Brito, V. N., Alves, J. L., Araújo, K. S., Leite, T. S., de Queiroz, C. B., Pereira, O. L. and de Queiroz, M. V. (2023). Endophytic *Trichoderma* species from rubber trees native to the Brazilian Amazon, including four new species. *Frontiers in Microbiology*, 14, p. 1095199.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Homology-based quantification of the fibrosis in the CT image: A proof of concept for CT image feature-assisted gene expression prediction

Kentaro Doi[1, 2], Hodaka Numasaki[2], and Yayoi Natsume-Kitatani[1,3,4]
1. National Institutes of Biomedical Innovation, Health and Nutrition
2. Graduate School of Medicine, Osaka University
3. Institute of Advanced Medical Sciences, Tokushima University
4. Institute for Protein Research, Osaka University

[Purpose]
This study aimed to establish a proof of concept (POC) for developing the computed tomography (CT) image feature-assisted gene expression predicting.

[Materials and Methods]
We collected 100 CT images of the patients with COVID-19 and lung cancer from the cancer image archive [1, 2]. Homology-based features are b0, b1, and b1/b0, indicating the number of isolated components (black pixels), holes (white pixels), and the ratio of the holes-isolated components in a binary image. These features represent the connectivity among objects in the image. This estimation was performed by changing the threshold value applied to the image binarization using the homology-profile method (HP) [3]. The calculation region of interest was set to 32×32 matrix and shifted by 8 pixels on the CT image of the 512×512 matrix. The correspondence region in the resulting image assigned the maximum value of each HP as a homology-based feature (HF) map. In this study, the effectiveness of the HF map in capturing the promising feature of fibrosis was evaluated by comparing the cases of COVID-19 and lung cancer.

[Results]
The b1 HF map showed the largest visual association with the fibrosis lesions, with the HF value of fibrosis lesions enhanced as higher than that of the lung field of a lung cancer patient. Here, the b1 HF value on the fibrosis lesion indicated more than 1500 in the lung field, although the value showed a lower value than 500 in the case of lung cancer.

[Conclusion]
We validated the effectiveness of the homology-based feature in the CT image to capture the promising feature as a POC for the CT image feature-assisted gene expression prediction. This study demonstrated that the HF value can quantitatively indicate the fibrosis. We expect that the HF value may be promising in predicting gene expression prediction.

Keywords: *radiomics, image analysis, fibrosis, CT image*

References

1. An, P., Xu, S., Harmon, S. A., et al. (2020). CT Images in COVID-19 [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/TCIA.2020.GQRY-NC81
2. Aerts, HJWL., Rios, VE., Leijenaar RTH., et al. (2015). Data From NSCLC-Radiomics-Genomics. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2015.L4FRET6Z
3. Nakane, K., Takiyama, A., Mori, S., et al. (2015). Homology-based method for detecting regions of interest in colonic digital images. Diagnostic pathology 10: 1-5.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

In silico validation of Linear B-Cell epitopes using Machine Learning: A proposed approach with organisms of genus *Trypanosoma*

Bruna Caroline Russi[1], Renato Simões Moreira[2], Pablo Daniel Cuña Cabrera[3], Sílvio César Cazella[4]

1. Universidade Federal de Santa Catarina (UFSC), Brasil
2. Instituto Federal de Santa Catarina (IFSC), Brasil
3. Universidad Tecnológica del Uruguay (UTEC), Uruguay
4. Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Brasil

Introduction: Epitopes or antigenic determinants are parts of a pathogen or a foreign protein to the organism that can be recognized by B or T cells, playing a crucial role in the immune response [Abbas et al. 2021]. *In silico* approaches to predict epitopes using machine learning can help peptide-based vaccine research, as it allows the development of strategies that specifically target these entities – boosting the effectiveness of immune interventions [Bravi 2024]. However, identifying relevant epitopes among the predicted ones is still a challenge due to the need to consider biochemical characteristics. Improving this approach could be strategic for the study of Neglected Tropical Diseases (NTDs), such as Chagas disease (American trypanosomiasis), caused by *Trypanosoma cruzi*, which affects approximately 30,000 people annually in Brazil and is responsible for about 14,000 deaths per year [Ministério da Saúde 2022]. **Objective:** To develop a predictive model to validate predicted linear B-cell epitopes in organisms of the genus *Trypanosoma* using biochemical features. **Methodology:** The adopted methodology included creating a dataset using data collected from the IEDB [Vita et al. 2018] and UniProtKB [Magrane and Consortium 2011] databases. Initially, epitopes of *Trypanosoma* genus from the IEDB were selected, serving as a reference due to their validation in previous studies. Then, synthetic peptides were generated from data available in UniProtKB, complementing the database. Data preprocessing involved filtering only linear epitopes from IEDB and integrating valid and synthetic epitopes, incorporating scores and biochemical features obtained using the BepiPred-3.0 [Clifford et al. 2022] and EpiBuilder-1.0 [Moreira et al. 2022] tools. After preparing the dataset, experiments were conducted using the Orange Data Mining [Demsar et al. 2013] tool, where different machine learning algorithms were tested. The Decision Tree algorithm was selected as the most effective for the task. **Results:** The obtained results showed that the predictive model achieved an accuracy of 0.933, with an F1-Score also of 0.933, indicating a high ratio of true positives relative to the total predicted positives. The analysis of feature importance revealed that the Emini [Emini et al. 1985] feature was the most relevant, contributing 68.79% to the model's decisions. **Conclusion:** The model effectively validated the predicted epitopes, suggesting a contribution to *in silico* validation of epitopes. The developed approach has the potential to assist in studies involving new treatments for *Trypanosoma*-caused diseases, but could be adapted and generalized to other pathogens.

Keywords: *Immunoinformatics, Machine learning; Decision tree; Linear epitopes; Trypanosoma.*

References

Abbas, A. K., Lichtman, A. H., and Pillai, S. (2021). Imunologia Básica - Funções e Distúrbios do Sistema Imunológico. GEN Guanabara Koogan, 6ª edição.

Bravi, B. (2024). Development and use of machine learning algorithms in vaccine target selection. npj Vaccines, 9(1):15.

Clifford, J. N., Høie, M. H., Deleuran, S., Peters, B., Nielsen, M., and Marcatili, P. (2022). BepiPred-3.0: Improved B-cell epitope prediction using protein language models. Protein Science, 31(12):e4497.

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M., and Zupan, B. (2013). Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research, 14:2349–2353.

Emini, E. A., Hughes, J. V., Perlow, D. S., and Boger, J. (1985). Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. Journal of Virology, 55(3):836–839.

Magrane, M. and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. Database, 2011(0):bar009–bar009.

Ministério da Saúde (2022). Pacto Nacional para a Eliminação da Transmissão Vertical de HIV, Sífilis, Hepatite B e Doença de Chagas como Problema de Saúde Pública. Brasília, DF.

Moreira, R. S., Filho, V. B., Calomeno, N. A., Wagner, G., and Miletti, L. C. (2022). EpiBuilder: A Tool for Assembling, Searching, and Classifying B-Cell Epitopes. Bioinformatics and Biology Insights, 16:117793222210952.

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., and Peters, B. (2018). The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res, 46(D1):D339–D343.

Brazilian Symposium on Bioinformatics 2024

Metabarcoding reveal *Fusarium decemcellulare* as the potential causal agent of the emergent disease in *Coffea canephora*

Miquéias Fernandes[1], Iana Pedro da Silva Quadros[1], Leandro Fonseca de Souza[1], Inorbert de Melo Lima[2], José Aires Ventura[2], Marcia Flores da Silva Ferreira[1]

[1]Departamento de Agronomia - Universidade Federal do Espírito Santo, Brazil.
[2]Incaper - Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural, Brazil.

Coffea canephora is produced in almost all of Espírito Santo state in Brazil and corresponds to 70% of the country's production. Since 2014, great damage to coffee plantations has been recorded by the “coffee tree branch canker”. The pathosystem of this disease is complex and the conclusion of what is the causal agent of the disease has not yet been reached, with previous research indicating *Fusarium* spp. as the etiologic cause. To elucidate the plant-pathogen interaction of this system, genotypes of conilon coffee plants in the state of Espírito Santo were evaluated by metabarcoding with primers for ITS region (ITS1F-ITS2). The Sequencing was performed in Illumina NovaSeq 6000 using DNA from pooled leaves and stem of 5 plants as replicates, 10 cm up/down the damaged tissue, of the clones: Susceptible: K61 and MP3 (asymptomatic and symptomatic samples) and Resistant A1 and P2 (Asymptomatic). The sequences were processed with softwares Cutadapt 4.6, Trimmomatic 0.39, BBMap 38.32 and Qime2 2024.2, with ITSxpress and Jupyterlab for bioinformatics analyses. The BBSplit tool of BBMap suite was used to decontamination and the *C. canephora* assembly NC_030053.1 was used to mapping sequences on chloroplast and nuclear genome. The 59 samples have 0.1 million sequences in average, with mean read length of 218 bases. After quality control, 99,8% of the reads lasted (total of 18,6 million reads). The rarefaction analyses showed sufficient sequencing coverage to represent fungal diversity, with at least 70.000 reads per samples. The reads classification was performed on qiime2 using Unite10 database, for singletons clustered sequences at minimum of 99% identity. The Differential abundance tested with ANCOM-BC analysis on level 7 (species classification) showed enrichment of the fungal plant pathogen *Albonectria rigidiuscula* in stem tissue of symptomatic plants of the two genotypes tested (K61 and MP3) compared to asymptomatic plants of the same genotype and to the resistant genotypes P2 and A1. No fungal species were detected as differential in leaves of symptomatic plants, compared to asymptomatic or resistant genotypes. The anamorph of *A. rigidiuscula,* the fungus *Fusarium decemcellulare,* is associated with inflorescence wilt and vascular necrosis in fruit tree crops such as Mango, Longan and Rambutan. These results highlight *Fusarium decemcellulare* as the potential causal agent of coffee tree branch canker, open venues to phytopathology trials for disease control strategies. Next steps are the analyses of 16S rRNA e 18S rRNA metataxonomy to evaluate if there are potential Bacteria or Protozoa related to the disease.

Keywords: microbial ecology, *metabarcoding,* phytopatology, ecogenomics.

Acknowledgment: FAPES, CAPES, CNPq, Embrapa Café e Incaper.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Molecules of the Amazon: Integration and Centralization of Data on the Amazon Flora and its Biomolecules

Carolina Barros da Costa [1], Gustavo Casagrande Borges [1], Marcos Reis Dutra [1], Márcio Rodrigues Miranda [1], Kaio Alexandre da Silva [1]

1. Federal Institute of Education, Science and Technology of Rondônia- IFRO

The Brazilian flora is considered one of the richest and most diverse in the world. According to the Flora and Fungi of Brazil project, a total of 52,772 species were recorded as of October 2024, including native, naturalized, or cultivated species. The Amazon stands out as the Brazilian region richest in biological diversity. Its flora is composed of species that represent a vast source of bioactives with biotechnological potential, arousing the interest of industries such as pharmaceuticals, cosmetics, and food. However, there is no database that centralizes information on the Amazon flora that includes species, biomolecules, and digital structures, regardless of file formats, structural level, or visualization format, as these are dispersed across different databases. This fragmentation limits the potential for compound prospecting, as researchers need to consult multiple sources to obtain information, which not only wastes a lot of time in this activity, but also results in greater financial and natural resource expenditure by performing bioprospecting experimentally. Thus, the creation of a centralized database allows researchers to quickly access complete data, contributing to reducing the time and costs involved in the research process. Thus, the objective of this study is to present the development of "Molecules of the Amazon", an online repository that centralizes data on Amazonian flora species and their biomolecules. To this end, it was necessary to develop a web scraping algorithm that was capable of collecting and classifying data from a set of databases for feeding or repository: GBIF, for collecting data on species, NuBBEDB, NCBI and ChEMBL, for collecting molecular, chemical, bibliographic and patent data. As preliminary results, Molecules of the Amazon has already integrated data on Amazonian flora species and biomolecules, crossing taxonomic and geographic data of Amazonian species with molecular, chemical data, and bibliographic and patent information on their biomolecules. Its main distinguishing feature is its ability to provide data on the geographic distribution of species and associated biomolecules. The repository offers filters that allow you to locate molecules using their common name, IUPAC, Inchikey, molecular formula or physical-chemical characteristics. In addition, it is possible to refine searches by species, with filters for taxonomy, life form, origin and geographic location, or refine by bibliographic or patent reference data. When searching, the information is presented clearly and concisely through intuitive layouts. When selecting the molecule of interest, the user is directed to a page with more detailed data. In addition, search data is stored in the search history. Thus, by centralizing previously dispersed data, Molecules of the Amazon results in a robust platform that can directly impact the advancement of bioinformatics applied to biodiversity and biotechnology, and can not only facilitate scientific research, but also assist in the discovery of new biotechnological applications.

Keywords: biodiversity, bioactives, database, biotechnology.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Network Pharmacology and UHPLC-ESI-Q-TOF-MS/MS Approaches to Explore Active Compounds and Mechanisms of Acerola Seed Hydroethanolic Extract in Obesity Treatment

Beatriz Paes Silva[1], Gabriel Arcanjo Viana Neto[1], Gustavo Henrique de Sousa[1], Vinicius Franco de Oliveira[1], Thales Yugo Kitahara[1], Livia Bracht[1], Jurandir Fernando Comar[1], Rosane Peralta[1], Adelar Bracht[1], Anacharis Babeto de Sá-Nakanishi[1]

1 Laboratory of liver metabolism, Department of Biochemistry, State University of Maringá, Brazil.
*Corresponding author: Beatriz Paes Silva
pg55746@uem.br, ORCID: https://orcid.org/0000-0003-1242-7856

Obesity is a disease with rising global prevalence and is linked to various health-related conditions including, metabolic syndrome and type 2 diabetes. Recent studies suggest that acerola byproducts, particularly seed extract, exhibit potential anti-diabetic properties. Brazil is the largest producer of acerola and disposes tons of byproducts from its juice industry. This work aimed to identify phytochemicals present in acerola seed extract (AS) that could contribute to managing obesity and its related conditions using network pharmacology. Chemical characterization of AS was performed through liquid chromatography-mass spectrometry (UHPLC-ESI-Q-TOF-MS/MS). Oral bioavailability, drug-likeness (Lipinski's Rule of Five), gastrointestinal absorption, Caco-2 permeability, and blood-brain barrier (BBB) penetration were evaluated using the SwissADMET Program. Obesity-related targets were obtained from the GeneCards database, while targets related to AS extract were sourced from SuperPred. Protein-protein interaction (PPI) analysis was conducted using the STRING platform and visualized with Cytoscape software. Additionally, GO and KEGG enrichment analyses were performed to elucidate the molecular mechanisms of AS's core genes. Forty-eight molecules were identified in AS, and thirty-six showed drug-likeness and gastrointestinal absorption potential. The active ingredient-target network revealed 171 nodes and 1106 interactions. Key targets in the PPI network included SRC, EGFR, ESR1, NFKB1, PTGS2, MMP9, TLR4, PPARA, STAT1, and MAPK1. KEGG pathway analysis indicated that AS action on obesity involves primarily key signaling pathways related to inflammation and insulin resistance, namely PI3K-Akt, HIF-1, mTOR, FoxO, Toll-like receptor, NF-kappa B, and calcium signaling pathways. Then, it was found that the AS is associated mainly with cell proliferation, responses to external stimuli, and regulation of oxidative stress. About cellular components, membrane microdomains like rafts and caveolae were identified as crucial elements in the aggregation of signaling proteins, facilitating inflammatory responses and metabolic regulation. Obesity and insulin resistance are linked to low-grade chronic inflammation. Proteins such as NFK-B, PTGS2, and TLR4 play in the inflammatory response, while ESR1, EGFR, and SRC are involved in oxidative stress and inflammation modulation. Membrane components like rafts and caveolae provide essential platforms for aggregation signaling proteins such as SRC, EGFR, and TLR4, thus facilitating responses to external stimuli and metabolic regulation. In conclusion, acerola seed extract can manage obesity by modulating chronic inflammatory processes and oxidative stress. Network pharmacology revealed the multi-target action of these compounds, suggesting a comprehensive mechanism for obesity treatment.
Keywords: Inflammation; Oxidative stress; Insulin resistance; Phytochemicals; Signaling pathways

References

1. Liu, Y., Luo, J., and Xu, B. (2024). Elucidation of Anti-Obesity Mechanisms of Phenolics in *Artemisiae argyi Folium (Aiye)* by Integrating LC-MS, Network Pharmacology, and Molecular Docking. Life (Basel), 14(6):656. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11205026/, May.

2. Xie, P., Guo, M., Xie, J.B., Xiao, M.Y., Qi, Y.S., Duan, Y., Li, F.F., and Piao, X.L. (2022). Effects of Heat-Processed *Gynostemma pentaphyllum* on High-Fat Diet-Fed Mice of Obesity and Functional Analysis on Network Pharmacology and Molecular Docking Strategy. *Journal of Ethnopharmacology*, 294:115335. https://www.sciencedirect.com/science/article/pii/S0378874122003749?via%3Dihub, August.

3. Wu, Z., Yu, W., Ni, W., Teng, C., Ye, W., Yu, C., and Zeng, Y. (2023). Improvement of Obesity by Liupao Tea is through the IRS-1/PI3K/AKT/GLUT4 Signaling Pathway According to Network Pharmacology and Experimental Verification. *Phytomedicine*, 110:154633. https://www.sciencedirect.com/science/article/pii/S0944711322007206?via%3Dihub, February.

4. Wang, Y., Yang, S.H., Zhong, K., Jiang, T., Zhang, M., Kwan, H.Y., and Su, T. (2020). Network Pharmacology-Based Strategy for the Investigation of the Anti-Obesity Effects of an Ethanolic Extract of *Zanthoxylum bungeanum Maxim*. *Frontiers in Pharmacology*, 11:572387. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7751641/, November.

Brazilian Symposium on Bioinformatics 2024

Nuclear Segmentation of Oncology Microscopy Images through Convolutional Neural Networks: A Comparative Analysis

Antônio Vithor Prinz Moraes[1], Felipe dos Santos Passarela[1], Matheus Correia Casotti[1], Daniel Cruz Cavalieri[2], Débora Dummer Meira[1], Iúri Drumond Louro[1]

1. Systems and Computational Biology Group (SCBG), Center for Human and Molecular Genetics (NGHM), Federal University of Espírito Santo (UFES)
2. Instituto Federal do Espírito Santo - Campus Serra (IFES Serra)

Nuclear segmentation is a fundamental step in cancer diagnostics and classification as it directly impacts the precision of morphological analysis, thereby improving diagnostic accuracy, objectivity, and reproducibility. This study presents a comparison between two CNN architectures, U-Net and a modified DeepLab V3+—for segmenting nuclei in microscopic images from the 2018 Data Science Bowl dataset, encompassing both brightfield and fluorescence microscopy. The study (source code: https://github.com/FelipePassarela/cancer-cell-segmentation) was conducted in the PyTorch library using an NVIDIA RTX 4060 Ti GPU (8GB VRAM) and 16GB of RAM. The hyperparameters included 50 training epochs, a learning rate of 3x10^-4, a batch size of 8, and image resizing to 256x256 pixels. A ReduceLROnPlateau scheduler, patience set to 5, dynamically adjusted the learning rate, while the AdamW optimizer was used with default configurations. A hybrid loss function, combining Binary Cross Entropy and Soft Dice, balanced learning across edges and interiors of the cell nuclei. To boost training robustness, we incorporated data augmentation (Gaussian blur, adaptive autocontrast, affine transformations) and batch normalization followed by GELU activation in each convolutional layer. Data augmentation includes: Gaussian blur (3x3 kernel, sigma between 0.1 and 2.0), adaptive autocontrast, and affine transformations (rotation of ±5°, translation of ±10%, and scaling between 0.9-1.1). Validation and inference stages applied normalization and resizing only. The DeepLab V3+ used a simplified backbone with only three convolutional layers, resulting in 3,897,249 trainable parameters, approximately 12% the size of the implemented U-Net (31,043,521 parameters). The models were evaluated based on three metrics: Dice coefficient and Hausdorff distance. U-Net showed slightly better performance, with Dice coefficients of 0.9187 (training), 0.9015 (validation), and 0.9131 (test). The modified DeepLab V3+ achieved values of 0.9044 (training), 0.8839 (validation), and 0.8921 (test). Meanwhile, the Hausdorff distances were comparable between the models: U-Net with 21.7319 (training), 19.5527 (validation), and 20.1712 (test), and DeepLab V3+ with 21.7826 (training), 19.5527 (validation), and 20.1712 (test). The high efficacy observed (approximately 90%) demonstrates the viability of CNNs in nuclear segmentation, promoting automation and accuracy in cancer diagnosis, minimizing human intervention, and providing traceable information. In summary, the simplified implementation of DeepLab V3+ achieved comparable metrics to traditional U-Net, even with a significant reduction in the number of parameters. This optimization suggests potential applications, especially due to the high demand for robust and adaptable segmentation methods, driven by the need to analyze large volumes of data, aiming to provide significant innovations for diagnostic and prognostic practices in Oncology, by accelerating workflows, reducing diagnostic errors, and adapting therapies. Thus, delivering speed, precision, and safety in oncology, even with minimal computational resources.

Keywords: Digital Pathology. Oncology. Computer-Assisted Image Processing. Convolutional Neural Networks. Deep Learning.

References

Chen, L. C. et al. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).

Goodman, A. et al. (2018). 2018 Data Science Bowl. https://kaggle.com/competitions/data-science-bowl-2018, 2018. Kaggle.

Hayakawa, T. et al. (2021). Computational nuclei segmentation methods in digital pathology: a survey. Archives of Computational Methods in Engineering, 28, 1-13.

Irshad, H. et al. (2013). Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential. IEEE reviews in biomedical engineering, 7, 97-114.

Ronneberger, O., Fischer, P. and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18 (pp. 234-241). Springer International Publishing.

Serag, A. et al. (2019). Translational AI and deep learning in diagnostic pathology. Frontiers in medicine, 6, 185.

Xing, F. and Yang, L. (2016). Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. IEEE reviews in biomedical engineering, 9, 234-263.

Brazilian Symposium on Bioinformatics 2024

Predicting side effects of drug combinations in realistic experimental settings

Rubén Jiménez[1] and Alberto Paccanaro[1,2]

1. Escola de Matemática Aplicada, Fundação Getúlio Vargas, Rio de Janeiro, Brazil
2. Department of Computer Science, Centre for Systems and Synthetic Biology, Royal Holloway University of London, Egham, UK

Drug-Drug Interactions (DDIs) present significant challenges in healthcare due to their potential to cause harmful side effects [1]. Knowledge of these complex interactions is limited because they are not often detected during clinical testing [2]. Here, we introduce DCSE (Drug Combinations Side Effects), a novel machine learning method for predicting side effects caused by pairs of drugs. DCSE learns signatures (latent feature vectors) which encode the biological interplay between drug pairs and side effects. Signatures are learned separately for each drug and then combined nonlinearly using a deep neural network to obtain the signature for the drug pair. Side effect predictions are obtained through the linear combination of drug pair and side effect signatures, and it can be interpreted as the probability of the side effect being associated with the drug pair. We first evaluated DCSE's performance in the experimental settings that are most commonly adopted in the literature. Next, we introduced more realistic settings in which experiments were separated into two distinct categories: warm-start, where the drug pairs in testing are the same as in training but some of their side effect associations are unseen, and we aim to predict these side effects for the pair; and cold-start, where the pairs of drugs in testing are unseen during training, and we predict the potential side effect of a new combination of drugs. Finally, we performed a prospective evaluation where we predicted associations reported between 2009 and 2014 that were not listed in the 2009 snapshot that we used for training. Our extensive evaluation shows that DCSE consistently outperforms state-of-the-art methods, both in the traditional settings as well as in our novel settings, demonstrating its robustness and efficacy in real-world applications.

