Preface

The Brazilian Symposium on Bioinformatics (BSB, Simpósio Brasileiro de Bioinformática) is an international scientific conference focused on Bioinformatics, Computational Biology, Systems Biology, Biomedical Informatics and related areas. It is organized annually by the Brazilian Computer Society (Sociedade Brasileira de Computação – SBC), under the steering of the Special Committee on Computational Biology (Comissão Especial de Biologia Computacional – CE-BioComp). The 18th edition of BSB was held from September 29 to October 2, 2025, at the Gran Mareiro Hotel, in the city of Fortaleza, Brazil. Fortaleza is the capital of Ceará, located in the northeastern region of Brazil. BSB 2025 was co-located with several events, most notably the Brazilian Conference on Intelligent Systems (BRACIS 2025) and the Brazilian Symposium on Databases (SBBD 2025). Participants registered for BSB were able to attend activities of all co-located events, and vice versa.

BSB 2025 was coordinated by Sérgio Lifschitz (PUC-Rio), together with Daniel de Oliveira (UFF) and Kele Belloze (Cefet/RJ) as Chairs. The Technical Program Committee (TPC) for this edition consisted of 32 members from Brazil as well as from Germany and Mexico. This year, the symposium accepted contributions in the form of full papers, short papers and abstracts (posters), receiving a total of 73 submissions. From these, 18 full papers, 7 short papers and, 22 abstracts were accepted. All submitted manuscripts were evaluated through a single-blind peer-review process, with each paper receiving at least two independent reviews. The accepted full and short papers were presented at the conference by one of their authors in one of the five technical sessions held during BSB 2025, and their archival versions are included in these proceedings. The accepted abstracts were presented during a poster session and are included after this preface.

BSB 2025 featured three distinguished keynote speakers: Jing Qin (University of Southern Denmark, Denmark), Antonio Pedro Camargo (University of São Paulo, Brazil), and André C. Ponce de Leon F. de Carvalho (University of São Paulo, Brazil). In this edition, André C. Ponce de Leon F. de Carvalho was honored for his dedication and outstanding contributions to the advancement of Bioinformatics, as well as for founding the Workshop on Bioinformatics (WOB) in 2002, a precursor to the Brazilian Symposium on Bioinformatics (BSB).

We would like to thank everyone who contributed to making BSB 2025 a successful event: the members of the Technical Program Committee, the local organizing teams of BRACIS and SBBD, the volunteers, the keynote speakers, the session chairs, and the researchers who assisted in the evaluation of the works presented during the poster sessions. We also gratefully acknowledge the Brazilian Computer Society (SBC) and the National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq) for their support. Finally, we thank all the authors who contributed to this conference, either through paper submissions or poster abstracts, as well as all conference participants. We sincerely appreciate their contributions and hope to welcome everyone again at BSB 2026.

Daniel de Oliveira and Kele Belloze

Organization

Steering Committee for Special Committee for Computational Biology of the Brazilian Computer Society (CE-BioComp)

- Sergio Lifschitz (coordinator)
- Kele Belloze (vice-coordinator, Cefet/RJ)
- Daniel de Oliveira (UFF)
- Diogo Tschoeke (UFRJ)
- João Carlos Setubal (USP)
- Marcelo da Silva Reis (UNICAMP)
- Raquel Minardi (UFMG)
- Sérgio Nery Simões (IFES)

Technical Program Committee

- Daniel de Oliveira (UFF) - Chair
- Kele Belloze (Cefet/RJ) - Chair
- Adriano Cortes (UFRJ)
- Adriano V. Werhli (FURG)
- Allysson Farias (UFC)
- Ana Carolina Guimaraes (FioCruz)
- Bruno de Oliveira (UFF)
- Daniel Saad Nogueira Nunes (IFB)
- Danielo G. Gomes (UFC)
- Deborah Antunes (FioCruz)
- Diego N. Brandão (CEFET/RJ)
- Diogo Tschoeke (COPPE/UFRJ)
- Eduardo Bezerra (CEFET/RJ)
- Fabrício Martins Lopes (UTFPR)
- Felipe A. Louza (UFU)
- Giuseppe Leite (UNIFESP)
- Glauber Wagner (UFSC)
- Graciela Maria Dias (UFRJ)
- Joao C. Setubal (USP)
- Luis Cunha (UFF)
- Luiz Gadelha (German Cancer Research Center)
- Marcelo Macedo Brigido (UnB)
- Maria Emilia Machado Telles Walter (UnB)
- Mariana Recamonde-Mendoza (UFRGS)
- Maribel Rosales Hernandez (CINVESTAV Irapuato – México)
- Raquel Lopes Costa (INCA)
- Sabrina Azevedo Silveira (UFV)
- Said Sadique Adi (UFV)
- Sergio Campos (UFMG)
- Sergio Lifschitz (PUC-Rio)
- Waldeyr Mendes Cordeiro da Silva (IFG)
- Zanoni Dias (UNICAMP)