Keywords: *machine learning, deep neural networks, representation learning, polypharmacy side effects, drug-drug interactions*

References

1. Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 34(13), i457–i466. doi:10.1093/bioinformatics/bty294.
2. Bansal, M., Yang, J., Karan, C., Menden, M.P., Costello, J.C., and Tang, H. (2014). A community computational challenge to predict the activity of pairs of compounds. Nature Biotechnology, 32(12), 1213–1222.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Sialic Acid Enzymes Database

Moura, T. A. F.[1], Zaramela, L. S.[1]

1. Ribeirão Preto Medical School (University of São Paulo)

Sialic acids are a large family of nine-carbon carboxylated monosaccharides present in animals' and microorganisms' membrane glycoconjugates. They are located at the outermost end of *N*-linked and *O*-linked carbohydrate chains of membrane glycoproteins and glycolipids, and this characteristic is reflected in their functions, as they act as ligands or receptors for cell-cell communication and host-parasite interaction. Several commensal and pathogenic bacteria have the ability to metabolize, synthesize, and incorporate sialic acids through catabolism, anabolism, and assimilation pathways. Bacteria capacity to synthesize and assimilate sialic acids enables them to evade the innate immune system from hosts through the synthesis of sialylated capsules and sialic acid incorporation onto their membrane glycoconjugates. Also, the ability of bacteria to catabolize sialic acids allows them to use it as a carbon and nitrogen source, conferring an advantage in competitive growth. To date, it is still unknown how prevalent the presence of sialic acid metabolic pathways is in microbial communities. Since sialic acid metabolism is associated with pathogenicity, this knowledge will greatly contribute to the future development of alternative medical interventions. In order to broaden our knowledge in this field, we aim to characterize all bacterial genomes available in public databases. We will perform a comparative genomic analysis to evaluate the presence and absence of sialic acid metabolic pathways in these microbes. In addition, we will determine their origins (environmental, host-associated) and their potential interactions with their hosts. A challenging step in this analysis is the annotation of the genomes. The protein sequences available in databases are not well annotated, not curated, and their indiscriminate use can lead to error propagation. A suitable approach is to use curated databases rather than general ones[1]. Therefore, in order to achieve an accurate genomic annotation, a local database is being established. This database will include data from RefSeq, Uniprot, dbCAN-PUL, Cazzy, KEGG, MetaCyc, and BV-BRC. In addition, several quality control steps are going to be performed to ensure the reliability of the sequences. So far, 853838 sequences have been collected from RefSeq, Uniprot and dbCAN-PUL, 74.2% of which are catabolic enzymes, 3.5% anabolic enzymes, 13.2% transport proteins, 8.2% sialidases, and 0,8% transcriptional factors. These numbers bring insight into the relative availability of information relating to sialic acid metabolism in the databases. We are currently working on improving data selection to establish a gold standard for our local database.

Keywords: *Sialic Acid, Database, Genomics, Metagenomics.*

References

GHOSH, Shyamasree (2020). Sialic acid and biology of life: An introduction. Sialic acids and sialoglycoconjugates in the biology of life, health and disease, p. 1.

Del Angel, V. D., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Pettersson, O. V., ... & Lantz, H. (2018). Ten steps to get started in Genome Assembly and Annotation. F1000Research, 7.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

The potential role of the JAK/STAT pathways in the progression of depressive and anxiety disorders in Long COVID

Flávio dos Santos Alvarenga[1], Bianca Paulino Campanharo[1], Isabele Pagani Pavan[1], Matheus Correia Casotti[1], Débora Dummer Meira[1], Iuri Drumond Louro[1]

1. Universidade Federal do Espírito Santo (UFES)

After the end of the pandemic, several consequences from SARS-CoV-2 infection have been reported. These are persistent multisystemic sequelaes, a condition referred to as Long COVID (LC). Among them are depression and anxiety, possibly associated with the cytokine storm during the acute phase, which, over time, influences the persistence and progression of mental disorders. Therefore, to elucidate the molecular mechanisms involved, a holistic analysis of genes and proteins potentially involved in depression and anxiety in LC was conducted, aiming to support future *in vivo* and *in vitro* studies. With this purpose, a manual curatorship was performed following this search strategy: (Depression OR Anxiety) AND (Long COVID OR Post COVID) AND (Protein OR Gene). The search yielded 48 articles, published between 2020 and 2023, containing the searched descriptors in the title and/or abstract. Through this process, 50 proteins were identified, each cited at least twice in the articles, and found to be related to mental disorders. Subsequently, the identification codes of the collected proteins were obtained from *UNIPROT*, followed by the construction of the protein interaction network and its topological and functional analysis using *Cytoscape*, based on the protocol by Casotti *et al.* (2022). In the functional analysis, pathway enrichment linked the proteins to the Janus Kinase/Signal Transducer and Activator of Transcription (JAK/STAT) pathway, an important viral clearance mechanism associated with SARS-CoV-2, which, by binding to angiotensin-converting enzyme 2 (ACE-2), undergoes endocytosis and activates biochemical cascades to induce the expression of interferons α and β (IFN-α and IFN-β), as well as pro-inflammatory cytokines, which help sustain acute inflammation. Additionally, interferons interacting with the JAK-STAT pathway can induce cascades that result in the expression of interferon-stimulated genes (ISGs) to combat SARS-CoV-2. In this process, IFN-α, involved in serotonin reduction, increases the production of neurotoxic compounds through the conversion of tryptophan into quinolinic acid and 3-hydroxykynurenine, perpetuating neuroinflammation associated with cognitive decline. Furthermore, the progression of inflammation is likely linked to an increase in interleukin-6 (IL-6), amplifying the permeability of the blood-brain barrier (BBB) and recruiting interleukin-1B (IL-1β), C-reactive protein (CRP), and tumor necrosis factor-alpha (TNF-α), furthering neuroinflammation. In summary, this research, along with numerous other studies, supports the notion that the chronic inflammatory storm triggered by the COVID-19 virus may induce depressive and anxiety disorders through interference with

neuroplasticity, neurogenesis, and synaptogenesis in emotion-regulating areas such as the hippocampus, amygdala, and prefrontal and subgenual anterior cortex, following a complex and multifactorial mechanism that becomes less elusive with the findings of this study.

Keywords: System Biology, COVID, Depression, Anxiety.

References

Ahmad, S. J *et al.* (2022). Neurological sequelae of COVID-19. *Journal of integrative neuroscience*, p. 77.

Casotti, M. C. *et al.* (2023). Translational Bioinformatics Applied to the study of Complex diseases. *Genes*, p.419.

Kou, Y *et al.* (2022). Therapeutic potential of plant iridoids in depression: a review. *Pharmaceutical Biology*, p. 2167-2181.

Lorkiewicz, P., & Waszkiewicz, N. (2021). Biomarkers of post-COVID depression. *Journal of Clinical medicine*, p.4142.

Tan, P. H. *et al*. (2022). Emerging roles of type-I interferons in neuroinflammation, neurological diseases, and long-haul COVID. *International journal of molecular sciences*, p.14394.

Brazilian Symposium on Bioinformatics 2024

The role of polyploid giant cells in cancer progression and their potential as therapeutic targets: a bioinformatic overview

Isabele Pagani Pavan[1], Débora Gonçalves Barbosa[1], Bianca Paulino Campanharo[1], Matheus Correia Casotti[1], Débora Dummer Meira[1], Iuri Drumond Louro

1. Universidade Federal do Espírito Santo

Polyploid giant cells (PGCCs) are characterized by possessing multiple complete sets of chromosomes, unlike the two copies (diploid) typically found in humans, resulting in a much larger cell, present in certain types of cancer, such as subtypes of breast cancer (the most common in the female population). Robust evidence indicates that they play a crucial role in tumor survival and progression, as they are capable of generating daughter cells with an irregular number of chromosomes (aneuploidy), resistant and able to escape senescence, thus aiding in the perpetuation of the tumor. A deeper understanding of the pathways involved in the formation of PGCCs and their progeny has led to research into developing anti-PGCC therapies, some of which are promising in preclinical studies; however, the dynamic changes in the tumor microenvironment limit their success. Therefore, it is necessary to thoroughly understand the pathways that contribute to the success of PGCCs, establishing robust biomarkers for them and aiming for new combined treatments, especially for breast cancer. In this context, the protocol developed by the Computational and Systems Biology Group (SCBG) of the Human and Molecular Genetics Center (NGHM) at UFES was used, a methodology for constructing protein-protein interaction networks and performing topological and functional analyses. It includes a robust manual curation stage for bibliographic selection, selection of experimental proteins, construction and analysis of protein-protein interaction networks using the Cytoscape software and tools like stringApp, Network Analyzer, MCODE, and cytoHubba, as well as the enrichment of biological pathways in databases such as KEGG and Reactome. The constructed network exhibited typical characteristics of natural biological networks, such as high heterogeneity and a low average degree, indicating that most nodes are not hubs (have few connections). The short path length, high clustering coefficient, and small diameter suggest a "small-world" network, indicating that the proteins have high cohesion and significant interactions with each other. The studied network largely contains proteins that interact with kinases, especially cyclins, which act in the cytosol, are involved in cell cycle progression and checkpoints, and when deregulated, can lead to the formation of PGCCs. Other proteins present are associated with cancer signaling pathways. These results align with the accumulating data over the last decade on PGCCs, as previously mentioned. Finally, through the mentioned applications, the five most relevant proteins in the network were selected: CCNA2, CDK1, MAPK3, PLK1, and P53, with CDK1, MAPK3, and PLK1 already described in the literature as potential biomarkers for PGCCs. Future in vitro and in vivo studies are necessary to verify the viability of these proteins as pharmacological targets for new treatments against breast cancer.

Keywords: PGCCs; Breast cancer; Biomarkers; Aneuploidy; Protein-protein interaction networks;

References

Bharadwaj, D. and Mandal, M. (2020). Senescence in polyploid giant cancer cells: A road that leads to chemoresistance. In: *Cytokine & Growth Factor Reviews*, Vol. 52, pp. 68-75.

Casotti, M. C., *et al*. (2021). Construindo Redes de Interação Proteína-Proteína por Curadoria Manual. In: *BioInfo*, 23 Nov., https://bioinfo.com.br/construindo-redes-de-interacao-proteina-proteina-por-curadoria-manual/, Acess in september 2024.

Fu, J., *et al*. (2022). The role of cell division control protein 42 in tumor and non-tumor diseases: A systematic review. In: *Journal of Cancer*, Vol. 13, No. 3, pp. 800.

Liu, K., *et al.* (2020). Association and clinicopathologic significance of p38MAPK-ERK-JNK-CDC25C with polyploid giant cancer cell formation. In: *Medical Oncology*, Vol. 37, pp. 1-11.

Liu, P., Wang, L., and Yu, H. (2024). Polyploid giant cancer cells: Origin, possible pathways of formation, characteristics, and mechanisms of regulation. In: *Frontiers in Cell and Developmental Biology*, Vol. 12, Article 1410637.

Saini, G., *et al*. (2022). Polyploid giant cancer cell characterization: New frontiers in predicting response to chemotherapy in breast cancer. In: *Seminars in Cancer Biology*, Academic Press, pp. 220-231.

Zhao, Y., *et al*. (2024). Surviving the storm: The role of poly and depolyploidization in tissues and tumors. In: *Advanced Science*, Article 2306318.

Zhou, M., *et al.* (2023). Single-cell morphological and transcriptome analysis unveil inhibitors of polyploid giant breast cancer cells in vitro. In: *Communications Biology*, Vol. 6, No. 1, Article 1301.

Brazilian Symposium on Bioinformatics 2024

Transcriptomic analysis of *Crassostrea gigas* oysters exposed to tamoxifen demonstrates alterations in cancer-associated metabolic pathways

Ramon Diedrich[1,2], Fernanda Luiza Ferrari[1,3], Juliana Tisca[1,2], Tomás Bohn[1,2] Pessatti, Guilherme Toledo e Silva[1,3], Afonso Celso Dias Bainy[1,2]

1. Federal University of Santa Catarina
2. Center of Biological Sciences, Biochemistry Department
3. Center of Biological Sciences, Cellular Biology, Embriology and Genetics Department

Tamoxifen (TAM) is a drug used to treat breast cancer in premenopausal patients with estrogen receptor-positive tumors, acting as an antagonist of this receptor. TAM is metabolized by P450 isoforms CYP2D6 and CYP3A4, producing metabolites such as afimoxifen and endoxifen, which are absorbed by the intestinal system and eliminated via urine and feces. TAM can bioaccumulate through trophic transfer via sediments and water. Bivalves are recognized as sentinel organisms, allowing molecular evaluation of transcriptomic responses to detect potential molecular changes. For the experiment, five *Crassostrea gigas* oysters were used per aquarium, with a total of six aquariums. The experimental groups included one exposed to tamoxifen at a concentration of 100 ng/L diluted in 0.0001% DMSO and a control group treated with DMSO only. Digestive glands and gills were collected after 24 hours of exposure. RNA was extracted using the Trizol®/QIAzol® reagent and treated with the Qiagen® RNase-free DNase kit, precipitated with ammonium acetate, and quantified using Nanodrop. The transcriptome was assembled via RNA-Seq. The quality of reads was assessed using the FastQC program, and mapping was performed against a reference genome using the STAR program. Gene counts were determined using the htseq-count program. Differential expression analysis was carried out using the DESeq2 library in R. Principal Component Analysis (PCA) was performed for both digestive gland and gill samples using the FactoMineR and factoextra packages. Gene ontology enrichment was conducted using the weight01 algorithm from the R topGO library, with significant terms defined as $p < 0.05$. Enriched KEGG metabolic pathways were identified using the Benjamini-Hochberg multiple testing correction ($p.adj < 0.05$). A total of 26,598 genes were identified from the mapped reads, with GC content between 40% and 42%. Mapping coverage ranged from 75.80% to 80%, indicating good mapping quality. PCA results showed clustering of samples exposed to tamoxifen, whereas the control group data demonstrated separation, with no clear similarities to the TAM-exposed group. Differential expression analysis revealed 2,492 genes in the digestive gland and 154 genes in the gills, with a predominance of underexpressed genes in both tissues. Gene ontology enrichment analysis in the gills highlighted biological processes related to cell adhesion and intracellular movement, while calcium ion binding was notable in the molecular function category. In the digestive gland, molecular function enrichment was related to transmembrane transport, and biological process terms such as "cellular communication," "transmembrane transport," and "carbohydrate metabolism" were significant. KEGG pathway enrichment, performed only on the digestive gland due to the role of tamoxifen biotransformation, identified lysosomal pathways, with two genes associated with apoptosis being overexpressed and one involved in cancer transcriptional deregulation. Glycosphingolipid biosynthesis pathways were also enriched, with overexpressed genes linked to cellular receptors for pathogens and ovarian cancer. Retinol metabolism, although antioxidant, was implicated in promoting tumor growth in some contexts. These findings indicate alterations in key cancer-related pathways.

Keywords: Tamoxifen, oyster, transcriptomics, cancer

References

Ahern, T *et al* (2020). Metabolic Pathway Analysis and Effectiveness of Tamoxifen in Danish Breast Cancer Patients. In *Cancer Epidemiology, Biomarkers & Prevention*, pages 582-590. AACR.

Alam, S. et al (2017). Altered (neo-) lacto series glycolipid biosynthesis impairs α2-6 sialytion on *N*-glycoproteins in ovarian cancer cells. In *Scientific reports*. Vol. 7, pages 1-18. Nature.

Balbi, T. et al (2021). Immunological Responses of Marine Bivavalves to Contaminant Exposures: Contribution of the -Omics Approach. In *Frontiers in Immunology*, pages. 1-11. Frontiers.

Fonseca, T. G *et al* (2024). Impacts of *in vivo* and *in vitro* exposures to tamoxifen: Comparative effects on human cells and marine organisms. In *Environment International*, pages 256-272. Science Direct.

Klein, D. J *et al*. PharmGKB summary: tamoxifen pathway, pharmacokinetics. In *National Institutes of Health*, pages 1-9. NIH.

Kuan, C. T. *et al* (2010). Multiple phenotypic changes in mice after knockout of the B3gnt5 gene, encoding LC3 synthase – a key enzyme lacto -neolacto ganglioside lipid. In *BMC Development Biology*, pages 1-20. Science Direct.

Pucci, M. *et al* (2022). Glycosyltransferases in Cancer: Prognostic Biomarkers of Survival in Patients Cohorts and Impacto in Malignancy in Experimental Models. In *Cancers*, pages 1-19. MDPI.

Transcriptomic Analysis Reveals a Novel MicroRNA in Porcine Fetuses from Gilds Supplemented with L-Arginine

Francelly G. Campos[1], Natanieli S. Máximo[1], Isabela de Oliveira Eiterer, Simone E. F. Guimarães[1]

[1]Departamento de Zootecnia – Universidade Federal de Viçosa (UFV)
Caixa Postal 36.570-900 – Viçosa – MG – Brazil

natanieli.maximo@ufv.br, francellycampos@gmail.com,
isabela.eiterer@ufv.br, sfacioni@ufv.br

Abstract

MicroRNAs (miRNAs) are small non-coding RNAs that regulate gene expression post-transcriptionally. They are involved in biological processes such as apoptosis, cell differentiation, growth, metabolism, and immune response. Arginine plays various biological roles, including protein synthesis, nitric oxide production, and the regulation of blood flow, with significant importance during pregnancy, as it is involved in fetal development. Given the limited understanding of arginine's role in miRNAs, especially during the fetal stage in pigs, the aim was to identify differentially expressed miRNAs in fetuses at 35 days of gestation, with and without L-arginine supplementation. RNA was extracted using Trizol from 10 pig fetuses, sourced from six sows supplemented and non-supplemented with L-arginine during pregnancy. After extraction, quantification was performed with a QUBIT fluorometer (Thermo Scientific, Waltham, MA, USA), and integrity was assessed using 1.0% agarose gel electrophoresis. Additionally, the Agilent 2100 BioAnalyzer (Agilent Technologies, Santa Clara, CA, USA) was employed to measure RNA integrity, with samples showing an RNA Integrity Number (RIN) above eight being used for library preparation with the Illumina TruSeq Small RNA kit, followed by sequencing on the Illumina HiSeq2500 platform. Post-sequencing, data quality control was carried out using Trimmomatic and Cutadapt tools. Initial mapping was performed with Bowtie and the Rfam database to eliminate tRNA and rRNA sequences. The remaining reads were mapped to the pig reference genome (Sscrofa11.1/susScr11) using the miRDeep2 tool through the mapper.pl script, which was also used to identify potential novel miRNAs. Differential expression analysis was conducted using the limma package in R. The results revealed a novel miRNA, chr2_17249, which was downregulated in fetuses from sows supplemented with L-arginine. This potential miRNA is located on SSC2, in intron 14 of the ATG2A (Autophagy Related 2) gene, and is predicted to target the ACACA (Acetyl-CoA Carboxylase Alpha) gene, which regulates fatty acid synthesis, suggesting that miRNAs may serve as modulators of its expression and fatty acid production. These findings demonstrate that miRNA activity can be altered by diet, broadening our understanding of swine biology and opening new avenues for future research on fetal development.

Keywords: bioinformatics, diet, non-coding RNA

References

Ambros, V. (2004). “The functions of animal microRNAs”, In: *Nature*, 431(7006), 350-355.

Aryal, B., Singh, A. K., Rotllan, N., Price, N., and Fernández-Hernando, C. (2017). “MicroRNAs and lipid metabolism”, In: *Current opinion in lipidology*, 28(3), 273-280.

Chen, L., Song, J., Cui, J., Hou, J., Zheng, X., Li, C., and Liu, L. (2013). “microRNAs regulate adipocyte differentiation”, In: *Cell biology international*, 37(6), 533-546.

DeVeale, B., Swindlehurst-Chan, J., & Blelloch, R. (2021). “The roles of microRNAs in mouse development”, In: *Nature Reviews Genetics*, 22(5), 307-323.

Gebert, L. F., & MacRae, I. J. (2019). “Regulation of microRNA function in animals”, In: *Nature reviews Molecular cell biology*, 20(1), 21-37.

He, L., and Hannon, G. J. (2004). “MicroRNAs: small RNAs with a big role in gene regulation”, In: *Nature reviews genetics*, 5(7), 522-531.

Komatsu, S., Kitai, H., and Suzuki, H. I. (2023). “Network regulation of microRNA biogenesis and target interaction”, In: *Cells*, 12(2), 306.

Lin, N., Chang, K. Y., Li, Z., Gates, K., Rana, Z. A., Dang, J., ... and Rana, T. M. (2014). “An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment”, In: *Molecular cell*, 53(6), 1005-1019.

Nicoli, S., Standley, C., Walker, P., Hurlstone, A., Fogarty, K. E., and Lawson, N. D. (2010). “MicroRNA-mediated integration of haemodynamics and Vegf signalling during angiogenesis”, In: *Nature*, 464(7292), 1196-1200.

O'Brien, J., Hayder, H., Zayed, Y., and Peng, C. (2018). “Overview of microRNA biogenesis, mechanisms of actions, and circulation”, In: *Frontiers in endocrinology*, 9, 402.

Su, L., Zhao, S., Zhu, M., and Yu, M. (2010). “Differential expression of microRNAs in porcine placentas on days 30 and 90 of gestation”, In: *Reproduction, Fertility and Development*, 22(8), 1175-1182.