Realization

- Brazilian Computer Society (Sociedade Brasileira de Computação – SBC)

Financial Support

- National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq)

Poster Abstracts

P01 - A Robust Transcriptomic Pipeline Identifies Novel LncRNAs Associated with Sexual Dimorphism in Schistosoma mansoni
Caio Freire (USP), Thalles Souza-Lopes (USP), Ana Tahira (Instituto Butantan), Sergio Verjovski (USP)

P02 - A shuffle-based statistical approach for robust pseudogene annotation
Pedro Barcelos (PUC-Rio), Marcos Catanho (FioCruz), Antonio B de Miranda (FioCruz), Edward Hermann Haeusler (PUC), Sergio Lifschitz (PUC-Rio)

P03 - Integrating Clinical Guidelines and Genomic Data with AI: Towards Adaptive RAG in Healthcare
Jean Paes Landim de Lucena (UFRN), Patrick Terrematte (UFRN)

P04 - Classification and Annotation of Metagenome-Assembled Genomes
Izabel Gomes (USP)

P05 - CuratedGeneDis: Towards a Sentence-Level Gene–Disease Relation Extraction from Oncology Abstracts
Bryan Khelven Barbosa (UFSCAR), Maria Julia Porto (UEPB)

P06 - Exploring soil microbial diversity of PARNA Tijuca
Beatriz Moura da Silva (UFRJ), Danielly Mariano (UFRJ), Lucia Bahiense (UFRJ), Paulo Bisch (Federal University of Rio de Janeiro), Diogo Antonio Tschoeke (UFRJ), Graciela Maria Dias (UFRJ)

P07 - Exploring TabPFNv2 as a Novel Baseline for ADMET Prediction in Drug Discovery
Oroel Ipas (University of Granada, Spain), Ignacio Suárez Martín (Universidad de Granada), Guillermo Gomez-Trenado (Universidad de Granada), Isaac Triguero (University of Granada), Rocío Romero Zaliz (University of Granada)

P08 - Forecasting binding affinity of paratope-epitope surfaces with a novel decision tree framework
Shawnak Shivakumar (MAHS)

P09 - Identification and annotation of alternative splicing of microexon genes in version 10 of the Schistosoma mansoni genome
Karolline Silva (USP), Ana Tahira (Instituto Butantan), Sergio Verjovski-Almeida (USP)

P10 - Identification of IP3 Pathway Components in Plasmodium falciparum and P. chabaudi Using Gene Co-expression Networks.
Lyang Higa Cano (USP), Celia Regina da Silva Garcia (USP), Ronaldo Fumio Hashimoto (USP)

P11 - Impact of ITIM1 Polymorphisms on SHP-2-LILRB2 Complex
Laura Maria de Araújo Pereira (USP), Silvana Giuliatti (USP)

P12 - Improving Clustering Coherence in scRNA-seq Data with Prior Knowledge and Pairwise Constraints
Davi Guimarães (Cefet/RJ), Mateus Pereira (IBM Research), Kele Belloze (Cefet/RJ), Marcel Pedroso (FioCruz), Eduardo Bezerra (Cefet/RJ)

P13 - In silico identification of SEPs encoded by long non-coding RNAs
Rafael Nascimento (USP), Eduardo Reis (USP), Joao C. Setubal (USP)

P14 - Mesoscopic model application to 2-O-methylribose/RNA hybrids reveals base fraying and salt concentration effects
Daniel Jesus (UFMG), Gerald Weber (UFMG)

P15 - Molecular Modelling as a Strategy to Specific Antibody Design for ApoE4's Carries
Victor Andrade (USP), Silvana Giuliatti (USP)

P16 - Multivariate conformational patterns may distinguish HLA-DRB1 alleles associated with Multiple Sclerosis susceptibility
Levy Bueno Alves (USP), Silvana Giuliatti (USP)

P17 - Optimization of genomic prediction in young nellore bulls: balancing production and resilience for tropical cattle production
Leonardo Melchior (UFAC)

18 - Simple yet robust annotation of homologous and non-homologous isofunctional enzymes
Emanuel Umbelino, Alexander da Franca Fernandes (Fundação Oswaldo Cruz Brasil), Sergio Lifschitz (PUC-Rio), Edward Hermann Haeusler (PUC-Rio), Ana Carolina Guimaraes (FioCruz), Marcos Catanho (FioCruz), Antonio B de Miranda (FioCruz)

P19 - SynDRA: Synonym Drug Repurposing Alignment
Tolga Corbaci (Beth Israel Deaconess Medical Center, Harvard Medical School, Bahçeşehir University), Katjusa Koler (University of Sheffield), Erlend Skaga (University Hospital, Oslo), Pourya Naderi Yeganeh (Beth Israel Deaconess Medical Center, Harvard Medical School), Winston Hide (Beth Israel Deaconess Medical Center, Harvard Medical School)

P20 - Towards a Roadmap for Translational Bioinformatics in the Era of AI and Personal Data Governance
Luana Alvarenga (UFRN), Leonardo César Teonácio Bezerra (University of Stirling), Renan Moioli (UFRN), Cesar Renno-Costa (UFRN)

P21 - Type VI Secretion System in Pseudomonas aeruginosa – distribution and comparative analysis
Hadassa Loth de Oliveira (UFRJ), Graciela Maria Dias (UFRJ), Bianca Cruz Neves (UFRJ)

P22 - Uncovering the hidden structure of subspecies in large metagenomic datasets
Arthur Henrique Barrios Solano (USP), Joao C. Setubal (USP)

A Robust Transcriptomic Pipeline Identifies Novel LncRNAs Associated with Sexual Dimorphism in *Schistosoma mansoni*

Caio Felipe Freire de Sousa[1,2], Thalles Souza-Lopes[1,2], Ana C. Tahira[1], Sergio Verjovski-Almeida[1,3]

1. *Laboratório de Ciclo Celular, Instituto Butantan (São Paulo, Brasil)*
2. *Programa Interunidades de Pós-Graduação em Bioinformática, Universidade de São Paulo (São Paulo, Brasil)*
3. *Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo (São Paulo, Brasil)*

Human schistosomiasis, caused by the trematode *Schistosoma mansoni*, remains a major neglected disease. With no vaccine and reliance on a single drug, praziquantel, with known limitations, there is an urgent need for novel therapeutic targets. Long non-coding RNAs (lncRNAs) are known regulators of the parasite's biology, but their characterization has been challenging due to their transcription from genomic regions rich in repetitive elements, low conservation among species, high tissue-specificity and the availability of incomplete genome assemblies. To address this, we developed a robust, reproducible bioinformatic pipeline, to construct the most comprehensive lncRNA catalog for *S. mansoni* to date using the recently, fully assembled V10 genome and performing an initial functional characterization of these transcripts. A dataset of 1,797 public bulk RNA-seq libraries, comprising diverse developmental stages and tissues across 49 BioProjects, was processed with our pipeline. Libraries from single-cell RNA-seq and from small RNAs were excluded. Read quality control (QC) was assessed with fastp, followed by STAR alignment to the genome, and post-alignment QC with RseqC; libraries with low Transcript Integrity Number (TIN<70), with low gene body coverage (3'-bias) and non-stranded libraries were removed. Libraries with reads mapping to less than 70% of exonic regions were also excluded due to possible genomic DNA contamination. A total of 1,107 bulk RNA-seq libraries were retained. Hierarchical guided assembly of these libraries with Ryūtō was performed, to generate 12 stage- and tissue-specific primary transcriptomes. Ryūtō was benchmarked against other assemblers, Scallop+TACO or StringTie+StringTieMerge, and chosen for the reliable assembly of transcripts based on the F1 score, the harmonic mean of precision and recall of protein-coding genes used as reference. Assemblies were subjected to a lncRNA annotation workflow combining a three-tool consensus (FEELnc, CPC2, CPAT), orthology filtering with eggNOG database, and removal of assembly artifacts. The resulting assemblies were merged with StringTie to create a final consensus transcriptome. This pipeline successfully identified 15,482 lncRNA genes, of which 9,490 (61%) are completely novel genes. To infer function and verify their involvement in important sex-biased regulatory processes, we performed differential gene expression (DGE) and Weighted Gene Co-expression Network Analysis (WGCNA) on adult male and female control samples (those without treatment). Batch effects originated from library heterogeneity and experimental bias across BioProjects were corrected using ComBat-seq with sexes as 'group' and Bioprojects as 'batch' parameters. DGE analysis revealed 1,813 sex-biased genes, including 607 lncRNAs (100 of which are novel). WGCNA,