Brazilian Symposium on Bioinformatics 2024

Unlocking The Anti-aging Potential: *In silico* Analysis of Astaxanthin, Curcumin, Quercetin, and Resveratrol in Modulating Skin Aging Pathways

Debora Gonçalves Barbosa[1], Karen Ruth Michio Barbosa[1], Lorena Souza Castro Altoé[1], Matheus Correia Casotti[1], Rahna Gonçalves Coutinho da Cruz[1], Yasmin Moreto Guaitolini[1], Iúri Drumond Louro[1], Débora Dummer Meira[1]

1. Systems and Computational Biology Group (SCBG), Center for Human and Molecular Genetics (NGHM), Federal University of Espírito Santo (UFES)

Aging is associated with diverse intrinsic and extrinsic factors, among them UV exposure. Both internal and external stimuli cause short-term and long-term skin inflammation by stimulating the production of reactive oxygen species (ROS), designated as the first line of defense against external agents, which promote the inflammatory process. As a consequence of ROS accumulation, there is an imbalance in the body between production of oxidant molecules and their neutralization by antioxidants, causing oxidative stress. In addition to the decrease in antioxidant capacity, oxidative stress is related to DNA damage, inflammation, and the production of matrix metalloproteinases (MMPs), enzymes that damage skin dermal proteins. In this context, the concept of inflammaging emerged as a chronic low-level inflammation associated with aging, manifested by high levels of pro-inflammatory factors such as IL-6. Thus, there is a recognized need for cosmeceuticals that modulate inflammation pathways to prevent and treat aging. In this sense, this study aimed to evaluate whether the bioactive compounds astaxanthin, curcumin, quercetin, and resveratrol are effective in treating the effects of skin aging, using *in silico* analyses. For this purpose, protein-protein interaction networks (PPINs) related to skin aging and the bioactive compounds were generated using the *Cytoscape plug-in.*

The key genes for each network were identified from the Bottleneck (BN) analysis: *IL-6* (general and astaxanthin), *TAB1* (curcumin), *TNF-α* (quercetin), and *TP53* (resveratrol). IL-6, a pro-inflammatory cytokine, plays a pivotal role in the development of inflammaging and is inhibited by astaxanthin. In this context, the excessive accumulation of ROS in the skin, combined with the imbalance of the redox state due to intrinsic aging, can increase the release of pro-inflammatory cytokines, such as IL-6 and TNF-α. These interleukins have the potential to positively regulate the expression of mRNA, proteins, and the enzymatic activity of MMPs, thus aggravating the skin aging process. Therefore, it is understood that the anti-inflammatory effects of quercetin inhibit TNF-α activity, protecting the skin from damage caused by inflammaging.

Curcumin's BN network, on the other hand, has TAB1 as its central gene. This protein acts as a key adaptor with fundamental action in the activation of NF-κB, as well as in the production of pro-inflammatory cytokines in response to stimuli with toll-like receptors and cytokines. As NF-κB activation triggers the production of MMPs, it is inferred that curcumin negatively regulates TAB1, presenting anti-aging potential.

Finally, resveratrol BN showed TP53 as the most relevant gene, which is related to apoptosis inductions. Given that apoptosis is one of the aging cascades promoted by ROS, it is possible to correlate the action of resveratrol with apoptotic pathways, which may play an important role in maintaining skin integrity.

In short, aging is directly related to the inflammatory response, therefore the bioactive compounds studied may represent promising elements to be included in cosmetic products for topical application and/or candidates for oral drugs, as they are capable of reducing excessive oxidative stress and inflammatory processes, thereby preventing or delaying cellular aging.

Keywords: Bioinformatics. Aging. Inflammation. Natural Products. Systems Biology.

References

Barbosa, K. B. F., Costa, N. M. B., Alfenas, R. de C. G., De Paula, S. O., Minim, V. P. R., & Bressan, J.. (2010). Estresse oxidativo: conceito, implicações e fatores modulatórios. Revista De Nutrição, 23(4), 629–643. https://doi.org/10.1590/S1415-52732010000400013

Hossain, M. R., Ansary, T. M., Komine, M., & Ohtsuki, M. (2021). Diversified Stimuli-Induced Inflammatory Pathways Cause Skin Pigmentation. International journal of molecular sciences, 22(8), 3970. https://doi.org/10.3390/ijms22083970

Kohandel, Z., Farkhondeh, T., Aschner, M., & Samarghandian, S. (2021). Nrf2 a molecular therapeutic target for Astaxanthin. Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie, 137, 111374. https://doi.org/10.1016/j.biopha.2021.111374

Mohd Zaid, N. A., Sekar, M., Bonam, S. R., Gan, S. H., Lum, P. T., Begum, M. Y., Mat Rani, N. N. I., Vaijanathappa, J., Wu, Y. S., Subramaniyan, V., Fuloria, N. K., & Fuloria, S. (2022). Promising Natural Products in New Drug Design, Development, and Therapy for Skin Disorders: An Overview of Scientific Evidence and Understanding Their Mechanism of Action. Drug design, development and therapy, 16, 23–66. https://doi.org/10.2147/DDDT.S326332

Pallela, R., Na-Young, Y., & Kim, S. K. (2010). Anti-photoaging and photoprotective compounds derived from marine organisms. Marine drugs, 8(4), 1189–1202. https://doi.org/10.3390/md8041189

Pilkington, S. M., Bulfone-Paus, S., Griffiths, C. E. M., & Watson, R. E. B. (2021). Inflammaging and the Skin. The Journal of investigative dermatology, 141(4S), 1087–1095. https://doi.org/10.1016/j.jid.2020.11.006

Subedi, L., Lee, T. H., Wahedi, H. M., Baek, S. H., & Kim, S. Y. (2017). Resveratrol-Enriched Rice Attenuates UVB-ROS-Induced Skin Aging via Downregulation of Inflammatory Cascades. Oxidative medicine and cellular longevity, 2017, 8379539. https://doi.org/10.1155/2017/8379539

Vollono, L., Falconi, M., Gaziano, R., Iacovelli, F., Dika, E., Terracciano, C., Bianchi, L., & Campione, E. (2019). Potential of Curcumin in Skin Disorders. Nutrients, 11(9), 2169. https://doi.org/10.3390/nu11092169

Zhou, X., Cao, Q., Orfila, C., Zhao, J., & Zhang, L. (2021). Systematic Review and Meta-Analysis on the Effects of Astaxanthin on Human Skin Ageing. Nutrients, 13(9), 2917. https://doi.org/10.3390/nu13092917

Brazilian Symposium on Bioinformatics 2024

Large scale analysis of sialic acid incorporation mechanisms of microorganisms from intestinal microbiota

NASCIMENTO-SILVA; E.A. [1], CALOBA; P. [2], ALISSON-SILVA; F. [2], ZARAMELA, L.S. [1]

1. Ribeirão Preto Medical School (USP)
2. Carlos Chagas Filho Biophysics Institute(UFRJ)

Sialic acids are monosaccharides composed of nine carbons and they are mostly present in the cell surface of vertebrates. Among them, N-Acetylneuraminic acid (Neu5Ac) and the N-Glycolylneuraminic acid (Neu5Gc) are the major classes of sialic acid in mammals. However, in humans, the Neu5Gc is not synthesized due to one mutation in the CMAH gene, which encodes the enzyme responsible for the hydroxylation of Neu5Ac into Neu5Gc. The relevance of sialic acid has grown over the years, for example, levels of Neu5Gc can be found in the human body after dietary consumption with food sources rich in this class of sialic acid. The presence of Neu5Gc as xeno-autoantigen is linked to inflammation and cancer as tumor cells highly express Neu5Gc onto their surfaces. Sialic acids found in the terminal moiety of glycoconjugates have an important role in cell signaling and also in the interaction among different cells. Furthermore, pathogenic microorganisms present in the microbiota can scavenge the host's sialic acid and incorporate it on their cell surface. With this mechanism, the microorganisms are capable of evading the host's immune system and continue their proliferation in the organism without interference. This raises many questions: Overall, which microorganisms are capable of sialylating their cell walls and if microorganisms could incorporate Neu5Gc, which impact could this have in pathological states such as inflammation. For this purpose, this project aims to identify microorganisms capable of incorporating sialic acid into their cell walls. To achieve this, the project includes *in silico*, *in vitro*, and *in vivo* analysis. For the bioinformatics component, Hidden Markov Models (HMMs) were generated using the HMMER program for each protein involved in the sialylation process. RefSeq genomes from NCBI were then downloaded, and after filtering, they were matched with the protein models to identify relevant protein sequences in the genomes. Subsequently, a phylogenetic tree was generated and annotated for better understanding. After the filtration steps, 235 genomes were identified as having a complete sialylation pathway. From bioinformatic analysis, 77% of genomes do not possess any virulence factor, which can be associated with commensal microorganism. This unexpected finding is surprising, as the literature typically associates sialylation more with pathogenic microorganisms. This could suggest that potential commensal microorganisms could have a more impactful role in inflammation, particularly when incorporating the xeno-autoantigen Neu5Gc. Although, this is a preliminary result and further analysis is required. To solidify this hypothesis, *in vivo* analysis using a $Cmah^{-/-}$ mice model will be crucial to conclude our findings.

Keywords: Sialic acid, metabolism, host-pathogen interaction, microbiota

References

EDDY, S. R (2011). Accelerated Profile HMM Searches. PLOS Computational Biology, v. 7, n. 10, p. e1002195.

GHOSH, S (2020). Sialic acid and biology of life: An introduction. Sialic Acids and Sialoglycoconjugates in the Biology of Life, Health and Disease, p. 1.

JENNINGS, M. P.; DAY, C. J. and ATACK, J. M (2022). How bacteria utilize sialic acid during interactions with the host: snip, snatch, dispatch, match and attach. Microbiology, v. 168, n. 3, p. 1157.

SAMRAJ, A. N. et al (2015). A red meat-derived glycan promotes inflammation and cancer progression. Proceedings of the National Academy of Sciences of the United States of America, v. 112, n. 2, p. 542–547.

SAMRAJ, A. et al (2014). Involvement of a Non-Human Sialic Acid in Human Cancer. Frontiers in Oncology, v. 4.

Brazilian Symposium on Bioinformatics 2024

POSTER SESSION BETA ABSTRACTS

(Abstracts are sorted by title ascending order)

Cidade da Inovação (IFES), Vitória,
December 3, 2024

Summary

P30 – Harnessing Integrated Informatics and Molecular Simulation to Predict Antibody Epitopes on Viral Envelope Glycoproteins

P32 – Identification of Genetic Alterations in Patients Who Developed Physical Fatigue as a Long COVID Condition

P34 – Investigation of UDE-based approaches for cell cycle modeling

P36 – Metagenomics and Bioinformatic tools in Agricultural Microbiome

P38 – Multi-omics systems biology approach identifies novel signature genes for neuropsychiatric disorders

P40 – Non coding variants near the NOTCH1 gene are associated with frailty criteria in Brazilians older adults

P42 – Predicting aggregation region in proteins with machine learning based on tertiary structure: web platform

P44 – Predictive Modeling Of Post-covid-19 Hair Loss: Insights From Machine Learning And Logistic Regression

P46 – SP crime: A Python package for merging São Paulo criminal and medical data

P48 – The Role of Indel Variants in COVID-19: Unveiling Frequency Patterns and Potential Clinical Significance

P50 – Topology-based pan-cancer analysis of DLK1-DIO3-derived microRNA roles

P52 – Transcriptomic analysis of the aged female omentum. Insights on metastatic invasion in a mouse model of ovarian cancer

P54 – Tumor-Regeneration Interplay: Systems Biology and New Models in Comparative Study with Therapeutic Insights

P56 – Verdict: An Interactive Web Tool for Exploring Disease Modules and Drug Targets within the Human Interactome

P58 – Improvement of the assembly and Annotation of the Trypanosoma cruzi Genome Using Hi-C Data

P60 – Helix-Shifts and Incongruence in RNA Evolution

BSB 2024 Brazilian Symposium on Bioinformatics 2024

A Curated Dataset for Machine Learning Training to Predict Novel Peptide Inhibitors of Voltage-Gated Sodium Channels in Drosophila suzukii (Matsumura, 193)

Jailan da Silva Sousa [1], Joicymara Santos Xavier [1,3], Bruno Silva Andrade[2,4]

1. Federal University of Minas Gerais [1], Southwest Bahia State University [2], Federal University of the Jequitinhonha and Mucuri [3], INRAe UMR STLO, Rennes [4]

Drosophila suzukii (Matsumura, 1931) is a highly destructive agricultural pest, primarily affecting soft-skinned fruits. Current control strategies rely heavily on chemical insecticides, which pose significant risks to non-target species, including pollinators, and contribute to environmental contamination. To address these challenges, the use of natural toxins, particularly peptides, has emerged as a promising alternative. These molecules demonstrate high specificity and selectivity, offering a more sustainable approach to pest management. Voltage-gated sodium channels (Navs) are critical molecular targets for insecticides, and natural toxins modulate these channels by binding to three distinct sites, providing an avenue for the development of novel bio-insecticides. In this study, we curated a dataset of 125 Nav-inhibiting toxins from databases such as UniProt, NCBI, and ArachnoServer, as well as relevant literature, filtered using LitSuggest. Among these, 54 toxins were confirmed to interact with one of the three Nav binding sites in insects. However, other toxins exhibited insecticidal activity without an identified mechanism or targeted Navs in non-insect organisms. To elucidate the mode of action of these additional toxins and expand the dataset, molecular docking simulations were performed using HADdock and Zdock, two of the most advanced tools for protein-ligand interaction prediction. For validation, decoys were generated by randomly permuting amino acid residues and fully reversing sequences of active toxins. A second round of decoy generation involved substituting 50% of amino acids with similar types based on blosum62 matrix scores. Both active toxins lacking structural data in the Protein Data Bank (PDB) and decoys were modeled using AlphaFold2, and docking simulations were conducted for each toxin in its respective Nav binding site. HADdock achieved area under the curve (AUC) values of 0.59, 0.51, and 0.38 for sites 3 (α-toxins), 4 (β-toxins), and E (pore domain), respectively. Zdock yielded AUC values of 0.61, 0.33, and 0.36 for the same sites. For decoys with 50% residue substitutions, AUC values improved to 0.63, 0.53, and 0.52 for HADdock, and to 0.69, 0.41, and 0.41 for Zdock. Given the limitations of docking-based approaches, an alternative method leveraging Foldseek was employed. This technique, which is based on sequence and structural similarity, identified 172 additional toxins with TM-scores ≥ 0.5 and Fiden scores ≥ 0.7. These toxins were subjected to docking with Zdock and their performance compared to the original active toxin dataset. The final dataset, consisting of 298 toxins (125 from the original selection and 172 from Foldseek), forms the basis for training machine learning models to identify novel insecticidal toxins with higher specificity for *D. suzukii* Navs.

Keywords: Natural toxins, Curated dataset, Molecular Docking, Bio-insecticides

References

Allot, A. et al. (2021). LitSuggest: a web-based system for literature recommendation and curation using machine learning. *Nucleic acids research*,

Andreazza, F. et al. (2017). Drosophila suzukii in southern neotropical region: current status and future perspectives. *Neotropical entomology*, p. 591-605.

Coudert, E., et al. (2023). Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinformatics*.

de Lera Ruiz, M., and Kraus, R. (2015). Voltage-gated sodium channels: structure, function, pharmacology, and clinical indications. Journal of medicinal chemistry, p. 7093-7118.

Eng, J. (2005). Receiver operating characteristic analysis: a primer1. *Academic radiology*, p. 909-916.

Ffrench-Constant, R. et al. (2016). Ion channels as insecticide targets. Journal of Neurogenetics, p. 163-177.

Garcia, F. (2020). Drosophila suzukii management. New York: *Springer International Publishing*, p. 93-110.

Jumper, J. et al. (2020). AlphaFold 2. Fourteenth Critical Assessment of Techniques for Protein Structure Prediction.

Honorato, R. et al. (2021). Structural biology in the clouds: the WeNMR-EOSC ecosystem. *Frontiers in molecular biosciences*.

LI, Z., WU, Q., YAN, N. (2024). A structural atlas of druggable sites on Nav channels. Channels, p. 2287832.

Pierce, B., Hourai, Y., Weng, Z. (2011). Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PloS one*.

Pineda, S. et al. (2018). ArachnoServer 3.0: an online resource for automated discovery, analysis and annotation of spider toxins. *Bioinformatics*, p. 1074-1076.

Van Kempen, M. et al. (2024). Fast and accurate protein structure search with Foldseek. *Nature biotechnology*, p. 243-246.

Brazilian Symposium on Bioinformatics 2024

A graphical bioinformatics tool for delineating viral taxonomic levels

Igor C. Santos[1], Arthur Gruber[2]

1. Biotechnology undergraduate course, Escola de Artes, Ciências e Humanidades, EACH/USP, São Paulo, Brazil
2. Instituto de Ciências Biomédicas, ICB/USP, São Paulo, Brazil

Viruses display a far greater diversity than that observed in cellular organisms and viral taxonomy remains poorly organized. Given the significant advancements in viral genome sequencing, particularly from metagenomic samples, the development of reliable taxonomic classification tools has become essential. Tools like GRAViTy, vConTACT, VICTOR, PASC, DEmARC, and Sequence Demarcation Tool, employ biological sequences to determine viral relationships. In recent years, accurate 3D structure prediction methods have been developed, and programs like Dali, TM-align, and Foldseek can generate pairwise structural alignments. Over the course of evolution, proteins can maintain their functional domains through 3D structure, even when primary sequences are highly divergent. Thus, structural homology of viral proteins can unravel relationships at deep taxonomic levels, such as orders and classes. The development of an integrated tool, employing various metrics and presenting a graphical output, would be highly beneficial for taxa delineation and visualization of viral groups. In this work, we present 3A-DGT (All-Against-All Distance Graphical Tool), a toolbox for the analysis and graphical visualization of biological distance data using primary sequences and 3D protein structures. Five metrics are currently implemented: sequence identity and similarity, maximum likelihood (ML) distance, structural similarity (TM-score), and 3Di character similarity. For each metric, the program performs sample clustering and generates heatmaps and distance trees. Additionally, 3A-DGT can generate FASTA sequence files of the classified groups.

We employed five different viral sequence datasets to validate the program: dsRNA viruses of the *Amalgaviridae* and *Totiviridae* families, ssDNA viruses of the *Microviridae* family, and dsDNA viruses of tailed phages of the *Caudoviricetes* class. In the case of the dsRNA viruses, the results confirmed that RNA-dependent RNA polymerase (RDRP) is much more conserved than the ORF1 protein whose function remains unclear, within *Amalgaviridae*. Moreover, the resulting RDRP distance tree confirmed that *Amalgaviridae* and *Totiviridae* are sister clades, in agreement with our previous phylogenetic analyses. In the case of *Microviridae*, heatmap graphs exhibited high conservation, especially on metrics based on structural comparisons, and the resulting distance tree presented good congruence with a maximum likelihood tree. Caudoviricetes are highly diverse, encompassing members of numerous distinct families. Indeed, the heatmaps of all metrics revealed considerably lower conservation compared to *Totiviridae*, *Amalgaviridae*, and *Microviridae*. Across all datasets, we

observed increasing conservation among viruses in the following order of metrics: identity, similarity, maximum likelihood distance, 3Di character similarity, and TM score. These findings substantiate that structural comparisons can be more sensitive in identifying taxonomic relationships than amino acid sequences.

In comparison to SDT and Dali programs, 3A-DGT integrates a more extensive range of metrics within a single toolbox, rendering it suitable for the comparison and demarcation of viral sequences with varying degrees of conservation. We are currently implementing the generation of multiple sequence alignments (MSA) with structural information, utilizing either the Foldmason program with 3Di characters or reseek to generate a mega alphabet and MUSCLE5 to align the sequences. These MSAs will be employed to estimate ML distances.

Acknowledgements

ICS received a Scientific Initiation fellowship from FAPESP. Correspondence: argruber@usp.br

Keywords: *Virus taxonomy, pairwise distance, heatmap, Totiviridae, Amalgaridae*

Brazilian Symposium on Bioinformatics 2024

ANALYSIS OF RBD-SPIKE INTERACTIONS OF SARS-COV-2 OMICRON VARIANTS WITH ACE2 THROUGH MOLECULAR DYNAMICS

Raphaella Luisa Fernandes de Almeida, Heberth de Paula, Greiciane Gaburro Paneto

Universidade Federal do Espírito Santo/Departamento de Farmácia e Nutrição, Alto Universitário, s/n° - 29500-000 – Alegre-ES – Brasil.

raphaella.almeida@edu.ufes.br, heberth.paula@ufes.br, greiciane.paneto@ufes.br

Abstract

The fast-spreading transmission of Omicron results from mutations that increase its infectivity, facilitating the interaction between the virus's RBD and the ACE2 receptor on the target cell (MOURA et al., 2022; FREITAS; GIOVANETTI; ALCANTARA, 2021). These mutations in the S protein enhance the viral affinity for the receptor, impacting transmission (TEGALLY et al., 2020). With over 16 million SARS-CoV-2 genomes in databases such as GISAID, this study aims, through bioinformatics, to analyze the mutations of the Omicron variant from Espírito Santo and its interactions with ACE2, seeking to identify interactions of greater stability. This study investigates the impact of mutations within the SARS-CoV-2 Omicron variant identified in Espírito Santo, Brazil, on its binding affinity to the human ACE2 receptor, a crucial step in viral infection. The research leveraged the complete genomic sequence of the Omicron variant obtained from the GISAID database (https://www.gisaid.org/) (ID: 1383367). The RBD was identified through sequence alignment using the NCBI database (https://www.ncbi.nlm.nih.gov/). Its three-dimensional structure was subsequently predicted using AlphaFold3 (https://alphafoldserver.com/). Molecular dynamics simulations were performed using GROMACS 2024.2 (ABRAHAM et al., 2024), employing the CHARMM36 (BEST et al., 2012) force field. This computationally intensive process involved meticulous system preparation, including solvation with TIP3P (BEST et al., 2012) water molecules, neutralization with 0.15 M NaCl, and minimized using the SD method until the maximum force was below 1,000 kJ/mol·nm and subjected to 100 ps under constant volume (NVT) and constant pressure (NPT) conditions at 300 K and 1.0 N/m². The 50 ns simulation was visualized in PyMOL, with RMSD, RMSF, and SASA analyses conducted using GROMACS tools. Interactions between RBD and ACE2 were evaluated with PLIP software (SALENTIN et al., 2015). The resulting data were organized into an interaction map using in-house Python scripts, and the ±SEM was calculated using the PBSA method in gmx_MMPBSA (VALDÉS-TRESANCO et al., 2021). The analysis revealed an RMSD variation between 0.25 and 0.35 nm. The SASA of the RBD domain remained around 377 nm², peaking at 70 ns. The RMSF showed fluctuations corresponding to loop regions, indicating intermittent interactions and suggesting greater flexibility of residues in this region compared to the 7KMB. Key interactions included ionic bonds involving residues ARG800, LYS837 and LYS855. Cation-π interactions between ARG800 and HIS34, as well as π-π stacking between PHE883 and TYR83. Binding free energy calculations yielded a value of -15.89 kcal/mol (±1.52 kcal/mol), compared to the -32.44 kcal/mol (± 0.67 kcal/mol) calculated for the wild-type interaction. Comparative sequence analysis revealed fourteen distinct mutations in the Espírito Santo Omicron RBD compared to the wild-type, including

D805N, R808S, K817N, G846S, E884A, and Q893R. An amino acid insertion was also observed in the wild-type sequence between positions 897-911. Residues MET82, LYS441, VAL442, and PHE483, exhibiting negative energy values while LYS31, GLU35, ASP38, and LYS353 displayed positive energy values, a potentially destabilizing effect. This study highlights the importance of investigating the mutational landscape of SARS-CoV-2 variants and their impact on ACE2 binding dynamics, offering insights into the molecular mechanisms of viral adaptation.

Keywords: *Bioinformatics. SARS-CoV-2. Angiotensin II Converting Enzyme. Molecular Dynamics Simulation.*

References

ABRAHAM, M. et al. (2024). GROMACS 2024.2 Manual (2024.2). Zenodo. https://doi.org/10.5281/zenodo.11148638.

BEST, R. B. et al. "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi1 and chi2 dihedral angles", Journal of Chemical Theory and Computation, 8: 3257-3273, 2012, PMC3549273.

FREITAS, A. R. R, GIOVANETTI, M., ALCANTARA, L. C. J. Variantes emergentes do SARS-CoV-2 e suas implicações na saúde coletiva. InterAm J Med Health, v. 4, p. 1-8, 2021.

JORGENSEN, W. L. et al. (1983). Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics, 79(2), 926-935. doi:10.1063/1.445869.

MOURA, E. C. et al. Covid-19: evolução temporal e imunização nas três ondas epidemiológicas, Brasil, 2020–2022. Revista de Saúde Pública, v. 56, 2022.

SALENTIN, S. et al. PLIP: fully automated protein-ligand interaction profiler. Nucleic Acids Res. 2015 Jul 1;43(W1). doi: 10.1093/nar/gkv315. Epub 2015 Apr 14. PMID: 25873628; PMCID: PMC4489249.

SHU, Y., MCCAULEY, J. GISAID: Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance, v. 22, n. 13, p. 30494, 2017.

TEGALLY, H. et al. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. medrxiv, p. 2020.12. 21.20248640, 2020.

VALDÉS-TRESANCO, M. S. et al. gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS. Journal of Chemical Theory and Computation, 2021 17 (10), 6281-6291. https://pubs.acs.org/doi/10.1021/acs.jctc.1c00645.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Application of Machine Learning Algorithms for Identification of Viruses in Dark Matter from Next-Generation Sequencing

Gabriel Montenegro de Campos[1], Luan Gaspar Clemente[2], Alex Ranieri Jerônimo Lima[3], Milton Yutaka Nishiyama Junior[3], Eneas de Carvalho[3], Sandra Coccuzzo Sampaio[3], Maria Carolina Elias[3], Svetoslav Nanev Slavov[3]

1. Ribeirão Preto Medical School, University of São Paulo
2. College of Agriculture "Luiz de Queiroz", University of São Paulo
3. Butantan Institute

Metagenomic methods represent one of the most potent tools for identifying emerging or lesser-known viruses (SANTIAGO-RODRIGUEZ e HOLLISTER, 2022. With the advent of next-generation sequencing technologies (NGS) and taxonomic classifiers, it has become feasible to discern their genetic makeup and correlate the identified sequences with their respective taxa. However, a subset of sequences remains unassociated with known taxa, commonly called "dark matter." Dark matter poses a significant impediment to achieving a comprehensive understanding of the metagenome (KRISHNAMURTHY e WANG, 2017; SANTIAGO-RODRIGUEZ e HOLLISTER, 2022). Content residing within the dark matter can potentially unveil novel pathogens capable of infecting humans. Hence, the primary aim of this study is to delineate the viral content within the unclassified portion utilizing sequences obtained from samples collected via nasopharyngeal swabs of pediatric patients of Hospital das Clínicas de Ribeirão Preto, who tested negative for SARS-CoV-2 in 2021. The obtained samples were sequenced using the NGS technique and the raw data was subjected to quality control. Next, the entire human genome present was mapped and removed, the unmapped reads were then taxonomically classified. The entire unclassified part (dark matter) was subjected to a search for viral protein families that were used for training and testing of supervised machine learning techniques, namely Naïve Bayes, Random Forest, XGBoost, and LightGBM. Our approach revealed the presence of four predominant virus groups—Caudovirales, Enterovirus, Respiratory Syncytial Virus, and Torque Teno Virus—within the dark matter, indicating that certain genomic sequences evade taxonomic classification. Moreover, our findings indicate that while XGBoost exhibited superior performance, Random Forest yielded the most reliable outcomes.

Keywords: Viral metagenomics; Viral dark matter; Machine Learning; Bioinformatics.

References

Santiago-Rodriguez, T.M. and Hollister, E.B. (2022). Unraveling the viral dark matter through viral metagenomics. Frontiers in Immunology, v. 13, 2022. DOI: 10.3389/fimmu.2022.1005107.

Krishnamurthy, S.R. and Wang, D. (2017). Origins and challenges of viral dark matter. Virus Research, v. 239, p. 136–142, 2017. DOI: 10.1016/j.virusres.2017.02.002.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Bioinformatics analysis revealed that NOTCH1 expression in Glioblastoma Multiforme patients and Glioma Stem Cells is associated with impaired cellular OXPHOS and low immune infiltration

Iris Moreira da Silva [1,2], Matheus Correia Casotti [2], Rafael Rezende Borges [3], Flavia de Paula [2], Eldamária de Vargas Wolfgramm dos Santos [2], Débora Dummer Meira [2], Iúri Drumond Louro [2], Flávia Imbroisi Valle Errera [2], Estevão Carlos Silva Barcelos[4]

1. Fundação Doutor Amaral Carvalho, Jaú - SP, Brasil
2. Núcleo de Genética Humana e Molecular (NGHM) - Universidade Federal do Espírito Santo (UFES), Vitória - ES, Brasil
3. Instituto Federal do Espírito Santo (IFES), Vila Velha - ES, Brasil
4. Università degli Studi di Perugia (UNIPG), Perugia, Itália

Glioblastoma (GBM) is the most common brain tumor in adults, characterized by low survival rates and significant heterogeneity. A specific subset of GBM cells, known as glioma stem cells (GSCs), plays a crucial role in tumor progression and treatment resistance by maintaining stem cell-like characteristics. The *NOTCH1* signaling pathway is important for regulating stem cells in normal and neoplastic tissues within the central nervous system, potentially contributing to the aggressiveness of GBM. Given the challenges posed by GSCs, there is increasing interest in exploring their metabolic reprogramming and its influence on tumor microenvironment, highlighting the potential for novel therapeutic strategies. This study aims to: (i) assess whether *NOTCH1* expression differs between healthy brain tissue and GBM samples; (ii) characterize the GBM cohort and identify pathways or processes associated with varying levels of *NOTCH1* expression; (iii) analyze the signaling pathway changes that occur upon *NOTCH1* silencing in GSCs. We analyzed GBM samples on GEPIA2 to compare *NOTCH1* expression between TCGA-GBM and GTEx healthy samples. TCGA-GBM transcriptomic and DSS data were downloaded from UCSC Xena, and samples were grouped into *NOTCH1*-high and -low based on a DSS-based cutpoint calculated with *maxstat* package. Gene set enrichment analysis (GSEA) was employed to predict potential pathways and biological processes of the differentially expressed genes (DEGs) between different groups. Additionally, we analyzed transcriptomics data of GSCs (GSM2300615, GSM2300618, GSM2300616 and GSM2300619) from the study GSE86348 available on the GEO platform. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses were performed using the clusterProfiler package in the R environment. Our results revealed *NOTCH1* is highly expressed in brain tumors (~7X) and in all four GBM subtypes ($p<0.05$,

all) compared to healthy control. The Reactome signature “Signaling by *NOTCH1*” also exhibited an increasing trend in GBM cases compared to healthy samples. Among 631 patients in the TCGA-GBM cohort, 172 had available *NOTCH1* expression data, with 140 presenting disease-specific survival (DSS) data and classified as primary tumor. Of these, 99 (65%) were male and 54 (35%) female, with median ages at initial pathological diagnosis of 60 and 62, respectively. One hundred fifteen samples were identified as exhibiting high *NOTCH1* expression. GSEA and GO analysis comparing NOTCH1-high and -low groups revealed differences related to the cell metabolism pathway, demonstrate that samples with elevated levels of *NOTCH1* exhibit reduced oxidative phosphorylation (OXPHOS), reactive oxygen species (ROS) production, immune infiltration signatures and high tumor plasticity. The cellular metabolism profile altered by *NOTCH1* was further examined in GSC samples with *NOTCH1* silenced, revealing restoration of OXPHOS-related genes (NES=2.10) and suppression of carbon metabolism (NES=-1.80). GO analysis of these samples also indicated enhanced mitochondrial function, evidenced by heightened expression of mitochondrial genes (ND1, ND2, ND3, ND5, ND6). Additionally, we observed diminished expression of G6PD, a pivotal enzyme in the pentose phosphate pathway parallel to glycolysis, underscoring the *NOTCH1* in metabolic reprogramming. These findings suggest *NOTCH1* involvement in metabolic changes in GBM and GSCs, particularly in OXPHOS, potentially influencing alternative metabolism and pathogenic mechanisms.

Keywords: *Brain Tumor, Notch, Metabolism, Immune infiltration.*

References

Hothorn, T. (2022). Package “maxstat” title maximally selected rank statistics, https://cran.r-project.org/web/packages/maxstat/maxstat.pdf, October.

Lo Dico, A., Salvatore, D., Martelli, C., Ronchi, D., Diceglie, C., Lucignani, G. and Ottobrini, L. (2019). Intracellular Redox-Balance Involvement in Temozolomide Resistance-Related Molecular Mechanisms in Glioblastoma. Cells, 8(11), 1315.

Neftel, C. et al. (2019). An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. Cell, v. 178, n. 4, p. 835-849.

Brazilian Symposium on Bioinformatics 2024

Classification of HIV genomes through graph comparison and analysis

Tatiana Mari Saita[1], Matheus H. Pimenta-Zanon[1], Glaucia Maria Bressan[1], Arthur T. L. de Queiroz[2], Fabricio Martins Lopes[1]

1. Universidade Tecnológica Federal do Paraná (UTFPR) - Cornélio Procópio - PR - Brazil
2. Instituto Gonçalo Moniz, Fundação Oswaldo Cruz (Fiocruz Bahia) - Salvador - BA - Brazil

The comparison and analysis of genetic sequences allow researchers to identify similarities and differences among organisms, understand molecular evolution, and reveal significant genetic patterns; however, the large volume of data generated by next-generation sequencing technologies presents a significant challenge for conducting these comparisons and analyses using bioinformatics methods. In particular, methods based on sequence alignment don't cope very well with analyzing whole genomes in large quantities. Recently, alignment-free methods are being developed as a potential solution to address this issue. In such a manner, this study aims to present a free-alignment method for comparing genomic sequences using a graph representation of genomes.

The DNA sequences are initially transformed into graphs using the BASiNET method, where the FASTA format sequence is processed with a sliding window to identify subsequences of three nucleotides, with a word and step size of 3 nucleotides. Edges are established between adjacent triplets, resulting in a weighted directed graph. The weight of each edge signifies the frequency of co-occurrence of the corresponding triplets. Subsequently, this graph is transformed into a weighted adjacency matrix, which reflects the relationships between vertices and is employed to calculate various metrics for assessing sequence similarity. The comparison of adjacency matrices can be performed using metrics that quantify the distance between the matrices, facilitating the analysis of similarity between the sequences. In this work, to simplify the calculations, the matrices are vectorized and then the Hamming distance, Euclidean distance, Hausdorff distance, Jaccard distance, and Pearson correlation are applied. With the distance matrices, it is possible to conduct a cluster analysis and examine the similarity relationships among the sequences.

For this study, sequences of the HIV virus will be analyzed, which possesses a high mutation capacity primarily due to its error-prone reverse transcriptase enzyme, leading to significant genetic diversity. This genomic variability has necessitated the classification of HIV into groups, subtypes, and recombinant forms.The proposed approach was evaluated considering a set of HIV genomes, producing promising results in the classification of HIV-1 and HIV-2 types and the analysis of virus subtypes.

Keywords: Graph, adjacency matrix, HIV, subtypes.

Brazilian Symposium on Bioinformatics 2024

Comparative Analysis of Supervised Classifiers in Predicting COVID-19 Severity Using Data from 239 Exomes

Aléxia Stefani Siqueira Zetum[1], Felipe Passarela[1], Felipe Ataídes Míon[1], Flávio Rosendo da Silva Oliveira[3],Bartolomeu Aciolli Santos[2], Túlio de Lima Campos[2], Débora Dummer Meira[1], Iuri Drumond Louro[1].

1. Universidade Federal do Espírito Santo (UFES)
2. Fiocruz-PE
3. Instituto Federal de Pernambuco (IFPE)

Introduction: Machine learning algorithms effectively handle high-dimensional genomic data, such as exome data, addressing challenges such as prediction, classification, and dimensionality reduction. Objective: This study aimed to compare classifiers to predict the severity of COVID-19 using exome data from 239 samples. Methodology: The database included the exome of 239 patients, with 150 in the "Non-severe COVID" class and 89 in the "Severe COVID" class (CAAE: 37094020.6.0000.5060). Variants with up to 10% missing data were retained (4505). Python (v.3.12.5) was used to train the models, with libraries such as “pandas”, “scikit learn”, applying the linear kernel Support Vector Classifier (SVC). After data cleaning and removing variables that could introduce bias, the Recursive Feature Elimination (RFE) method was implemented to select the 20 most influential features. Five algorithms were tested: KNeighborsClassifier(KNN), LogisticRegression(LR), and DecisionTreeClassifier(DT), SVM with RBF kernel, and SVM with linear kernel. SimpleImputer was used to handle missing values, and evaluation metrics included AUC-ROC, accuracy, F1 Score, sensitivity, specificity, and cross-validation. Results and Discussion: With the 20 best features selected by SVM-RFE, kNN achieved an AUC of 0.60, an accuracy of 0.65, an F1 Score of 0.47, sensitivity of 0.80, and specificity of 0.40, indicating high sensitivity but challenges in identifying true negatives. The linear kernel SVM achieved an AUC of 0.83, accuracy of 0.85, F1 Score of 0.85, sensitivity of 0.90, and specificity of 0.77, demonstrating excellent overall performance. The RBF kernel SVM had an AUC of 0.79, accuracy of 0.82, F1 Score of 0.73, sensitivity of 0.93, and specificity of 0.64, standing out for its high sensitivity but with lower specificity. LR presented an AUC of 0.75, accuracy of 0.78, F1 Score of 0.68, sensitivity of 0.86, and specificity of 0.64, showing good distinction between classes and competing with linear SVM-RFE. DT achieved an AUC of 0.71, accuracy of 0.72, F1 Score of 0.64, sensitivity of 0.73, and specificity of 0.70, with slightly better performance than kNN, particularly in terms of specificity. The ideal model choice depends on the most relevant metric for the problem, such as sensitivity versus specificity, and the dataset's characteristics. The linear kernel SVM stood out for its ability to reduce complexity and avoid the curse of dimensionality, being preferable in situations with many attributes (Wiyono et al., 2019). According to Huang et al. (2003), SVM tends to perform better in terms of AUC, maintaining high true positive rates and minimizing false positives. Conclusion: The linear kernel SVM-RFE had the best overall performance, with the highest AUC, accuracy, F1 Score, and specificity. LR also proved competitive, being a good alternative depending on the specific application context.

Keywords: Machine learning, COVID-19, Genetic.

References

Wiyono S. et al. (2019). Comparative study of machine learning knn, svm, and decision tree algorithm to predict students performance. International Journal of Research-Granthaalayah, v. 7, n. 1, p. 190-196.

Huang, J., LU, J., LING, C. (2003). Comparing naive Bayes, decision trees, and SVM with AUC and accuracy. In: Third IEEE International Conference on Data Mining. IEEE, p. 553-556.

Asteris, P. et al. (2022) Genetic prediction of ICU hospitalization and mortality in COVID-19 patients using artificial neural networks. Journal of Cellular and Molecular Medicine, v. 26, n. 5, p. 1445–1455.

ANDRÉ FILIPE PASTOR et al. (2023). Human Genome Polymorphisms and Computational Intelligence Approach Revealed a Complex Genomic Signature for COVID-19 Severity in Brazilian Patients. Viruses, v. 15, n. 3, p. 645–645.

VAPNIK, V. (1997). The support vector method. In: International conference on artificial neural networks. Berlin, Heidelberg: Springer Berlin Heidelberg, p. 261-271.

GOSWAMI, M., SEBASTIAN, N. (2022). Performance analysis of logistic regression, KNN, SVM, Naïve Bayes classifier for healthcare application during COVID-19. In: Innovative data communication technologies and application: proceedings of ICIDCA 2021. Singapore: Springer Nature Singapore, p. 645-658.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Comparative Genomic Analyses reveal key characteristics for the Biocontrol and the Promotion of Plant Growth in *Paenibacillus* Strains

Luciano Nascimento de Almeida [1], Mirelly Jady Fernandes e Silva [1], Osiel Silva Gonçalves [1], Sumaya Martins Tupy [1], João Paulo Lopes da Rocha [1], Blenda de Freitas Rodrigues Jesuino [1], Gabriela Amaral Xavier [1], Mateus Ferreira Santana [1]

1. Eco-evolutionary Microbial Genomics Group, Molecular Genetics of Microorganisms Laboratory, Department of Microbiology, Institute of Applied Biotechnology to Agriculture, Federal University of Viçosa, Minas Gerais, Brazil.

Paenibacillus is a genus of Gram-positive, facultatively anaerobic, endospore-forming bacteria, comprising at least 240 species. Over the past two decades, interest in its biotechnological potential has increased, although there is still limited genomic information available about the genus. In this study, we performed a comparative genomic analysis of *Paenibacillus* genomes and investigated their potential to interact with plants and act as antagonists of phytopathogens. 1,500 *Paenibacillus* genomes available in August 2024 were downloaded from GenBank and evaluated using CheckM v1.0.13, considering parameters (>95% completeness and <0.5% contamination). A total of 428 genomes were selected for phylogenetic and global distribution analyses. The alignment of the genomes of *Paenibacillus* was constructed using CheckM v1.0.13. The phylogenetic tree was obtained using maximum likelihood with IQ-TREE v1.6.11. A set of 102 complete genomes was selected for subsequent analyses. The protein files were used in pan-genome analyses and categorization of core, accessory, and unique gene families according to the COG database using the Bacterial Pan Genome Analysis v. 1.3 pipeline. The identification of CAZymes was performed using the dbCAN 3. Plant growth-promoting genes was carried out using the BLASTP and HMMER tools on the PGPT-Pred database from PLaBAse, and a heatmap was generated using the Clustvis v.1.0 tool. Secondary metabolite biosynthetic gene clusters were predicted using antiSMASH v. 6.0. *Paenibacillus polymyxa* was the most predominant species (46). Global distribution was represented in a map generated using the 'ggplot 2' package in R, which indicated that the areas of greatest abundance were concentrated primarily in North America. 11.71% of the genomes were derived from soil samples, with *P. polymyxa* being the most abundant species, also found in rhizosphere samples, suggesting its role in plant growth promotion. The phylogenetic tree was built based on 43 marker genes using the LG+I+G4 model. The genome sizes ranged from 3.839 to 9.080 Mb, while the GC content varied from 40.0% to 63.5%. The lineages exhibited an average of 174, 4,620, and 775 core, accessory, and unique genes, respectively. According to Heaps Law, the pan-genome was open and increasing (b = 0.56). Genes related to nitrogen acquisition (*nifU, nifS* and *nifF*), iron acquisition (*lipA* and *lipB*),

indole-3-acetic acid (IAA) production (*trpD, trpA, trpF, trpB, trpS, trpE, trpC* and *trpG*), and phosphate solubilization (*ptsB, ptsS, ptsA, ptsC, phoU, phoR, phoH* and *phoA*) were ubiquitously present. *P. polymyxa* exhibited higher frequencies of phosphate-solubilizing genes. *Paenibacillus* synthesizes ten types of secondary metabolite clusters, with non-ribosomal peptide synthetase (NRPS) being the most predominant. The most frequently identified CAZyme family in the genomes was glycoside hydrolases (GH). By exploring these genomes, we provide new insights into ubiquitous characteristics of the genus and unique traits among species. Our research emphasizes the role of *Paenibacillus* in plant-microorganism interactions, highlighting its impact on plant growth promotion and biocontrol.

Keywords: Biological control, Secondary metabolites, Phytohormones, Sustainable agriculture, Bionformatics.

References

Blin, K., Shaw, S., Kloosterman, A. M., Charlop-Powers, Z., Wezel, G. P., Medema, M. H., Weber, T. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic acids research*, v. 49, n. W1, p. W29-W35.

Chaudhari, N. M., Gupta, V. K., Dutta, C. (2016). BPGA-an ultra-fast pan-genome analysis pipeline. *Scientific reports*, v. 6, n. 1, p. 24373.

Dai, X., Shi, K., Wang, X., Fan, J., Wang, R., Zheng, S., Wang, G. (2019*). Paenibacillus flagellatus* sp. nov., isolated from selenium mineral soil. *International Journal of Systematic and Evolutionary Microbiology*, v. 69, n. 1, p. 183–188, 1 jan.

Lal, S. and Tabacchioni, S. (2009). Ecology and biotechnological potential of *Paenibacillus polymyxa*: A minireview. *Indian Journal of Microbiology*, v. 49, n. 1, p. 2–10, mar.

Ming, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Haeseler, A., Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, v. 37, n. 5, p. 1530-1534.

Parks, D., Imelfort, M., Skennerton, C. T., Hugenholtz, P., Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, v. 25, n. 7, p. 1043-1055.

Patz, S., Gautam, A., Becker, M., Ruppel, S., Rodríguez-Palenzuela, P., Huson, D. H. (2021). PLaBAse: A comprehensive web resource for analyzing the plant growth-promoting potential of plant-associated bacteria. *BioRxiv*, p.12. 13.472471.

Xie, J., Shi, H., Du, Z., Wang, T., Liu, X., Chen, S. (2016*).* Comparative genomic and functional analysis reveal conservation of plant growth promoting traits in *Paenibacillus polymyxa* and its closely related species. *Nature - Scientific Reports*, v. 6, 9 fev.

Zheng, J., Ge, Q., Yan, Y., Zhang, X., Huang, L., Yin, Y. (2023). dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Research*, v. 51, n. W1, p. W115-W121.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Development of a web system responsible for automating the process of validating three-dimensional proteins

Carolina Barros da Costa [1], Márcio Rodrigues Miranda [1], Kaio Alexandre da Silva [1]

1. Federal Institute of Education, Science and Technology of Rondônia - IFRO

The analysis of the structures of a protein is a fundamental step for its better understanding, since through this it is possible to understand not only its functionality, but also its interaction with other molecules and its relationship with different biological activities. One tool that allows this analysis to be performed is the creation of Ramachandran graphs. Through this, it is possible to evaluate the stereochemical quality of protein structures, represented by the association of phi (φ) and psi (ψ) angles present in polypeptide residues, which are located in energetically favorable and unfavorable regions of the protein. A protein is composed of three main angles: omega (Ω), which is a fixed angle, and phi (Φ) and psi (Ψ), which are responsible for the conformational variation of the main chain, making it flexible. Furthermore, the creation of these graphs also helps in the evaluation of the quality of theoretical and experimental models involving proteins. Thus, the use of this approach becomes fundamental for this process. However, the manual analysis of these graphs, combined with the lack of tools that automate this process, makes this a time-consuming and error-prone activity, especially in cases where there is a large volume of experimental data. Therefore, the objective of this work is to develop a web system responsible for automating the process of generating Ramachandran graphs of proteins contained in Protein Data Bank - PDB format files, in order to facilitate their three-dimensional analysis and assist research in the area of structural bioinformatics and proteomics. The prototyping development process was used to develop the system, where the necessary requirements were raised to automate the processing for generating Ramachandran graphs on a large scale. Then, the Figma tool was used to create prototypes of screens, which served as the basis for the development of the web platform. For the development of the code and the construction of the database, the MySQL system was used. As for the languages, Python was used to generate graphs, PHP with Laravel framework for the development of the back-end and TypeScript for the development of the front-end. The web system developed allows users to submit PDB files, generating four different graphs that allow detailed analysis of the torsion angles of protein residues. In addition, the system sends by email and displays a report indicating the outlier residues responsible for facilitating the identification of unusual conformations and contributing to the assessment of the quality of the protein structure. Therefore, this web system stands out for its ability to automate the process of generating Ramachandran graphs efficiently, making the process of analyzing protein structures faster, easier and more practical. Its creation contributes to the advancement of studies in proteomics, especially in the understanding and validation of protein structures, and is a promising tool for bioinformatics.