performed with a soft-threshold power of 7, identified 18 co-expression modules significantly associated with parasite sex through module-trait correlation analysis (adjusted p-value ≤0.05). Male-associated modules were enriched for genes related to morphogenesis and female-associated modules for genes related to cell proliferation. Within these networks, we identified the lncRNA genes with highest module membership that was also categorized as differentially expressed in DEG analysis, resulting in 277 high-confidence lncRNA hubs related to the 'Sex' trait, of which 44 are entirely novel discoveries. This work provide the most comprehensive catalog of *S. mansoni* lncRNAs to date and uncover novel regulatory hubs with direct links to sexual dimorphism.

Keywords: *RNA-seq analysis, Systems Biology, Neglected Tropical Diseases, Computational Biology, Parasite Reproduction*

Financial Support: FAPESP grant 2023/14590-6.

References

Buddenborg, S. K. et al. (2021). Assembled chromosomes of the blood fluke Schistosoma mansoni provide insight into the evolution of its ZW sex-determination system. *bioRxiv*, https://doi.org/10.1101/2021.08.13.456314.

Dobin, A. et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, v. 29, n. 1, p. 15-21.

Gatter, T. and Stadler, P. F. (2019). Ryūtō: Network-flow based transcriptome reconstruction. *BMC Bioinformatics*, v. 20, n. 1, p. 1-14.

Langfelder, P. and Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, v. 9, n. 1, p. 559.

Leek, J. T. et al. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, v. 28, n. 6, p. 882-883.

Love, M. I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, v. 15, n. 12, p. 550.

Pertea, M. et al. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, v. 33, n. 3, p. 290-295.

Silveira, G. O. et al. (2023). Long non-coding RNAs are essential for Schistosoma mansoni pairing-dependent adult worm homeostasis and fertility. *PLOS Pathogens*, v. 19, n. 5, p. e1011369.

Wucher, V. et al. (2017). FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Research*, v. 45, n. 2, p. e7.

Yu, G. et al. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, v. 16, n. 5, p. 284-287.

A shuffle-based statistical approach for robust pseudogene annotation

Pedro M. Barcelos [2] , Marcos Catanho [1,] , Antônio Basílio de Miranda [1,] , Edward H. Haeusler [2] & Sérgio Lifschitz [2]

1. Instituto Oswaldo Cruz, Fiocruz, Brazil
2. Pontifícia Universidade Católica do Rio de Janeiro, Brazil

Pseudogenes are stretches of genomic DNA that resemble known genetic sequences but have lost their ability to produce an active (or putatively active) product, usually a protein, due to mutations or other disabling changes. These "genetic fossils" often arise from duplication or retrotransposition events, or from once-effective genes that accumulated inactivating mutations: premature stop codons, frameshift mutations, missing promoters and/or start codons, or truncations that prevent normal transcription/translation. The most common approach to identify pseudogenes is the similarity to known or putative proteins. However, the accurate annotation of pseudogenes is a significant challenge in genomics, as their decaying sequences often fall into a "twilight zone" of similarity to known genetic sequences that confounds automated methods. One major limitation is the arbitrary similarity thresholds applied, leading to a critical problem: the criteria used to define a significant sequence similarity are often not uniform across different studies, resulting in inconsistent and frequently incomplete annotations. This lack of a standardized foundation for a biologically and statistically-grounded significance motivates the need for a more objective approach. To address this, we developed a genome-wide comparative methodology that does not rely on pre-existing gene models to discriminate true sequence remnants from stochastic background noise, providing a reliable framework for annotating protein-coding sequence (CDS) pseudogenes in large-scale genomic projects [Abrahim et al., 2022]. It employs a customized, statistically derived filter to confidently identify genomic sequence regions that exhibit significant sequence similarity to computationally predicted or experimentally verified protein sequences (Barcelos et al., 2025), providing systematic and automated discovery of both preserved and degraded CDS genome-widely. Applying our approach to the reference genome of E. coli str. K-12 substr. MG1655 [Blattner et al. 1997] with the protein sequences in UniProtKB reviewed (Swiss-Prot) database as references, we found 3.354 sequences with no observable DNA strand preference, displaying statistically significant sequence similarity to functional proteins, ranging from 72 to 11.865 nt in length, comprising 386 small ORF (smORF) [Guerra-Almeida et al., 2021] (< 300 nt) and 2.967 CDS (>= 300 nt).