Keywords: Ramachandran, bioinformatics, automation, proteomics.

Brazilian Symposium on Bioinformatics 2024

EEG-Based Schizophrenia Classification Using Vision Transformers and Microstate Analysis

João Vitor Maciel Vianna[1], Karin Satie Komati[1]

1. Instituto Federal do Espírito Santo (IFES), Campus Serra

Schizophrenia is a complex mental disorder characterized by disruptions in thought, perception, and social functioning [Patel et al. 2014]. Diagnosing schizophrenia is challenging due to its variable symptomatology and the absence of specific biological markers [Hany et al. 2024]. While traditional diagnostic approaches are grounded in behavioral assessments, recent advances in neuroimaging, particularly Electroencephalography (EEG), offer potential support for diagnostic processes [Sun et al. 2021]. EEG enables the tracking of neural oscillations across various frequencies and amplitudes, providing a means to observe specific energy fluctuations in the brain. Through the analysis of these fluctuations, researchers can compute Global Field Power (GFP), a measure of the brain's overall electrical activity, which can be used to generate topographic images known as microstates [Skrandies, 1990]. Microstates represent brief, spatially stable intervals of brain activity and reflect fundamental building blocks of cognitive processing. Studies have shown that patients with schizophrenia exhibit abnormal microstate patterns [Andreou and Leicht, 2020; Alves et al. 2024], such as an increased duration of microstate class C (linked to attention networks) and a reduced occurrence of class B (related to language functions). These abnormalities may serve as biomarkers, providing insight into cognitive disruptions associated with schizophrenia [Michel & Koenig, 2018]. Additionally, patients often show a higher prominence of microstate class D, a potential indicator of psychosis risk, and a reduction in alpha power (8–12 Hz) in frontal brain regions, associated with impaired attention and working memory [Tomescu et al., 2014; Berman & Stern, 2017]. These findings support EEG microstate abnormalities as early markers for schizophrenia risk, particularly as similar patterns are observed in individuals at high risk. This work hypothesizes that the classification of EEG microstates can be realized by machine learning models such as Vision Transformers (ViTs) [Dosovitskiy et al. 2020] applied to analyze microstate images. ViTs divide input images into patches and apply self-attention mechanisms. This approach allows ViTs to capture subtle spatial dependencies and identify patterns that may relate to schizophrenic traits. The application of Vision Transformers in analyzing EEG microstate images has demonstrated classification accuracies of over 60% in diagnosing schizophrenia; however, these results are based on preliminary experiments, utilizing a dataset with 28 patients, of whom 14 have schizophrenia and 14 are healthy [Olejarczyk and Jernajczyk, 2017], and an additional 84 patients from [Gorbachevskaya and Borisov, 2019], where 39 patients are healthy and 45 have schizophrenia. Further research is needed to validate these findings across diverse populations.

Keywords: Global Field Power, Neural Network, Machine learning in diagnostics.

References

Alves, L. M., Côco, K. F., De Souza, M. L., and Ciarelli, P. M. (2024). Identifying adhd and subtypes through microstates analysis and complex networks. Medical & Biological Engineering & Computing, 62(3):687–700.

Andreou, C., & Leicht, G. (2020). EEG microstates in schizophrenia: Findings, perspectives, and clinical implications. In Clinical Neurophysiology, 131(1), 42-51. doi:10.1016/j.clinph.2019.08.008.

Berman, R. A., & Stern, E. R. (2017). Functional neuroimaging of anxiety: A meta-analysis of emotional processing in PTSD, social anxiety disorder, and specific phobia. In American Journal of Psychiatry, 174(10), 1017-1025. doi:10.1176/appi.ajp.2016.16040439.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Gorbachevskaya, K. and Borisov, S. (2019). Eeg of healthy adolescents and adolescents with symptoms of schizophrenia. Available <http://brain.bio.msu.ru/eeg_schizophrenia.htm>.

Hany, M., Rehman, B., Rizvi, A., et al. (2024). Schizophrenia. StatPearls Publishing, Treasure Island (FL). Updated 2024 Feb 23.

Michel, C. M., and Koenig, T. (2018). EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: A review. In NeuroImage, 180, 577-593. doi:10.1016/j.neuroimage.2017.11.062.

Olejarczyk, E., & Jernajczyk, W. (2017). *EEG in schizophrenia* (Version V1) [Data set]. RepOD. https://doi.org/10.18150/repod.0107441.

Patel, K. R., Cherian, J., Gohil, K., and Atkinson, D. (2014). Schizophrenia: overview and treatment options. Pharmacy and Therapeutics, 39(9):638.

Skrandies, W. (1990). Global field power and topographic similarity. Brain topography, 3:137–141.

Sun, Q., Zhou, J., Guo, H., Gou, N., Lin, R., Huang, Y., Guo, W., and Wang, X. (2021). EEG microstates and its relationship with clinical symptoms in patients with schizophrenia. Frontiers in Psychiatry, 12:761203.

Tomescu, M. I., Rihs, T. A., Schneider, M., Gruber, C., Cattapan-Ludewig, K., Klaver, P., and Michel, C. M. (2014). Deviant dynamics of EEG resting state pattern in 22q11.2 deletion syndrome adolescents: A vulnerability marker of schizophrenia? In Schizophrenia Research, 157(1-3), 34-40. doi:10.1016/j.schres.2014.05.032.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

EVALUATION OF MACHINE LEARNING MODELS IN IDENTIFYING NEUROLOGICAL COMPLICATIONS OF COVID-19: AN INTEGRATED AND COMPARATIVE ANALYSIS

Basílio LA[1], Silva DRC[1], Zetum ASS[1], Ventorim VP[1], Altoé LSC[1], Morelatto SA[1], Meira DD[1], Louro ID[1],

1 Universidade Federal do Espírito Santo (UFES)

Objective: This study aimed to compare five machine learning algorithms in classifying neurological complications in COVID-19 patients to identify the best-performing model. **Methods:** 206 patient samples with neurological sequelae (brain fog, memory loss, mood changes, and sleep disorders). Sociodemographic variables , comorbidities, and biomarkers were also included in the analysis. The R programming language was used to develop the models, and the Synthetic Minority Over-sampling Technique (SMOTE) was applied to address class imbalance. Performance was assessed using F1 score, AUC, accuracy, precision, and sensitivity. The ML tools used included SVM, Elastic-Net, RF, XGB, and LGBM, employing the SMOTE resampling technique. Risk prediction was performed using Stepwise Logistic Regression (RLS). **Results: ML: Brain Fog:** Best results were observed with EN (F1:0.74; AUC: 0.64), XGB (F1:0.73; AUC: 0.58), LGBM (F1:0.71; AUC:0.59), RF (F1:0.7; AUC:0.59), and SVM (F1:0.69; AUC:0.56). Important variables for EN: obesity and clinical spectrum(CS). **Memory Loss:** XGB (F1:0.57; AUC:0.68), LGBM (F1:0.56; AUC:0.65), EN (F1:0.53; AUC:0.64), SVM (F1:0.52; AUC:0.58) and SVM (F1:0.69; AUC:0.61). Important variables for XGB: CS, gender and age. **Mood Changes:** SVM (F1:0.74; AUC:0.74), EN (F1:0.71; AUC:0.66), LGBM (F1:0.69; AUC:0.69), XGB (F1:0.69; AUC:0.6) and RF (F1:0.53; AUC:0.6). Important variables for SVM: CS and Chronic Cardiovascular Disease (CCD). **Sleep Disorders:** XGB (F1:0,68; AUC:0.6), SVM (F1:0.68; AUC: 0.54), LGBM (F1:0.58; AUC:0.63), RF (F1:0.56; AUC:0.61) and EN (F1:0.45; AUC:0.63). Important variables for XGB: obesity, gender and such as TNF-α. **Stepwise Logistic Regression (P<0.05): Brain Fog:** Severe CS (OR:3.0140, CI:1.2714–7.5743), Female Gender (OR:2.4720, CI:1.2771–4.9214), Chronic Lung Disease (CLD) (OR:3.1492, CI:1.1936–8.7073), Diabetes Mellitus (OR:3.1546, CI:1.3067–7.8031), Obesity (OR:1.9775, CI:1.0032–3.9263), TNF-α (OR:2.1883, CI:1.1548–4.2219). **Memory Loss:** Severe CS (OR:4.8569, CI:2.2780–10.838), Gender (OR:3.6473, CI:1.9451–7.1157). **Mood Changes:** Moderate CS (OR:7.9958, CI:2.9263–24.632), Severe CS (OR:8.2091, CI:3.0150–25.497), Gender (OR:2.2266, CI:1.1329–4.4884), Age over 55 years (OR:0.3638, CI:0.1500–0.8456), CCD (OR: 2.6222, CI: 1.2749–5.5398), D-dimer (OR:0.4040, CI:0.1576–0.9623), TNF-α (OR:2.2469, CI:1.1721–4.3839). **Sleep Disorders:** CLD (OR:3.1725, CI:1.2074–8.5764), Diabetes Mellitus (OR: 2.7034, CI:1.0772–6.8683), Obesity (OR:2.1574, CI:1.0845–4.3484), TNF-α (OR:2.4231, CI:1.2733–4.7083). **Discussion:** XGB performed best for **memory loss and sleep disorders**, while SVM excelled in **mood changes** and EN in **brain fog**. The importance of biomarkers such as TNF-α and IL-6 highlights the potential of integrating biological data into predictive models. **Conclusion:** Artificial intelligence tools can optimize the diagnosis and management of neurological sequelae in severe COVID-19 cases. However, the selection of the most appropriate algorithm depends on dataset characteristics. Continued development of automated tools to test multiple classifiers remains essential to refine these predictions.

Keywords: Machine learning, Long COVID, Artificial intelligence, Logistic regression, Neuropathy.

References

Alballa, N., & Al-Turaiki, I. (2021). Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. Informatics in medicine unlocked, 24, 100564.

Asadi-Pooya, A. A., Akbari, A., Emami, A., Lotfi, M., Rostamihosseinkhani, M., Nemati, H., ... & Shahisavandi, M. (2022). Long COVID syndrome-associated brain fog. Journal of medical virology, 94(3), 979-984.

Binka, M., Klaver, B., Cua, G., Wong, A. W., Fibke, C., Velásquez García, H. A., ... & Janjua, N. Z. (2022, December). An elastic net regression model for identifying long COVID patients using health administrative data: a population-based study. In Open Forum Infectious Diseases (Vol. 9, No. 12, p. ofac640). US: Oxford University Press.

Nitesh, V. C. (2002). SMOTE: synthetic minority over-sampling technique. J Artif Intell Res, 16(1), 321.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Davis, H. E., McCorkell, L., Vogel, J. M., & Topol, E. J. (2023). Long COVID: major findings, mechanisms and recommendations. Nature Reviews Microbiology, 21(3), 133-146.

Brazilian Symposium on Bioinformatics 2024

Exploring hybrid dynamic modeling of ordinary differential equations and data-driven models: From validation to expansion assisted by high-resolution mass spectrometry.

Fabio Montoni [1 2 3 6 *], Rosangela A. Moreira [1 2], Toshiki Hanada [7], Ronaldo N. de Sousa[1 2 3], Thompson E. P. Torres [1 2], Jun Adachi [7], Natsume Yayoi-Kitatani [6], Hugo A. Armelin [1 2 5] and Marcelo S. Reis [1 4]

[1] Center of Toxins, Immune-Response and Cell Signaling (CeTICS), Butantan Institute, Brazil;
[2] Cell Cycle Laboratory, Butantan Institute, Brazil;
[3] Bioinformatics Graduate Program, University of São Paulo, Brazil
[4] Laboratory of Artificial Intelligence and Inference in Complex Data (Recod.ai); Institute of Computing, University of Campinas, Brazil
[5] Biochemistry Department, Institute of Chemistry, University of São Paulo, Brazil
[6] the AI Center for Health and Biomedical Research (ArCHER), National Institutes of Biomedical Innovation, Health and Nutrition, Osaka, Japan.
[7] Laboratory of Proteomics for Drug Discovery, National Institutes of Biomedical Innovation, Health and Nutrition, Osaka, Japan.
[*] Corresponding authors. E-mail: msreis@unicamp.br, fabio.montoni.esib@esib.butantan.gov.br

Introduction
The inherent complexity of representing living organisms has been a current challenge in modern science. Several tools such as KEGG, Reactome and String store complex cell signaling reactions, allowing a static recreation of the biological events observed within years of research endeavors. Despite the efforts, this approach solely cannot bring accuracy when modeling complex cell signaling pathways. To overcome this, dynamic models can be used to bring reliable results, but with a high-cost of information. Our research group has opted for using Ordinary Differential Equations (ODE) model to represent the KRas-GTP activation by SOS in Y1 cells. This model has revealed that SOS could not solely explain the high levels of KRas-GTP on this lineage, and further investigation pointed RasGRP4 as an important GEF. Protein depletion by CRISPR followed by tumor growth assays showed that the depletion of this protein displayed reduced tumor size compared to controls, insight that could never be reached without the initial ODE modeling. This result gives us the glance that the role of GEFs is underrated, and much more can be learned from complex networks if the dynamic models could be extensively expanded. We believe that this expansion can bring a game-changing role in disease and drug discovery research.
Objectives
For this research, our goal is to expand the dynamic hybrid models and explore its limitations. To achieve that, we plan to combine the ODE with data-driven models to overcome the limitations of building a bigger network, that in turn will be fed by high-resolution mass spectrometry of normal and phosphoproteome kinetic data from Y1 cell, over a number of perturbations on the system.
Methods
To confirm the initial theoretical aspect of the aforementioned background, we depleted the protein RasGRP4 and KRas protein with the CRISPR technique and performed tumoral growth assay in nude mice. As for the high-resolution mass spectrometry, we analyzed the Y1 cell TMT-Phosphoproteome (DDA) and Normal proteome (DIA) on Thermo Lumos™ and Q-Exactive™ mass spectrometers respectively, in a 0 to 30 minutes timespan under FGF2, SOS inhibitor + FGF2, with or without FBS in triplicates. After that, an exploratory analysis was performed in the data, which will be further used to compose the future dynamic hybrid models.
CEUA #3976280723
Results and Conclusion.
The mouse tumor assays showed that CRISPR-depleted RasGRP4 protein cells exhibited lower tumor average and frequency, where only one achieved 1000 mm³, pointing for the complexity of the overlooked GEFs in cancer development. This result prompted us to expand the modeling using high-resolution mass spectrometry of the Y1 cell line, which is currently under exploratory analysis. This result will further compose the dynamic hybrid modeling, where the limitations and achievements will be explored, to give new insights for cell signaling studies as a whole.

Supported by: FAPESP, CAPES, UNICAMP and NIBIOHN

Keywords: *(maximum 5 words)*

Hybrid Modeling, Cancer Research, Mass Spectrometry, Cell Signaling

Brazilian Symposium on Bioinformatics 2024

Genetic study of patients with persistent neurocognitive sequelae after COVID-19

Yasmin Moreto Guaitolini[1], Felipe Ataides Mion[1], Danielle Ribeiro Campos da Silva[1], Aléxia Stefani Siqueira Zetum[1], Vinícius do Prado Ventorim[1], Saulo Almeida Morellato[1], Débora Dummer Meira[1], Iuri Drumond Louro[1]

1. Universidade Federal do Espírito Santo (UFES)

Introduction: COVID-19 is an acute respiratory infection caused by SARS-CoV-2 that might affect various systems, including the nervous system, leading to persistent sequelae. SARS-CoV-2's ability to penetrate brain tissue may trigger neurological complications, especially in those with genetic predispositions. **Objective:** To examine the relationship between genetic variants in participants affected by COVID-19 and their post-infection neurocognitive issues. **Methodology:** From 277 participants, 167 reported neurocognitive issues after the acute COVID-19 phase. They completed questionnaires for sociodemographic and clinical data, and provided blood samples for DNA sequencing. The analysis focused on genes involved in nervous system dysfunctions. Partial least squares (PLS) regression was used, with cross-validation and variable selection via the R statistical software (version 4.4.0). **Results and Discussion**: Significant results ($p<0.05$) were found for the *HTR2A* gene, with the rs6314 variant (OR=2.4856; IC 95%=1.1420-5.6362), and for the *HLADRB1* gene, with the rs17886882 (OR=2.1758; IC 95%=1.0701-4.5211) and rs9269960 (OR=2.1493; IC 95%= 1.0364-4.5709) variants. These alterations were identified as risk factors for post-infection neurocognitive sequelae. *HTR2A* codes for serotonin receptor 2A and plays crucial roles in cognition and mental health. Jokela et al. (2007) linked an *HTR2A* polymorphism to increased chances of developing depression. Modulation of this gene has also been related to anxiety reduction in mice due to decreased mRNA expression and receptor formation (Rhon et al., 2023), while Rhon et al. (2024) found that knocking down 5HT-2A improved memory in mice, suggesting its potential in treating neurocognitive disorders. *HLADRB1* has also been linked to cognitive issues, particularly in age-related dementia and neurodegenerative diseases. Many *HLADRB1* variants are linked to cognitive decline and Alzheimer's (Cătană et al., 2024). Payton et al. (2016) associated an allelic variant of this gene with multiple sclerosis and cognitive decline. When the infection occurs, SARS-CoV-2 can penetrate the central nervous system (CNS) through the blood-brain barrier (BBB), potentially causing neurologic complications. Persistent CNS inflammation due to deoxygenation, coagulation, and neurotransmitter dysregulation in the acute infection phase (Lazzaroni et al., 2021; Rivas-Vazquez et al., 2022), followed by neuroinflammatory cascades, microclots, and glial cell activation in the residual phase (Shafqat et al., 2023; Sideratou; Papaneophytou, 2023), has been documented. These processes may worsen in individuals with genetic mutations (Siqueira et al., 2024). We hypothesize that viral infection in the CNS acts as a trigger for individuals with genetic variants in *HTR2A* and *HLADRB1* to develop neurocognitive sequelae. **Conclusion:** Genetic variants in the *HTR2A* and *HLADRB1* may increase the risk of neurocognitive complications, especially following SARS-CoV-2 infection. The variants rs6314 (*HTR2A*) and rs17886882, rs9269960 (*HLADRB1*) were linked to cognitive impairments, including memory loss and mental fatigue. The role of these genes in brain function, as shown in previous research, supports the hypothesis that viral-induced neuroinflammation could exacerbate genetic vulnerabilities, leading to more severe neurocognitive outcomes.

Keywords: ***Neurocognitive sequelae, COVID-19, Genetics.***

References

Cătană, C. S., Marta, M. M., Văleanu, M., Dican, L., & Crișan, C. A. (2024). Human leukocyte antigen and microRNAs as key orchestrators of mild cognitive impairment and Alzheimer's disease: A systematic review. *International Journal of Molecular Sciences, 25*(15), 8544. https://doi.org/10.3390/ijms25158544

Jokela, M., Keltikangas-Järvinen, L., Kivimäki, M., Puttonen, S., Elovainio, M., & Rontu, R. (2007). Serotonin receptor 2A gene and the influence of childhood maternal nurturance on adulthood depressive symptoms. *Archives of General Psychiatry, 64*(3), 356. https://doi.org/10.1001/archpsyc.64.3.356

Lazzaroni, M. G., Piantoni, S., Galli, M., Manfredi, A. A., Meroni, P. L., & Rampulla, V. (2021). Coagulation dysfunction in COVID-19: The interplay between inflammation, viral infection and the coagulation system. *Blood Reviews*, 46, 100745. https://doi.org/10.1016/j.blre.2020.100745

Payton, A., Dawes, P., Platt, H., Morton, C. C., Moore, D. R., Massey, J., Horan, M., Ollier, W., Munro, K. J., & Pendleton, N. (2016). A role for HLA-DRB1*1101 and DRB1*0801 in cognitive ability and its decline with age. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 171B*(2), 209-214. https://doi.org/10.1002/ajmg.b.32393

Rivas-Vazquez, R. A., Rey, G., Quintana, A., & Rivas-Vazquez, A. A. (2022). Assessment and Management of Long COVID. *Journal of Health Service Psychology, 48*(1), 21–30. https://doi.org/10.1007/s42843-022-00055-8

Rohn, T. T., Radin, D., Brandmeyer, T., Linder, B. J., Andriambeloson, E., Wagner, S., Kehler, J., Vasileva, A., Wang, H., Mee, J. L., & Fallon, J. H. (2023). Genetic modulation of the HTR2A gene reduces anxiety-related behavior in mice. *PNAS Nexus, 2*(6), pgad170. https://doi.org/10.1093/pnasnexus/pgad170

Rohn, T. T., Radin, D., Brandmeyer, T., Linder, B. J., Wagner, S., Kehler, J., Vasileva, A., Wang, H., Mee, J. L., & Fallon, J. H. (2024). Intranasal delivery of shRNA to knockdown the 5HT-2A receptor enhances memory and alleviates anxiety. *Translational Psychiatry, 14*, 154. https://doi.org/10.1038/s41398-024-02879-y

Shafqat, A., Omer, M. H., Ibrahem Albalkhi, Ghazi Alabdul Razzak, Humzah Abdulkader, Saleha Abdul Rab, Belal Nedal Sabbah, Khaled Alkattan, & Yaqinuddin, A. (2023). Neutrophil extracellular traps and long COVID. *Frontiers in Immunology, 14.* https://doi.org/10.3389/fimmu.2023.1254310

Sideratou, C.-M., & Papaneophytou, C. (2023). Persisting shadows: Unraveling the impact of long COVID-19 on respiratory, cardiovascular, and nervous systems. *Infectious Disease Reports, 15*(6), 806-830. https://doi.org/10.3390/idr15060072

Siqueira, S., Farias, L., Pereira, M., & Souza, R. (2024). Interseção viral: Análise dos efeitos neurológicos pós-COVID no sistema nervoso. *Journal of Neurological Research*, 45–79. https://doi.org/10.22533/at.ed.5082405034

Brazilian Symposium on Bioinformatics 2024

Genomic insights into the association between carbohydrate transporters and antimicrobial resistance in *Staphylococcus aureus*

João Vitor Wagner Ordine[1] , Edson Alexandre do Nascimento Silva[1] and Lívia Soares Zaramela[1]

[1] Department of Biochemistry, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, SP, Brazil.

Staphylococcus aureus is a versatile pathogen that colonizes various tissues in distinct vertebrate hosts, reflecting its mobile genome and metabolic adaptability. Genomic elements may have been primarily selected for roles crucial to microbial metabolism, with their function in counteracting antimicrobials being secondary. Since antibiotics co-opt bacterial transporters to enter bacterial cells, changes in the transporters' composition due to the adjustment of bacterial metabolism to an environmental shift might challenge the susceptibility to antibiotics. Conversely, mutations affecting the expression or the activity of porins due to antibiotic selective pressure might modify bacterial metabolism. The metabolic adaptation of *S. aureus* to distinct niches within its hosts results partially from the acquisition of multiple carbohydrate transporter types that allow maximal uptake of host sugars during infection. Thus, we aimed to cross-compare genetic elements related to antimicrobial resistance (AMR) and carbohydrate transport in the global *S. aureus* genomic population. To that end, we retrieved all complete genomes of the pathogen available at NCBI as of March 2024, filtering for high quality using CheckM v1.2.2 and annotating with Prokka v1.12. Sequence types (STs) were determined using mlst v2.11, and SCCmec types were identified with staphopia-sccmec v1.0. Additional resistance genes were detected using ABRICATE. For carbohydrate transporters, we curated a database from UniProt and identified transporters through a BLASTp search. We assessed the correlation between transporter abundance and AMR, performing statistical analyses such as Wilcoxon tests, Spearman's rank correlations, and odds ratio. Phylogenomic and pangenome analyses were conducted using Phylophlan v3.0.67 and Roary v3.13.0, with maximum likelihood (ML) trees built with IQ-TREE multicore v2.1.4. Our findings highlight the intricate complexity of the *S. aureus* sugar transportome and antimicrobial resistome, revealing high variability in presence and abundance across lineages. Statistical and phylogenomic analyses suggest a potential relationship between sugar transporters, multidrug-resistant efflux pumps, and the SCCmec element. Variations in transporter gene copies correlate with distinct resistance profiles in highly successful clones over the evolution of the pathogen's AMR. Together, these findings demonstrate the intricate complexity of the *S. aureus* sugar transportome and antimicrobial resistome, revealing high variability in presence and abundance across lineages. Overall, our study emphasizes the significance of taking into account multiple aspects of AMR in order to gain a deeper understanding of the genetic underpinnings of the evolution and dispersion of pathogenic bacteria and to develop more effective infection control strategies.

Keywords: Staphylococcus aureus, carbohydrate transporters, antimicrobial resistance, cross-genome comparison, efflux-pump.

Brazilian Symposium on Bioinformatics 2024

Harnessing Integrated Informatics and Molecular Simulation to Predict Antibody Epitopes on Viral Envelope Glycoproteins

Suyong Re[1] and Kenji Mizuguchi[1,2]

1 Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health, and Nutrition, Osaka, Japan.
2 Institute for Protein Research, Osaka University, Osaka, Japan.

An understanding of how an antiviral monoclonal antibody recognizes its target is important for the development of neutralizing antibodies and vaccines. Identification of epitopes, a part of antigen that antiviral antibody binds, is essential for this purpose, and several informatics tools have been developed for epitope prediction based on protein structure and sequence information. For example, BepiPred [1] is a sequence-based prediction method which uses a random forest algorithm trained on epitopes annotated from antibody-antigen protein structures. There is also a structure-based method, ElliPro [2], which uses antigenicity, solvent accessibility, and flexibility of protein structures. These methods are powerful and can identify epitopes in a short time. Given many viral proteins are extensively glycosylated, an obvious and critical limitation of these methods is that they do not take into account the effect of glycans. The glycans on the surface of a viral envelope protein potentially affect the antibody response and the interaction with host cell protein. However, the structural complexity inherent to glycans (heterogeneity and flexibility) hamper visualizing the overall structure of surface glycans using X-ray crystallography and Cryo-electron microscopy.

MD simulations provide atomistic structure information, incorporating dynamics, for complex biomolecules. Integrating Molecular Dynamics (MD) simulation with epitope prediction informatics tools provides a solution to overcome the limitation mentioned above [3,4]. The dynamical structure information of surface glycans is deduced from MD simulation, which is then combined with the results of informatics tools to construct a consensus score for predicting epitopes under the effect of glycans. Because glycans interact extensively on protein surfaces, sampling all of their possible conformations remains challenging. Here, in order to accurately characterize the conformational ensemble of surface glycans at affordable computational cost, we present an MD simulation using gREST (generalized Replica- Exchange with Solute Tempering) [5], which enhances conformational sampling by locally scaling the potential energy of interest (the “solute” region). We applied this method to a small-sized but densely glycosylated Lassa virus fully glycosylated envelope protein complex and compared the results with those obtained using conventional MD simulations. The model system was constructed using CHARMM-GUI Glycan Modeler [6]. gREST simulation was performed using GENESIS program package [7,8], where the surface glycans were selected as the solute region. Shortly, gREST simulations better characterized distinct glycan clusters on the protein surface and explored antibody-accessible regions. Integration of simulation results with immunoinformatic tools improved the accuracy of epitope prediction. This method can be extended to the development of molecular models of antigen−antibody interactions with details of explicit glycan involvement.

Keywords: *Molecular dynamics simulation, Enhanced sampling, Epitope prediction, Viral glycoprotein*

References

[1] Jespersen, M. C. et al. (2017). BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes, *Nucleic. Acids. Res.*, 45, W24-29.

[2] Ponomarenko, J. et al. (2008). ElliPro: a new structure-based tool for the prediction of antibody epitopes, *BMC Bioinf.*, 9, 514.

[3] Re, S. and Mizuguchi, K. (2021). Glycan Cluster Shielding and Antibody Epitopes on Lassa Virus Envelop Protein, *J. Phys. Chem. B*, 125, 2089-2097.

[4] Sikora, M. et al. (2021). Computational epitope map of SARS-CoV-2 spike protein, *PLoS. Comput. Biol.*, 17, e1008790.

[5] Kamiya, Motoshi, and Yuji Sugita. (2018). Flexible Selection of the Solute Region in Replica Exchange with Solute Tempering: Application to Protein-Folding Simulations, *J. Chem. Phys.*, 149, 072304–072304.

[6] Park, S. et al. (2019). CHARMM-GUI Glycan Modeler for modeling and simulation of carbohydrates and glycoconjugates, *Glycobiology*, 29, 320-331.

[7] Kobayashi, C.; Jung, J. et al. (2017). GENESIS 1.1: A Hybrid-Parallel Molecular Dynamics Simulator with Enhanced Sampling Algorithms on Multiple Computational Platforms, *J. Comput. Chem.*, 38, 2193-2206

[8] Jung, J.; Mori, T. et al. (2015). GENESIS: a hybrid-parallel and multi-scale molecular dynamics simulator with enhanced sampling algorithms for biomolecular and cellular simulations, *WIREs. Comput. Mol. Sci.*, 5, 310-323.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Identification of Genetic Alterations in Patients Who Developed Physical Fatigue as a Long COVID Condition

Felipe Ataides Mion[1], Aléxia Stefani Siqueira Zetum[1], Danielle Ribeiro Campos da Silva[1], Saulo Almeida Morellato[1], Vinícius do Prado Ventorim[1], Yasmin Moreto Guaitolini[1], Débora Dummer Meira[1], Iuri Drumond Louro[1]

1. *Universidade Federal do Espírito Santo (UFES)*

Introduction: Long COVID (LC) refers to persistent symptoms after infection with SARS-CoV-2, affecting millions of people. One of the most common sequelae is chronic fatigue syndrome, which occurs after viral infections and may be related to genetic factors that influence the inflammatory immune response. **Objective:** To identify genetic variants associated with predisposition to the development of post-COVID-19 physical fatigue. **Methodology:** A total of 277 participants were included, of whom 104 reported physical fatigue. All underwent a questionnaire to collect sociodemographic and clinical data, as well as blood collection for DNA extraction and sequencing. The analysis focused on genes related to cytokine signaling pathways, leukocyte recruitment, and endothelial injury markers. The variants were analyzed through partial least squares regression, with cross-validation and variable selection using R software version 4.3.3. **Results and Discussion:** Significant changes ($p < 0.05$) were found in the following genes: VWF (rs216325; OR=7.4792; 95% CI=1.5615-57.965); ADAMTS13 (rs2301614; OR=1.9656; 95% CI=1.0916-3.5728); HLA-B (rs200450928; OR=2.7742; 95% CI=1.1157-7.1936); and MICA (rs41553616; OR=3.3933; 95% CI=1.2429-9.8135). During the acute phase of COVID-19, the complement and coagulation systems can be activated and remain active in various tissues in patients with LC. Damage to endothelial cells results in the release of von Willebrand factor (VWF), which recruits platelets and promotes coagulation. The size of VWF is regulated by the enzyme ADAMTS13, responsible for its cleavage, preventing the formation of excessively large multimers that could promote thrombosis. The HLA-B gene presents peptides derived from foreign antigens to T cells, triggering an immune response to combat the invader. The literature indicates that patients with chronic fatigue show a higher VWF(ag)/ADAMTS13 ratio and downregulation of HLA-B. Thus, alterations in these genes may dysregulate their functions, favoring thrombus formation and impairing the recognition of pathogens by T cells. Moreover, the MICA gene, expressed in stressed cells and one of the ligands for the NKG2D receptor (natural killer group 2 member D), may undergo changes that lead to the release of soluble MICA during infection, allowing, along with the downregulation of HLA-B, the virus to evade the immune response and cause persistent hyperinflammation. Therefore, pathogenic genetic alterations may weaken the body's defenses and contribute to hyperinflammation, prolonging tissue damage and increasing the risk of complications. **Conclusion:** Genetic factors play crucial roles in the development and worsening of post-COVID-19 physical fatigue. The role of these genes in immune response and endothelial injury supports the hypothesis that virus-induced inflammation can overcome a weakened immune system and cause persistent damage. Thus, identifying these variants can aid in screening individuals at risk, enabling the development of more effective preventive and therapeutic treatments.

Keywords: ***Long COVID. Genetic. Myalgic Encephalomyelitis. Chronic Fatigue Syndrome***

References:

Castelli, E. C., de Castro, M. V., Naslavsky, M. S., Scliar, M. O., Silva, N. S. B., Andrade, H. S., Souza, A. S., Pereira, R. N., Castro, C. F. B., Mendes-Junior, C. T., Meyer, D., Nunes, K., Matos, L. R. B., Silva, M. V. R., Wang, J. Y. T., Esposito, J., Coria, V. R., Bortolin, R. H., Hirata, M. H., & Magawa, J. Y. (2021). MHC Variants Associated With Symptomatic Versus Asymptomatic SARS-CoV-2 Infection in Highly Exposed Individuals. *Frontiers in Immunology*, *12*. https://doi.org/10.3389/fimmu.2021.742881

Ely E.W., Brown, L. M., & Fineberg, H. V. (2024). Long Covid Defined. *New England Journal of Medicine*. https://doi.org/10.1056/nejmsb2408466

Gutiérrez-Bautista, J.F., Martínez-Chamorro, A., Rodríguez-Nicolás, A., Rosales-Castillo, A., Jiménez, P., Anderson, P., Miguel Ángel López-Ruz, López Nevot, & Ruíz-Cabello, F. (2022). Major Histocompatibility Complex Class I Chain-Related α (MICA) STR Polymorphisms in COVID-19 Patients. *International Journal of Molecular Sciences*, *23*(13), 6979–6979. https://doi.org/10.3390/ijms23136979

Peppercorn, K., Edgar, C. D., Kleffmann, T., & Tate, W. P. (2023). A pilot study on the immune cell proteome of long COVID patients shows changes to physiological pathways similar to those in myalgic encephalomyelitis/chronic fatigue syndrome. *Scientific Reports*, *13*(1), 22068. https://doi.org/10.1038/s41598-023-49402-9

Prasannan, N., Heightman, M., Hillman, T., Wall, E., Bell, R., Kessler, A., Neave, L., Doyle, A., Devaraj, A., Singh, D., Dehbi, H.-M., & Scully, M. (2022). Impaired exercise capacity in post–COVID-19 syndrome: the role of VWF-ADAMTS13 axis. *Blood Advances*, *6*(13), 4041–4048. https://doi.org/10.1182/bloodadvances.2021006944

Ruf, W. (2024). Immune damage in Long Covid. *Science*, *383*(6680), 262–263. https://doi.org/10.1126/science.adn1077

Ensiye Torki, Fahimeh Hoseininasab, Moradi, M., Sami, R., Mark, & Hamed Fouladseresht. (2024). The demographic, laboratory and genetic factors associated with long Covid-19 syndrome: a case–control study. *Clinical and Experimental Medicine*, *24*(1). https://doi.org/10.1007/s10238-023-01256-1

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Investigation of UDE-based approaches for cell cycle modeling

Ronaldo N. Sousa[1,2], Cristiano G.S. Campos[4], Ronaldo F. Hashimoto[1], Hugo A. Armelin[2,3], Marcelo S. Reis[4]

1. *Instituto de Matemática e Estatística – Universidade de São Paulo*
2. *CelTICS, Laboratório de Ciclo Celular – Instituto Butantan*
3. *Instituto de Química - Universidade de São Paulo*
4. *Recod lab, Instituto de Computação – Universidade Estadual de Campinas*

Cell signaling pathways regulate cellular activities, orchestrating processes such as growth, proliferation, and death through cascades of chemical reactions. Their deregulation is associated with diseases such as cancer [Hidalgo and Others 2018, Joo 2009]. In this context, dynamic modeling using ordinary differential equations (ODEs) can be employed to understand the behavior of individual proteins within these complex networks. However, these dynamic models capture only a portion of the processes occurring within the cell, assuming that the modeled proteins function in isolation. Some studies have been proposed to address the isolation assumption [Lee and Others 2020, Glass et al. 2021, Bangi and Others 2022, Santana and Others 2023]. One way to tackle the lack of isolation problem is by employing a hybrid model that integrates mechanistic and data-driven modeling, also known as universal differential equations (UDE) [Rackauckas et al. 2021]. Our previous work investigated the identifiability of such a model within the context of cell signaling pathways: we demonstrated, using toy models, that the hybrid approach was better than an ODE-based model [Sousa et al. 2023]. This ongoing work employs the hybrid model approach but with a real-world model instead of a toy model. It aims to investigate the performance of this approach when a data-driven component is used to capture the missing signals from a submodel of a real one, in this case, the cell cycle model [Iwamoto et al. 2011] with 54 proteins and 126 reactions. From this model, a submodel was created, compressing five proteins and ten reactions. A universal differential equation (UDE) was defined, where the neural network was employed to predict the missing signal from the ODE model from those five proteins. Initial results showed that the hybrid approach still got better than the ODE-based model, even though some of the protein predictions still need improvement. Thus, UDEs demonstrate the ability to capture missing signals. This ongoing work aims to improve the methodology for training these models and address the associated challenges. The goal is to advance hybrid modeling, making it a powerful tool in computational biology with implications for disease research and therapeutic development.

Keywords: *Scientific Machine Learning , First-principle Modeling , Universal Differential Equation , Inverse Problem , Cell Signaling Pathway*

References

Bangi, M. S. F. and Others (2022). Physics-informed neural networks for hybrid modeling of lab-scale batch fermentation for beta-carotene production using saccharomyces cerevisiae. *Chemical Engineering Research and Design*, 179:415–423.

Glass, D. S. et al. (2021). Nonlinear delay differential equations and their application to modeling biological network motifs. *Nature Communications*, 12(1):1788.

Hidalgo, M. R. and Others (2018). Models of cell signaling uncover molecular mechanisms of high-risk neuroblastoma and predict disease outcome. *Biology Direct*, 13(1):16.

Iwamoto, K., Hamada, H., Eguchi, Y., and Okamoto, M. (2011). Mathematical modeling of cell cycle regulation in response to dna damage: exploring mechanisms of cell-fate determination. *Biosystems*, 103(3):384–391.

Joo, J. D. (2009). The use of intra-cellular signaling pathways in anesthesiology and pain medicine field. *Korean Journal of Anesthesiology*, 57(3):277–283.

Lee, D. and Others (2020). Development of a hybrid model for a partially known intracellular signaling pathway through correction term estimation and neural network modeling. *PLOS Computational Biology*, 16(12):1–31.

Rackauckas, C. et al. (2021). Universal differential equations for scientific machine learning.

Santana, V. V. and Others (2023). Efficient hybrid modeling and sorption model discovery for non-linear advection-diffusion-sorption systems: A systematic scientific machine learning approach. *ArXiv Preprint ArXiv:2303.13555*.

Sousa, R. N., Campos, C. G. S., Wang, W., Hashimoto, R. F., Armelin, H. A., and Reis, M. S. (2023). Exploring identifiability in hybrid models of cell signaling pathways. In Reis, M. S. and de Melo-Minardi, R. C., editors, *Advances in Bioinformatics and Computational Biology*, pages 148–159, Cham. Springer Nature Switzerland.

Brazilian Symposium on Bioinformatics 2024

Metagenomics and Bioinformatic tools in Agricultural Microbiome

Inaiá Ramos Aguiar[1,2], Rafael Silva Rocha[3], Beatriz Bergamo[3], Cecilia Cerliani[4], María-Eugenia Guazzaroni[1]

1. Department of Biology, Faculty of Philosophy, Sciences and Letters of Ribeirão Preto, University of São Paulo, São Paulo, SP, Brazil
2. Department of Cell and Molecular Biology, Faculty of Medicine of Ribeirão Preto, University of São Paulo, São Paulo, SP, Brazil
3. ByMyCell Inova Simples. Av. Dra. Nadir Aguiar, 1805 – Supera Parque, Ribeirão Preto, SP, Brazil.
4. Faculty of Agronomy and Veterinary Medicine, National University of Río Cuarto.

Shotgun metagenomic sequencing allows for the detailed analysis of microbial communities directly from environmental samples, providing a comprehensive view of both microbial diversity and the present metabolic functions. The assembly and study of reconstructed genomes from metagenome-assembled genomes (MAGs) are crucial for characterizing uncultivable microorganisms, elucidating their ecological roles, and understanding their interactions within complex systems. In agricultural soils, MAG analysis is particularly relevant due to the essential role that microorganisms play in nutrient cycling, soil health, and plant growth promotion, directly impacting crop productivity and sustainability (Setubal, 2021). BASALT, a new bioinformatics tool (QIU *et al*., 2024), was used for metagenomic bin refinement. It offers an innovative approach to recovering microbial genomes from complex environmental samples. BASALT is expected to enhance genomic reconstruction efficiency, with lower contamination and higher completeness, outperforming traditional tools. Some bioinformatics tools (as MetaPhlAn e Kraken) were compared, aiming to address biologically relevant agricultural questions, such as the influence of the microbiome on soil fertility and crop performance.

This study focused on soil samples from three Argentine provinces (Santa Fé, Buenos Aires, and Córdoba) and from different regions in Brazil, covering crops such as soybean, corn, sugarcane, citrus and common bean. Preliminary results from MetaWrap bin refinement showed significant variations in microbial communities, influenced by both crop type and geographic location. Nitrogen-fixing species predominated in soil samples cultivated with soybeans, while other regions exhibited greater diversity of organisms involved in nitrogen cycling and organic matter decomposition.

These results were compared with those obtained after BASALT refinement, and an expressive enhancement in the number of generated bins was evident, providing more accurate data and enabling new insights into agricultural microbiology and bioinformatics.

Keywords: *Metagenomic sequencing, BASALT, Microbiome, Genomic binning*

References

Qiu, Z., Yuan, L., Lian, C.-A., Lin, B., Chen, J., Mu, R., Qiao, X., Zhang, L., Xu, Z., Fan, L., Zhang, Y., Wang, S., Li, J., Cao, H., Li, B., Chen, B., Song, C., Liu, Y., Shi, L., Tian, Y., Ni, J., Zhang, T., Zhou, J., Zhuang, W.-Q. and Yu, K. (2024). BASALT refines binning from metagenomic data and increases resolution of genome-resolved metagenomic analysis, *Nature Communications*, 15(1), p. 2179.

Setubal, J. C. (2021) Metagenome-assembled genomes: concepts, analogies, and challenges. *Biophysical reviews*, 13(6), p. 905-909.

Brazilian Symposium on Bioinformatics 2024

Multi-omics systems biology approach identifies novel signature genes for neuropsychiatric disorders

D. M. Gysi [1,*], K. Nowick [2,*]

[1]Department of Statistics, Federal University of Paraná, Curitiba, Brazil.
[2]Human Biology Group, Institute for Biology, Department of Biology, Chemistry, Pharmacy, Free University of Berlin, Konigin-Luise-Str. 1-3, D-14195 Berlin, Germany.
[*]D.M.G. deisy.gysi@ufpr.br [*]K.N. katja.nowick@fu-berlin.de

Introduction

Cognition involves a range of mental processes crucial for human functioning, with the prefrontal cortex (PFC) playing a central role. Several genes crucial for cognitive functions are implicated in neuropsychiatric disorders, i.e., autism spectrum disorder (ASD), Bipolar Disorder (BD), Major Depressive Disorder (MDD), Schizophrenia (SCZ), Alzheimer's Disease (AD) and Parkinson Disease (PD). The significant genetic overlap among disorders suggest a strong genetic basis. Yet, how these overlapping genes interact in individuals affected by neuropsychiatric remains largely unknown. Transcription factors (TFs), which play a role in brain development and gene regulatory processes, has shown enrichment in this context.

Cognitive disorders are intricate and polygenic. Therefore, network analyses are essential for unraveling the genetic interactions involved. In this study, we focus on the TFs' impact and their coexpression patterns on disease networks, aiming to pinpoint TFs specific to selected neuropsychiatric disorders (signature TFs). To achieve this, we conducted a thorough analysis of the genetic and regulatory overlap among mental disorders. We first assessed disease similarity by calculating gene-disease overlap and network separation for each disease pair based genetic association. Next, we construct their coexpression networks, employing rigorous statistical filtering and consensus network methods to ensure reliability. By comparing these networks, we identified biomarkers for each disorder.

Results

We mapped out 50 neuropsychiatric disorders to their associated genes, creating a bipartite network of disorders and their associated genes. We next assessed genetic disease similarity using Jaccard Similarity and disease separation based on protein-protein and their non-coding mediated interactions (PPI & NCI[1]). Diseases were considered connected if they had a significant overlap in their associated genes and if their networks were in the same topological region, revealing a strong genetic and network overlap among neuropsychiatric disorders. Additionally, we found diseases clusters in the same manner as the DSM-5 main chapters. We focus on the psychotic and neurodegenerative cluster, observing a high genetic overlap within the cluster. Notably, MDD shared a significant portion of its genes with AD, PD, SCZ, and BIP, suggesting a central role for MDD in this network of disorders.

To investigate the regulatory overlap among mental disorders in the psychotic and neurodegenerative cluster, we collected transcriptomic data from patients' PFC across 26 studies, covering six diseases and healthy controls. We ensure data integrity by processing each data independently. We identify next consistent gene relationships within each disease and disease-specific relationships, leading to the identification of signature genes.

Given the significant impact of non-coding GWAS signals on gene expression, and the importance of TFs to mental disorders, we focused our network analyses on approximately 3000 TFs. We constructed 61 independent TF-TF networks, one for each condition per study,

using the weighted topological overlap (wTO[2,3]). We then consolidated the most reliable links into six Consensus Networks (CNs[2]), combining TF-wTO networks from different datasets.
We next use the CNs to identify specific links and nodes for each mental disorder compared to healthy controls, employing coexpression Differential Network Analysis (CoDiNA[4,5]). CoDiNA integrates TF-wTO networks in a comparative manner, classifying links and nodes based on their specificity within each network. Specifically, α represents conserved functional relationships, β indicates functional relationships that have changed under certain conditions, and ɣ represents interactions specific to one network. We next focus on understanding the similarities, differences, and specificities of TF co-regulation patterns between controls and neurodegenerative disorders (AD and PD), as well as psychiatric disorders (ASD, BD, MDD, and SCZ). This analysis provides insights into the unique regulatory mechanisms underlying each disorder, revealing key biomarkers and therapeutic targets.
To understand how gene coexpression networks change in the neurodegenerative diseases, we compared TF-wTO networks of AD and PD with healthy controls using CoDiNA. We found 301 TFs specific to AD, 58 specific to PD, and 16 shared between both. The AD-specific TFs are enriched for cell cycle, differentiation, and brain development. PD-specific TFs are involved in myelination, mitosis regulation, and stress response. For psychiatric disorders, only 31 TFs are common across all networks. However, several specific TFs are identified ASD has 493 specific TFs, MDD has 248, while SCZ and BIP have 21 and 2, respectively. TFs enriched in SCZ are involved in dendrite morphogenesis, and synaptic development.

Conclusion

We investigated the genetic overlap of fifty mental disorders and their associated genes, finding that diseases cluster into modules which largely agree with DSM-5. For one disease module, which contains AD, PD, SCZ, MDD, ASD and BIP, we analyzed similarities and differences in their co-expression TF networks in the PFC. We found less overlap between the diseases at this regulatory level than at the genetic level, suggesting that more signature genes, here TFs with disease specific interactions, can be found by the analysis of coexpression patterns. These coexpression networks overlap very well with known PPI lending additional support for the validity of our results. Importantly, known gene functions and biological processes enriched among signature TFs and their disease specifically co-expressed genes agree very well with the known molecular underpinnings of the diseases.

Keywords: *Biomarker Discovery, Network Medicine, Transcriptomics*

References

1. Morselli Gysi, D. & Barabási, A.-L. Noncoding RNAs improve the predictive power of network medicine. *Proc. Natl. Acad. Sci.* **120**, e2301342120 (2023).
2. Gysi, D. M., Voigt, A., Fragoso, T. de M., Almaas, E. & Nowick, K. wTO: an R package for computing weighted topological overlap and a consensus network with integrated visualization tool. *BMC Bioinformatics* **19**, 392 (2018).
3. Nowick, K., Gernat, T., Almaas, E. & Stubbs, L. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proc. Natl. Acad. Sci.* **106**, 22358–22363 (2009).
4. Gysi, D. M., de Miranda Fragoso, T., Almaas, E. & Nowick, K. CoDiNA: Co-Expression Differential Network Analysis. *Httpscranr-Proj.* (2018).
5. Gysi, D. M. *et al.* Whole transcriptomic network analysis using Co-expression Differential Network Analysis (CoDiNA). *PLoS ONE* **15**, e0240523 (2020).

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Non coding variants near the *NOTCH1* gene are associated with frailty criteria in Brazilians older adults

Rafael Rezende[1], Michel Naslavsky[2], Izadora Silveira Fernandes[3], Maria Rita Passos-Bueno[2], Yeda Duarte[4], Mayana Zatz[2], Estevão[5] Carlos Silva Barcelos[3], Flavia Imbroisi Valle Errera[3]

1. Federal Institute of Espírito Santo (IFES)
2. Biosciences Institute, University of São Paulo (USP)
3. Federal University of Espírito Santo (UFES)
4. School of Nursing, University of São Paulo (USP)
5. Università degli Studi di Perugia (UNIPG)

INTRODUCTION: Frailty is an aging-related syndrome characterized by increased vulnerability and mortality, resulting from the loss of functional biological reserves. The genetics of the syndrome is not well understood, and its heritability varies between 19-46% in different studies (Sathyan; Verghese, 2020). *NOTCH1*, a gene that encodes a NOTCH receptor 1, plays a crucial role in the embryonic development, cell fate decision and maintenance of stem cells pool and appears to be important in the aging process. Although a recent study identified higher levels of NOTCH1 in plasma as a protective factor against frailty in middle-aged individuals (Liu et al., 2024), there are still no studies that have focused on the gene's involvement with the condition. **OBJECTIVE:** Determine whether single nucleotide polymorphisms (SNPs) in the *NOTCH1* gene are associated with the frailty phenotype. **METHODS:** Older adults from the Health, Well-being, and Aging (SABE) cohort in São Paulo were stratified into frail and robust groups according to the frailty classification based on the Fried Frailty Criteria (FF). Data were obtained from a standardized questionnaire and whole-genome sequencing to identify genetic variants. The variants and frequencies were deposited in ABraOM - Brazilian Online Archive of Mutations. We filtered SNPs located at the start and end positions of *NOTCH1* and within 50Kb on both sides. We excluded Indels and unannotated SNPs. We selected SNPs with a minor allele frequency (MAF) of 0.01 or greater, Hardy-Weinberg equilibrium (p>0.05) and $r^2 \geq 0.8$. Of the total 976 elderly individuals, 495 were excluded for being considered pre-frail. A total of 481 elderly individuals were included, with a median age of 69.12 years, of which 320 were women (66.5%), and 103 (21.14%) were frail. Statistical analyses were performed in R 4.4.1, and SNP-based association analysis was conducted using the R package SNPassoc. A total of 242 tag SNPs were analyzed. **RESULTS:** The intergenic SNPs located upstream *NOTCH1* rs56235385 and rs7021438 were associated respectively to FF and slowness, and weakness. Additionally, we found the SNP rs1127152, located at 3'UTR of *SEC16A*, in

association with low physical activity. The GTG haplotype was associated with lower risk of being frail (OR=0.7, 0.5-0.98) (p=0.039). *In silico* functional analysis predicted the regulatory potential of these SNPs in rSNPBase, as well as the presence of enhancer markers in muscle and brain in RegulomeDB, and eQTL data for rs7021438 and rs1127152, which also presented sQTL in GTEx portal. Alteration of transcription factor binding site (TFBS) analysis performed in FABIAN-Variant suggests that the alternative allele of the SNP rs56235385 increases the binding affinity of the region with SOD1, a transcription factor (TF) known to neutralize free superoxide radicals. **CONCLUSION:** The variants rs56235385, rs7021438 and rs1127152, located near the *NOTCH1* gene were associated with FF. The GTG haplotype appears to be protective against frailty. The results suggest that the regulatory region of NOTCH1 may be important for the development of frailty.

Keywords: frailty; aging; NOTCH

References

Sathyan, S. and Verghese, J. (2020). Genetics of frailty: A longevity perspective. In *Translational Research*, pages 83-96.

Liu, F., Schrack, J.A., Walston, J. (2024). Mid-life plasma proteins associated with late-life prefrailty and frailty: a proteomic analysis. In *GeroScience*, pages 5247–5265.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Predicting aggregation region in proteins with machine learning based on tertiary structure: web platform

Alexandro Tadeu Mathias de Souza[1], Carlos Alves Moreira [1], Ana Ligia Barbour Scott

[1] Computational Biology and Biophysics Laboratory – Federal University of ABC

P.O. Box 09280-560 – Santo Andre´ – SP – Brazil

Protein aggregation is a relevant biological problems. There are several tools on literature to predict protein aggregation, some ones are: i) **NET-Cssp**: uses amino acid sequences to predict secondary protein structures, helping to identify regions that could contribute to aggre- gate formation ii)Betascan that estimates the propensity of a sequence to form specific secondary structures, particularly β-sheets which are often associated with protein aggre- gation ii) **Tango**: Simulates mutations and environmental conditions to predict aggregate formation, while others, like Pasta, assess the propensity to form β-sheets and the over- all aggregation tendency, providing a comprehensive view of the likelihood of aggregate structure formation; iv) **Aggrescan** that a concentrates on calculating the average propen- sity in the vicinity of amino acids, providing insight into how the environment of residues might have influenced aggregation; v) **Zyggregator** is Similar to Aggrescan, it calculates the average propensity in the vicinity of amino acids, helping to identify regions with a higher likelihood of forming aggregates.[Navarro and Ventura 2022]. Our group have developed a tool called **MAGRE-I** for such predictions, based on the machine learning and sliding window techniques. We have applied the Support Vector Machine algorithm to generate classification models. We utilized information of primary structure - protein sequence - from the Amyloid Data [Moreira et al. 2019]. In this project, we are developing a web application to integrate the functionalities of MAGRE I and MAGRE II. This application will allow users to evaluate aggregation tendencies using amino acid sequences and/or tertiary structures.

Keywords: *Protein Aggregation, Machine Learning and WenServer*

References

Moreira, C. A., Philot, E. A., Lima, A. N., and Scott, A. L. (2019). Predicting regions prone to protein aggregation based on svm algorithm. *Applied Mathematics and Com- putation*, 359:502–511.

Navarro, S. and Ventura, S. (2022). Computational methods to predict protein aggrega- tion. *Current Opinion in Structural Biology*, 73:102343

Brazilian Symposium on Bioinformatics 2024

PREDICTIVE MODELING OF POST-COVID-19 HAIR LOSS: INSIGHTS FROM MACHINE LEARNING AND LOGISTIC REGRESSION

Basílio LA[1], Silva DRC[1], Zetum ASS[1], Ventorim VP[1], Mion FA[1], Rosa HP[1], Batista LS[1], Pegos AR[1], Alvarenga FDS[1], Reis RS[1], Barbosa KRM[1], Morais LC[1], Altoé LSC[1], Guaitolini YM[1], Giacinti GM1, Alves LNR[1], Morelatto SA[1], Meira DD[1], Louro ID[1],

1 Universidade Federal do Espírito Santo (UFES)

Objective:
This study aims to investigate post-COVID-19 hair loss by leveraging machine learning techniques to identify key factors contributing to its occurrence and propose prevention strategies[1]. Multiple algorithms will be evaluated to determine the best-performing model for predicting and understanding the primary drivers of hair loss following COVID-19 infection.

Methods:
A total of 206 patients were selected (Mild COVID: 55, Moderate: 60, Severe: 91) between November 2020 and December 2021, confirmed by RT-PCR, aged over 18 years, without acute symptoms for at least 30 days post-infection, and unvaccinated (not having received a two-dose series). All participants signed informed consent forms, completed an online questionnaire, and provided biological samples (CAAE: 37094020.6.0000.5060). Between March and July 2023, participants were contacted again to complete a follow-up questionnaire on Long COVID. A total of 19 parameters (sociodemographic, comorbidities, and biomarkers) were assessed. The machine learning tools used included Support Vector Machine (SVM), Logistic Regression with Elastic-Net (EN), Random Forest (RF), Extreme Gradient Boosting (XGB), and Light Gradient Boosting Machine (LGBM), utilizing the SMOTENC resampling technique Models were evaluated based on AUC, Accuracy, F1-score, Recall (R), and Precision (PR)[1,2]. Risk prediction was performed using Stepwise Logistic Regression (SLR) and Chi-square test or Fisher's exact test. The sequel analyzed was hair loss N=66.

Results: Machine Learning: Best results were observed with RF (F1=0.81; AUC: 0.78;), XGB (F1: 0.79; AUC: 0.79), EN (F1: 0.79; AUC: 0.79), SVM (F1: 0.79; AUC: 0.78), and LGBM (F1: 0.77; AUC: 0.79). Important variables for RF: gender. **Stepwise Logistic Regression (P<0.05):** Severe clinical spectrum (CS) (OR: 10.09, CI: 3.43-33.10), female gender (OR: 22.87, CI: 9.26-64.39). **Chi-square test or Fisher's exact test (P<0.05)**: female gender P < 0.001 and Von Willebrand Factor (VWF) (%) P=0.009.

Discussion: The RF model performed best, showing robustness for complex datasets with multiple variables. Logistic Regression highlighted female gender as the most significant predictor, which aligns with known associations between hormonal factors and hair loss. Moreover, VWF was a critical factor, potentially linking vascular injury to post-COVID-19 hair loss. This suggests that endothelial damage and microvascular dysfunction may play essential roles in the condition. Further studies are required to explore the precise role of VWF and other biomarkers in post-infection hair loss. These findings indicate that interventions targeting vascular health, nutrient replenishment, and environmental control could improve hair loss outcomes.

Conclusion: Selecting the most appropriate algorithm depends on the dataset characteristics, Nutrient replenishment and controlling environmental factors. However, further studies will be necessary to fully understand the long-term effects and the most effective interventions for hair loss management post-infection.

Keywords: Machine learning, Long COVID, Logistic regression, Hair loss.

References

1. PUROHIT, Manoj; MADIRAJU, Praveen. (2023). Predicting Mental Health Disorders Post Long COVID Diagnosis Using Advanced Machine Learning Techniques. In: 2023 IEEE International Conference on Big Data (BigData). **IEEE**. p. 4954-4962.
2. CHAWLA, Nitesh V. et al. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, v. 16, p. 321-357.
3. BREIMAN, Leo. Random forests. (2001). Machine learning, v. 45, p. 5-32.
4. Rahmati, M., Udeh, R., Yon, D. K., Lee, S. W., Dolja-Gore, X., McEVoy, M., Kenna, T., Jacob, L., López Sánchez, G. F., Koyanagi, A., Shin, J. I., & Smith, L. (2023). A systematic review and meta-analysis of long-term sequelae of COVID-19 2-year after SARS-CoV-2 infection: A call to action for neurological, physical, and psychological sciences. Journal of medical virology, 95(6).

Brazilian Symposium on Bioinformatics 2024

SP crime: A Python package for merging São Paulo criminal and medical data

Maria Laura Gabriel-Kuniyoshi[1], David Corrêa Martins-Junior[2], Sérgio Nery Simões[3], Helena Brentani[1]
[1]Instituto de Psiquiatria – Faculdade de Medicina da Universidade de São Paulo (FMUSP) São Paulo – SP – Brazil
[2]Centro de Matemática Computação e Cognição – Universidade Federal do ABC (UFABC) Santo André – SP – Brazil
[3]Instituto Federal do Espírito Santo (IFES) – Serra, ES – Brazil

Since a violent environment may relate to negative health outcomes, it is important to integrate criminal and health data. In São Paulo (SP) state, the Public Security Secretariat (SSP) publicly provides data, but it is unprocessed tables or through a purely visual interface. Scientists and journalists who analysed such data deliver the results, but not the code or any processed table-formatted database. Thus, healthcare scientists from SP urge for an open-source database with usable formats and updated information. Here, we present SPCrime, a Python package that processes SP criminal data and was designed to integrate it into medical databases. The tool processes raw data of the SP-SSP — which consists of individual records for each police occurrence — and calculates the frequency of 21 felony categories and three macro-categories (robbery/theft, lethal intentional violent crimes, and non-lethal intentional violent crimes). We normalized the rates by inhabitants of 645 municipalities of SP state and 96 districts of SP city. Given a dataset containing the postal codes of each patient, SPCrime returns the crime rates in their place of residence. The tool can be used along the Pandas library. In a test on a standard computer, SPCrime processed one year worth of SP-SSP data in 48 minutes. Then, it took <1 second to map the postal code of 307 patients and add the crime rates to the dataset. We envision that this tool will help healthcare scientists and that it can be extended to other areas such as criminology. The SPCrime package, source code, and documentation are freely available at https://github.com/marialgk/SPCrime.

Keywords: *Crime, Brazil, Medical Informatics, Data Management, Secondary Data Analysis.*

Brazilian Symposium on Bioinformatics 2024

The Role of Indel Variants in COVID-19: Unveiling Frequency Patterns and Potential Clinical Significance

Aléxia Stefani Siqueira Zetum[1], Vinícius do Prado Ventorim[1], Felipe Ataídes Míon[1], Karen Ruth Michio Barbosa[1], Lívia César Morais[1], Débora Dummer Meira[1], Danielle Ribeiro Campos da Silva[1], Iuri Drumond Louro[1]

1. Universidade Federal do Espírito Santo (UFES)

Introduction: Insertions and deletions (InDels) remain a largely unexplored frontier in structural biology. While frequently associated with various pathological phenotypes, studies on InDels and their structural impacts are still limited, particularly in the context of COVID-19 (Zheng Zhang et al. 2016). **Objective:** To map the percentage of inDels in a COVID-19 cohort. **Methodology:** A total of 300 individuals were sequenced (CAAE 37094020.6.0000.5060). The initial filtering of variants was performed using VCFtools (v0.1.16) (Danecek et al. 2011), isolating only inDels from the VCF file. Subsequently, a second filter was applied with bcftools (v1.9) to ensure variant quality (QUAL > 30 and DP > 10). The filtered variants were converted to ANNOVAR format and annotated using databases like refGene, enabling the identification of inDels and their potential functional consequences. The reference genome used was HG38. **Results:** A total of 27,974 inDels were identified. Of these, 24.82% occurred in exonic regions, with 21 inDels (0.08%) classified as exonic/splicing. Most exonic inDels were nonframeshift deletions (31.23%) and nonframeshift insertions (20.94%). Frameshift deletions accounted for 21.73% and frameshift insertions for 12.22%. Additionally, stopgain inDels (2.29%) and start loss inDels (0.28%) were identified. In ncRNA (exonic) regions, 799 inDels (2.86%) were found. In intronic regions, 56.53% of inDels were observed, with 5.37% and 2.78% in UTR3 and UTR5, respectively. Non-coding RNA (ncRNA) intronic regions represented 827 inDels (2.96%). Splicing regions contained 253 inDels (0.90%). Intergenic regions accounted for 720 inDels (2.57%), while 225 inDels (0.80%) and 70 inDels (0.25%) were found in upstream and downstream regions, respectively. An additional 12 inDels (0.04%) were identified in both upstream and downstream regions. **Discussion:** InDels may have a more severe impact than point mutations, particularly when they occur in regions essential for maintaining protein stability (Jilani et al. 2022). For example, frameshift inDels can severely disrupt protein function, potentially leading to pathogenic outcomes, while those in non-coding regions can affect gene regulation and splicing (Zheng Zhang et al. 2016). This analysis is especially relevant for understanding infectious diseases like COVID-19, where variants affecting key immune system genes or inflammatory response may influence the severity of infection. **Conclusion:** Studying inDels is important due to their impact on gene function and regulation across various diseases. Frameshift and stopgain inDels can lead to truncated or dysfunctional proteins. Future steps will involve comparing these results with control group data to determine whether InDel distribution varies in individuals with severe COVID-19, potentially clarifying the clinical roles of these elements in diverse infection outcomes.

Keywords: COVID-19, Indels, Genetic, Bioinformatics.

References

Yue, J., Andrei, L., Turinsky, M., and Brudno, M. (2015). The missing indels: an estimate of indel variation in a human genome and analysis of factors that impede detection, Nucleic Acids Research, 43(15), 7217–7228.

Jilani, M., Turcan, A., Haspel, N., and Jagodzinski, F. (2022). Elucidating the Structural Impacts of Protein InDels, Biomolecules, 12(10), 1435.

Zhang, Z., Huang, J., Wang, Z., Wang, L., and Gao, P. (2011). Impact of Indels on the Flanking Regions in Structural Domains, Molecular Biology and Evolution, 28(1), 291–301.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., and Durbin, R. (2011). The variant call format and VCFtools, Bioinformatics (Oxford, England), 27(15), 2156–2158.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., and Li, H. (2021). Twelve years of SAMtools and BCFtools, GigaScience, 10(2).

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Topology-based pan-cancer analysis of DLK1-DIO3-derived microRNA roles

Mylena Roberta dos Santos[1], Murilo Vieira Geraldo[2], André Santanchè[1]

1. Laboratório de Sistemas de Informação, Instituto de Computação, Universidade Estadual de Campinas (UNICAMP)
2. Departamento de Biologia Estrutural e Funcional, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP)

Cancer is recognized as a systems biology disease, requiring an integrated view for its study, such as that provided by network medicine [Hornberg et al. 2006, Laubenbacher et al. 2009]. By applying network science techniques, network medicine has significantly contributed to advances in cancer research, particularly by comparing gene expression profiles across different cancer types [Barabási et al. 2011]. Since the early 2010s, a comparative approach known as pan-cancer analysis has gained prominence, with network medicine analytical tools playing a pivotal role [Cancer Genome Atlas Research Network et al. 2013]. Additionally, microRNAs are crucial for cancer biology, acting as regulators of gene expression and establishing themselves as important cancer biomarkers and therapeutic targets [Hayes et al. 2014]. Our study focuses on the DLK1-DIO3 genomic region, a source of multiple microRNAs identified as tumor suppressors in distinct cancer types [Alves et al. 2024]. To our knowledge, no pan-cancer analysis grounded in network medicine has specifically targeted a genomic region before.
However, there are open challenges associated with using network-based approaches in pan-cancer research. In our proposal, we argue that a key challenge lies in applying topological network analysis in this context, especially when different cancer types are compared. To address this gap, we propose a topology-based axis to study and compare several cancer networks systematically. To uncover the recurrent topological patterns involving DLK1-DIO3-derived microRNAs, our comparative axis leverages a uniform application of topological metrics and structure detection strategies in these networks. We are using transcriptome profiling data from The Cancer Genome Atlas projects to model RNA-RNA interaction networks. Our topology-based comparative approach will elevate pan-cancer analysis to a holistic level, taking advantage of the complex system properties inherent to cancer. Instead of focusing on fine-grained interactions, we will compare these cancer networks based on their global organizational and structural characteristics. For instance, we will identify global network hubs, uncover the most recurrent ones, and determine their functional roles across cancers.

Keywords: *cancer, DLK1-DIO3, microRNA, network medicine, network science*

References

Alves, L. F., Marson, L. A., Sielski, M. S., Vicente, C. P., Kimura, E. T., and Geraldo, M. V. (2024). DLK1-DIO3 region as a source of tumor suppressor miRNAs in papillary thyroid carcinoma. *Translational Oncology*, 46, 101849.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews. Genetics*, *12*(1), 56–68.

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, *45*(10), 1113–1120.

Hayes, J., Peruzzi, P. P., and Lawler, S. (2014). MicroRNAs in cancer: biomarkers, functions and therapy. *Trends in Molecular Medicine*, *20*(8), 460–469.

Hornberg, J. J., Bruggeman, F. J., Westerhoff, H. v., and Lankelma, J. (2006). Cancer: A Systems Biology disease. *Biosystems*, *83*(2–3), 81–90.

Laubenbacher, R., Hower, V., Jarrah, A., Torti, S. v., Shulaev, V., Mendes, P., Torti, F. M., and Akman, S. (2009). A systems biology view of cancer. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, *1796*(2), 129–139.

Brazilian Symposium on Bioinformatics 2024

Transcriptomic analysis of the aged female omentum. Insights on metastatic invasion in a mouse model of ovarian cancer

Lorena Orellana [1], Carlos Chacon [1] and Ulises Urzúa [1]

1. Laboratorio de Genómica Aplicada, Departamento de Oncología Básico-Clínica, Facultad de Medicina, Universidad de Chile, Independencia 1027, Santiago, Chile

The development of genome-wide molecular methods in the last two decades has greatly improved our understanding of biological processes relevant to human health and the alterations that drive cells and tissues to a disease state. Our particular focus is the study of the molecular basis of aging and reproductive history as major risk factors of ovarian cancer (OC). As with several other cancers, the incidence and mortality of sporadic OC increases with aging -a non-modifiable risk factor- reaching a peak during the post-menopause. OC progression usually involves hemorrhagic ascites with intraperitoneal dissemination where a common metastatic niche is the greater omentum. In this work, we performed a transcriptomic study of the aged female mouse omentum aiming to identify dysregulated genes that might promote ovarian cancer metastatic invasion. Over 5-hundred differentially expressed genes were detected between aged and young omenta. Distinct genesets involved in cytokine response and neutrophil degranulation were up- and down-regulated with age. Genes involved in PPAR and insulin signaling, lipid metabolism and glucose homeostasis were up-regulated while genes related to collagen catabolism and structure, ECM remodeling, cell proliferation and ATP synthesis were down-regulated in the aged omentum. Dysregulated gene expression suggested age-dependent changes in the diverse cell types comprising the omentum, particularly macrophages, fibroblasts, endothelial and T-cells. GEO datamining uncovered 22 out of 42 of the lipid and glucose metabolism genes also upregulated in studies related to the OC response to azacitidine and talazoparib, olaparib resistance and the effect of PARP inhibitors on growth of BRCA2-deficient cells and tumors. Based on available gene-function knowledge, we conclude that the aged omentum develops a fibrotic, chronically immune-activated state supported by fatty acid catabolism with signs of insulin resistance and dysregulated glucose utilization. In this scenario, immunosuppression and chemotherapy resistance would be enhanced thus facilitating the invasion of metastatic OC cells.

Keywords: *Transcriptome, aging, omentum, mouse model, ovarian cancer*

References

Bella, Ángela et al. “Mouse Models of Peritoneal Carcinomatosis to Develop Clinical Applications.” Cancers vol. 13,5 963. 25 Feb. 2021, doi:10.3390/cancers13050963

Liu, Mingyong et al. “Specialized immune responses in the peritoneal cavity and omentum.” Journal of leukocyte biology vol. 109,4 (2021): 717-729. doi:10.1002/JLB.5MIR0720-271RR

Urzua, Ulises et al. “Parity History Determines a Systemic Inflammatory Response to Spread of Ovarian Cancer in Naturally Aged Mice.” Aging and disease vol. 8,5 546-557. 1 Oct. 2017, doi:10.14336/AD.2017.0110

Chacón, Carlos et al. “Transcriptomic Analysis of the Aged Nulliparous Mouse Ovary Suggests a Stress State That Promotes Pro-Inflammatory Lipid Signaling and Epithelial Cell Enrichment.” International journal of molecular sciences vol. 25,1 513. 30 Dec. 2023

Harper, Elizabeth I et al. “With Great Age Comes Great Metastatic Ability: Ovarian Cancer and the Appeal of the Aging Peritoneal Microenvironment.” Cancers vol. 10,7 230. 10 Jul. 2018.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Tumor-Regeneration Interplay: Systems Biology and New Models in Comparative Study with Therapeutic Insights

Matheus Correia Casotti[1], Bianca Paulino Campanharo[1], Débora Gonçalves Barbosa[1], Giulia Maria Giacinti[1], Isabele Pagani Pavan[1], Karen Ruth Michio Barbosa[1], Iúri Drumond Louro[1] and Débora Dummer Meira[1]

1. Systems and Computational Biology Group (SCBG), Center for Human and Molecular Genetics (NGHM), Federal University of Espírito Santo (UFES)

Metazoans, as multicellular organisms, face the constant challenge of cellular loss due to homeostatic renewal, injuries, or diseases. This involves cooperation between cells, which is guided by morphogenetic fields—chemical and physical patterns that direct growth, positioning, and shape. In this context, cancer can emerge when the organism fails to maintain order, also co-opting characteristics of biological regeneration, such as compensatory apoptosis and others. This abstract aimed to answer the following question: "How can the interactions between cancer and regeneration from a holistic perspective, through systems biology, provide valuable insights for using tumors in diverse treatments, demanding comparative methodologies?" To address this, protein interaction networks were constructed, according to the method standardized by our research group, using genes/proteins from the Regeneration Roadmap and REGene databases, in association with cancer-related genesets from Monarch (MONDO:0004992) and Disease (DOID:162). In summary, three networks were obtained—two for cancer (Monarch and Disease) and one for regeneration—and were compared, highlighting common genes/proteins with similar colors. A network with an mean degree of 19 was obtained, agreeing with mathematical parameters of graph and biological network theory. Functionally, pathway enrichment associated the regeneration-cancer relationship with processes/mechanisms linked to cellular and molecular structures (e.g., HuR binds and stabilizes mRNA, chromosome, membrane raft, cell cycle, DNA-binding, signal transduction), signaling pathways (e.g., MAPK, Wnt, TGF-beta receptor, PI3K-Akt, Notch, G protein-coupled serotonin receptor, SMAD binding and retinoic acid binding), development and regeneration (e.g., cell population proliferation, muscle organ development, axon guidance, neuron projection, and response to mechanical stimulus), metabolism and diseases (e.g., purine metabolism, peptide hormone metabolism, drug metabolism, chemical carcinogenesis, osteoblast), and other functions (e.g., apoptotic process and regulation, SUMOylation of DNA replication proteins, and chondroitin sulfate binding). Based on the results obtained, there is a shared mechanistic process in which there are cooperative and interactive arrangements permeating biochemical, biophysical, and biological regulations under an entropic trend inducing resilience capable of directing the dynamics of shared states between cancer and regeneration. The superexpression of the gamma retinoic acid receptor, which influences cell proliferation and anti-apoptosis. It also emphasizes the integration of cell cycle control with specific genes (FOS, JUNB, WNT4) and cellular senescence, regulated by epigenetic factors, tracing senescence as an intermediate state. Key signaling pathways like WNT, Notch, MAPK/ERK, and PI3K-Akt, along with hormone modulation, contribute to adaptive responses such as oxidative stress. Plasticity acts as a crucial link between cancer and regeneration, relying on stable post-transcriptional modifications and factors like HuD protein. And the common point between regeneration-cancer comes from the characteristic of cellular reprogramming that direct either pre-cancerous or regenerative/rejuvenating processes, depending on immune privileges and the normalization of morphogenetic fields. In light of this, this shared mechanistic program offers potential for innovative therapeutic approaches and demand for the inclusion of multidisciplinary and collaborative strategies in translational oncology, with constructivist algorithmic models, expanding the level of perception extracted from high-resolution genetic and functional data, as well as comparative studies with models that bring together these two themes, such as cnidarians (e.g. genus *Hydra*), offering potential innovative therapeutic approaches.

Keywords: Regeneration. Integrative Oncology. Comparative Study. Biological Evolution. Systems Biology.

References

Aktipis, C. A. *et al*. (2015). Cancer across the tree of life: cooperation and cheating in multicellularity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1673), 20140219.

Albuquerque, T. A. *et al*. (2018). From humans to hydra: patterns of cancer across the tree of life. *Biological Reviews*, 93(3), 1715-1734.

Baramiya, M. G. and Baranov, E. (2020). From cancer to rejuvenation: incomplete regeneration as the missing link (Part I: the same origin, different outcomes). *Future Science OA*, 6(3), FSO450.

Baramiya, M. G. *et al*. (2020). From cancer to rejuvenation: incomplete regeneration as the missing link (part II: rejuvenation circle). *Future Science OA*, 6(8), FSO610.

Boutry, J. *et al*. (2023). Spontaneously occurring tumors in different wild-derived strains of hydra. *Scientific Reports*, 13(1), 7449.

Casotti, M. C. *et al*. (2024). Integrating frontiers: a holistic, quantum and evolutionary approach to conquering cancer through systems biology and multidisciplinary synergy. *Frontiers in Oncology*, 14.

Echeverri, K. and Zayas, R. M. (2018). Regeneration: From cells to tissues to organisms. *Developmental biology*, 433(2), 109.

Goldman, J. A. and Poss, K. D. (2020). Gene regulatory programmes of tissue regeneration. *Nature Reviews Genetics*, 21(9), 511-525.

Huyghe, A. *et al*. (2024). Cellular plasticity in reprogramming, rejuvenation and tumorigenesis: a pioneer TF perspective. *Trends in cell biology*, 34(3), 255-267.

Ji, K. et a. (2023). Retinoic acid receptor gamma is required for proliferation of pancreatic cancer cells. *Cell Biology International*, 47(1), 144-155.

Kang, W. *et al*. (2022). Regeneration Roadmap: database resources for regenerative biology. *Nucleic Acids Research*, 50(D1), D1085-D1090.

Levin, M. (2011). The wisdom of the body: future techniques and approaches to morphogenetic fields in regenerative medicine, developmental biology and cancer. *Regenerative medicine*, 6(6), 667-673.

Levin, M. (2021). Bioelectric signaling: Reprogrammable circuits underlying embryogenesis, regeneration, and cancer. *Cell*, 184(8), 1971-1989.

Liu, L. *et al*. (2018). Transcriptomic analysis of Portunus trituberculatus reveals a critical role for WNT4 and WNT signalling in limb regeneration. *Gene*, 658, 113-122.

Milanovic, M. *et al*. (2018). The senescence–stemness alliance–a cancer-hijacked regeneration principle. *Trends in cell biology*, 28(12), 1049-1061.

Murugan, N. J. *et al*. (2024). Biophysical control of plasticity and patterning in regeneration and cancer. *Cellular and Molecular Life Sciences*, 81(1), 9.

Reiter, S. *et al*. (2012). Hydra, a versatile model to study the homeostatic and developmental functions of cell death. *International Journal of Developmental Biology*, 56(6), 593.

Riss, J. *et al*. (2006). Cancers as wounds that do not heal: differences and similarities between renal regeneration/repair and renal cell carcinoma. *Cancer research*, 66(14), 7216-7224.

Salinas-Saavedra, M. *et al*. (2023). Senescence-induced cellular reprogramming drives cnidarian whole-body regeneration. *Cell reports*, 42(7).

Srivastava, M. (2021). Beyond casual resemblance: rigorous frameworks for comparing regeneration across species. *Annual Review of Cell and Developmental Biology*, 37(1), 415-440.

Tran, A. P. *et al*. (2018). The biology of regeneration failure and success after spinal cord injury. *Physiological reviews*, 98(2), 881-917.

Yasmann, A. (2021). From receptor to organ: Serotonin's interaction with the 5-HT1B receptor and its role in skeletal muscle repair (Doctoral dissertation, lmu).

Zhao, M. *et al*. (2016). REGene: a literature-based knowledgebase of animal regeneration that bridge tissue regeneration and cancer. *Scientific reports*, 6(1), 23167.

Zimmerman, M. A. *et al*. (2013). Cell death–stimulated cell proliferation: A tissue regeneration mechanism usurped by tumors during radiotherapy. *Seminars in radiation oncology*, 23(4), 288-295.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Verdict: An Interactive Web Tool for Exploring Disease Modules and Drug Targets within the Human Interactome

María del Mar Sánchez Rojas[1], Mateo Torres[1,2], and Alberto Paccanaro[1,3]

1. Escola de Matemática Aplicada, Fundação Getúlio Vargas, Rio de Janeiro, Brazil
2. Department of Computational Biology, Weill Institute for Cell and Molecular Biology, Center for Innovative Proteomics, Cornell University, Ithaca, NY, USA
3. Department of Computer Science, Centre for Systems and Synthetic Biology, Royal Holloway University of London, Egham, UK

In Network Medicine, hereditary diseases are conceptualized as perturbations that originate within specific regions of the protein-protein interaction network (interactome), known as disease modules. Identifying these modules is crucial for understanding disease mechanisms and predicting disease genes [1]. Similarly, a drug binding to a protein target causes perturbations that propagate through the interactome.

We present Verdict, an interactive visualization tool designed to analyze diseases, drug targets, and custom protein groups within the human interactome. Verdict also enables researchers and clinicians to apply algorithms for predicting disease-related genes and drug targets, contextualizing these predictions within the interactome to facilitate hypothesis generation and experimental testing. In particular, disease genes predictions are obtained using Cardigan [2], a state-of-the-art prediction method and drug targets predictions are obtained using our in-house drug target prediction algorithm [3].

Verdict also offers metrics such as disease similarity based on the Caniza measure [4] and graph kernel methods that can quantify the similarity and proximity of nodes within the interactome, thus enhancing our understanding of the relationships between diseases. Functional profiling through over-representation analysis of Gene Ontology terms and Anatomical Therapeutic Chemical (ATC) categories [5] is also provided to further contextualize protein groups. Integrating data from OMIM [6], Entrez [7], UniProtKB [8], and DrugBank [9], and incorporating extensive similarity scores and gene predictions, Verdict provides a comprehensive overview of over 10,120 diseases, 19,700 proteins, 773,500 interactions, and 5,800 drugs within a single, accessible platform. This will enable scientists to formulate new research hypotheses through interactive exploration of the interactome, and enhance our understanding of disease pathology.

Keywords: *Network Medicine, Disease Modules, Interactome Visualization, Disease Gene Prediction, drug-target prediction*

References

1. Barabási, A.L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nature Reviews Genetics, 12, 56–68. doi: 10.1038/nrg2918.

2. Cáceres, J.J., and Paccanaro, A. (2019). Disease gene prediction for molecularly uncharacterized diseases. PLOS Computational Biology, 15(7), e1007078. doi: 10.1371/journal.pcbi.1007078.

3. Noto, S., Galeano, D., Jimenez, R., and Paccanaro, A. (Manuscript in preparation). An explainable self-expressive model for Drug Target Prediction.

4. Caniza, H., Romero, A., and Paccanaro, A. (2016). A network medicine approach to quantify distance between hereditary disease modules on the interactome. Scientific Reports, 5, 17658. doi: 10.1038/srep17658.

5. WHO Collaborating Centre for Drug Statistics Methodology. (2024). Guidelines for ATC classification and DDD assignment, 2024. Oslo.

6. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD. Available at: https://omim.org/

7. Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., Wang, J., Williams, R., Trawick, B.W., Pruitt, K.D., and Sherry, S.T. (2022). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 50(D1), D20–D26. doi:10.1093/nar/gkab1112.

8. The UniProt Consortium. (2023). UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research, 51(D1), D523–D531. doi:10.1093/nar/gkac1052.

9. Knox, C., Wilson, M., Klinger, C.M., et al. (2024). DrugBank 6.0: the DrugBank Knowledgebase for 2024. Nucleic Acids Research, 52(D1), D1265–D1275. doi:10.1093/nar/gkad976.

Brazilian Symposium on Bioinformatics 2024

Improvement of the assembly and Annotation of the *Trypanosoma cruzi* Genome Using Hi-C Data

Pedro Leonardo Carvalho de Lima[1,2], Natalia Karla Bellini[1], Pedro Gabriel Nachtigall[1], David da Silva Pires[1], Julia Pinheiro Chagas da Cunha[1]

1.University of São Paulo
2.Butantan Institute.

The protozoan *Trypanosoma cruzi*, the causative agent of Chagas disease, has a highly complex and repetitive genome, posing significant challenges for downstream analysis. Existing genome assemblies for T. cruzi, Dm28c strain, contain 636 contigs and 53.2 Mb, raising concerns about the quality and contiguity of the assembly. To address these limitations, we used chromosome conformation capture techniques (Hi-C dataset) to improve the genome assembly. By integrating Hi-C data with PacBio long reads, we reassembled the genome using Minimap2, scaffolded with Hi-C data through YaHS, and performed manual curation with Juicebox. This process removed duplicated contigs and refined the scaffold structure, resulting in a new assembly with 99 scaffolds (N50 835 Kbp, compared to 347 Kbp previously). We comprehensively annotated genomic features by applying an in-house script based on BLAST analysis leading to the identification of 802 non-coding RNAs (such as tRNAs, rRNAs, and snoRNAs) and 14 telomeric sequences identified by performing BLASTn searches. Additionally, we investigated centromeric regions using ChIP-seq data, focusing on co-localization with repetitive sequences enriched in these regions. In summary, the improved genome assembly demonstrated increased completeness, as evaluated by BUSCO, and the refined annotation provided deeper insights into telomeric and centromeric regions. This enhanced assembly and annotation represent a valuable resource for future studies on the biology and pathogenicity of *T. cruzi*.

Keywords: *(T. cruzi, Genome, Hi-C, Assembly, Annotation)*

References

El-Sayed, N.M., Myler, P.J., Bartholomeu, D.C., Nilsson, D., Aggarwal, G., Tran, A.N., Ghedin, E., Worthey, E.A., Delcher, A.L., Blandin, G., Westenberger, S.J., Caler, E., Cerqueira, G.C., Branche, C., Haas, B., Anupama, A., Arner, E., Åslund, L., Attipoe, P., Bontempi, E., Bringaud, F., Burton, P., Cadag, E., Campbell, D.A., Carrington, M., Crabtree, J., Darban, H., da Silveira, J.F., de Jong, P., Edwards, K., Englund, P.T., Fazelina, G., Feldblyum, T., Ferella, M., Frasch, A.C., Gull, K., Horn, D., Hou, L., Huang, Y., Kindlund, E., Klingbeil, M., Kluge, S., Koo, H., Lacerda, D., Levin, M.J., Lorenzi, H., Louie, T., Machado, C.R., McCulloch, R., McKenna, A., Mizuno, Y., Mottram, J.C., Nelson, S., Ochaya, S., Osoegawa, K., Pai, G., Parsons, M., Pentony, M., Pettersson, U., Pop, M., Ramirez, J.L., Rinta, J., Robertson, L., Salzberg, S.L., Sanchez, D.O., Seyler, A., Sharma, R., Shetty, J., Simpson, A.J., Sisk, E., Tammi, M.T., Tarleton, R., Teixeira, S., Aken, S.V., Vogt, C., Ward, P.N., Wickstead, B., Wortman, J., White, O., Fraser, C.M., Stuart, K.D., Andersson, B. (2005). The Genome Sequence of Trypanosoma cruzi, Etiologic Agent of Chagas Disease. Science, 309(5733), 409–415.

Berná, L., Rodriguez, M., Chiribao, M.L., Parodi-Talice, A., Pita, S., Rijo, G., Alvarez-Valin, F., Robello, C. (2018). Expanding an expanded genome: long-read sequencing of Trypanosoma cruzi. Microbial Genomics, 4(5).

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science, 33292:289–294.

Heng, L. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18), 3094–3100.

Zhou, C., McCarthy, S.A., Durbin, R. (2023). YaHS: yet another Hi-C scaffolding tool. Bioinformatics, 39(1).

Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., Lieberman Aiden, E. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Systems, 3(1).

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol., 215(3), 403-10. doi: 10.1016/S0022-2836(05)80360-2. PMID: 2231712.

Chiurillo, M.A., Cano, J.F.D.S., Ramirez, J.L. (1999). Organization of telomeric and sub-telomeric regions of chromosomes from the protozoan parasite Trypanosoma cruzi. Mol. Biochem. Parasitol., 100, 173–183.

Turowec, J.P., et al. (2010). Chapter 23 - Protein Kinase CK2 Is a Constitutively Active Enzyme that Promotes Cell Survival: Strategies to Identify CK2 Substrates and

Manipulate its Activity in Mammalian Cells. Methods Enzymol. Const. Act. Recept. Other Proteins, Part A, 484, 471–493.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210–3212.

Bellini, N.K., de Lima, P.L.C., Pires, D.d.S., da Cunha, J.P.C. (2024). Hidden origami in Trypanosoma cruzi nuclei highlights its nonrandom 3D genomic organization. bioRxiv.

BSB 2024 Brazilian Symposium on Bioinformatics 2024

Helix-Shifts and Incongruence in RNA Evolution

Maria Waldl[1,2], Peter F. Stadler[1,2,3,4,5,6]

1. *Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, Universität Leipzig, Leipzig, Germany*
2. *Institute for Theoretical Chemistry, University of Vienna, Vienna, Austria*
3. *Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany*
4. *Center for Non-Coding RNAs in Technology and Health, University of Copenhagen, Denmark*
5. *Faculty of Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia*
6. *Santa Fe Institute, Santa Fe, NM*

Since RNA function is often linked to structure, selection acts to preserve RNA structures throughout evolution. Typically, individual base pairs are conserved leading to similar structure formed by homologous sequence elements [Parsch et al. 2000, Washietl et al. 2005, Rivas et al. 2017]. In some cases, however, it has been observed that structural elements appear shifted with respect to the sequence so that analogous base pairs are no longer formed by homologous nucleotides. In order to better understand the prevalence of this kind of evolutionary incongruence, we explore here the possibility of generating shifted helices by introducing random mutations into RNA sequences. Our analysis reveals that helix shift in response to a small number of mutations is not an overly rare phenomenon, albeit it is less frequent than perfect structural conservation. Helix shifts thus are readily accessible to evolutionary processes. We propose that, alongside established models focusing on compensatory mutations, new approaches are required to detect and evaluate instances of incongruent evolution.

Towards that goal, we propose the concept of "Bi-alignments" as model for such incongruence [Waldl et al. 2020]: The key idea is to represent sequence and structure homology by separate alignments, U and V, linked together by an an intermediary alignment W that represents the shifts between the sequence and the structure of the individual RNAs. Bi-alignments can be computed efficiently, requiring only a constant-factor increase in effort compared to individual alignments when shift limits are imposed. By integrating established heuristics and sparsification techniques, we can implement an efficient Sankoff style bi-alignment model that can reliably compute (potentially shifted) consensus structures and sequence homology[Sankoff 1985, Will et al. 2007]. This approach provides a robust framework to further study incongruence in sequence-structure evolution.

Keywords: RNA secondary structure, RNA alignment, incongruent evolution

References:

Parsch, J., Braverman, J. M., and Stephan, W. (2000). Comparative sequence analysis and patterns of covariation in RNA secondary structures. Genetics, 154:909–921.

Rivas, E., Clements, J., and Eddy, S. R. (2017). A statistical test for conserved RNAstructure shows lack of evidence for structure in lncRNAs. Nature Methods, 14:45–48.

Sankoff, D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM J. Appl. Math. 45, 810–825.

Waldl, M., Will, S., Wolfinger, M., Hofacker, I. L., and Stadler, P. F. (2020). Bi-alignments as models of incongruent evolution of RNA sequence and secondary structure. In Cazzaniga, P., Besozzi, D., Merelli, I., and Manzoni, L., editors, Computational Intelligence Methods for Bioinformatics and Biostatistics, 16th International Meeting, CIBB'19, volume 12313 of Lect. Notes Comp. Sci., pages 159–170, Cham, CH. Springer Nature.

Washietl, S., Hofacker, I. L., and Stadler, P. F. (2005). Fast and reliable prediction of noncoding RNAs. Proc. Natl. Acad. Sci. USA, 102:2454–2459.

Will, S., Missal, K., Hofacker, I.L., Stadler, P.F., Backofen, R. (2007) Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comp. Biol. 3, e65.