Keywords: pseudogene, CDS, sequence similarity, genome annotation

References

Abrahim M, Machado E, Alvarez-Valín F, de Miranda AB, Catanho M. Uncovering Pseudogenes and Intergenic Protein-coding Sequences in TriTryps' Genomes. Genome Biol Evol. 2022 Oct 7;14(10):evac142. doi: 10.1093/gbe/evac142. PMID: 36208292; PMCID: PMC9576210.

Barcelos PM, Catanho M, de Miranda AB, Haeusler EH, Lifschitz S. A Shuffle-Based Statistical Approach for Robust Pseudogene Annotation. Short paper submitted to BSB 2025

Diego Guerra-Almeida, Diogo Antonio Tschoeke, Rodrigo Nunes-da-Fonseca, Understanding small ORF diversity through a comprehensive transcription feature classification, *DNA Research*, Volume 28, Issue 5, October 2021, dsab007, doi.org/10.1093/dnares/dsab007

Integrating Clinical Guidelines and Genomic Data with AI: Towards Adaptive RAG in Healthcare

Jean Paes Landim de Lucena[1], Patrick Terrematte[1,2]

1. Postgraduate Program in Bioinformatics, Bioinformatics Multidisciplinary Environment – BioME, Universidade Federal do Rio Grande do Norte – UFRN.
2. Metropole Digital Institute, UFRN.

The healthcare workflow demands the analysis of vast volumes of information, including medical records, laboratory results, and genomic data. The availability of such data grows daily, often overwhelming professionals tasked with aligning clinical and laboratory inputs with established protocols and guidelines to generate effective outputs for individualized diagnostic and therapeutic propaedeutics, tailored to the specific clinical context. In this scenario, tools based on large language models (LLMs) enable the handling of genomic data complexity and support clinical reasoning by optimizing information triage within an evidence-based practice framework. These tools aid in generating differential diagnoses and recommending personalized treatments. Within precision medicine, genomic data analysis facilitates the application of personalized therapies and preventive strategies. However, interpreting genomic data remains challenging, even for specialized professionals. Generative AI and large language models can assist in analyzing genetic variants and diseases, improving diagnostic accuracy and enhancing physician-patient communication. This project will present an API infrastructure based on intelligent agents implemented in LangGraph, utilizing an adaptive and self-corrective strategy via Retrieval-Augmented Generation (Adaptive RAG). The system employs the Llama 3.1 (70B) LLM model, with vectorized tokens stored in ChromaDB and metrics managed in LangSmith. The objective is to support clinical decision-making by integrating patient data with the Brazilian Ministry of Health's Clinical Protocols and Therapeutic Guidelines (PCDT), alongside continuously updated genomic data from the Ensembl Variation API. This infrastructure aims to enhance diagnostic propaedeutics across healthcare tiers and enable precise, individualized clinical follow-up, particularly in resource-limited settings. As a project outcome, we will demonstrate a proof-of-concept chatbot addressing queries related to Type 2 Diabetes Mellitus PCDT guidelines and corresponding genetic testing options available in MalaCards.This infrastructure is under construction, and the results reported here reflect preliminary work in progress.

Keywords:
Multi-Agent Systems. Generative AI. Large Language Models. Precision Medicine.

References

Ao, G. *et al.* (2025). *Comparative analysis of large language models on rare disease identification*. Orphanet Journal of Rare Diseases. Disponível em: https://doi.org/10.1186/s13023-025-03656-w. Acesso em: 14 jun. 2025.

Brasil. Ministério da Saúde. (2023). *Protocolos Clínicos e Diretrizes Terapêuticas (PCDT)*. Ministério da Saúde, Brasília, DF. Disponível em: https://www.gov.br/conitec/pt-br. Acesso em: 14 jun. 2025.

Chen, S. *et al.* (2024). *A genomic mutational constraint map using variation in 76,156 human genomes*. Nature. Disponível em: https://doi.org/10.1038/s41586-023-06045-0. Acesso em: 4 ago. 2025.

Ling, C.; Chen, H. *et al.* (2024). *Uncertainty Quantification for In-Context Learning of Large Language Models*. arXiv. Disponível em: https://doi.org/10.48550/arXiv.2402.10189. Acesso em: 14 jun. 2025.

López, L. J. L. *et al.* (2025). *Uncertainty Quantification for Machine Learning in Healthcare: A Survey*. arXiv. Disponível em: https://arxiv.org/abs/2505.02874. Acesso em: 15 jun. 2025.

Rappaport, N. *et al.* (2017). *MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search*. Nucleic Acids Research. Disponível em: https://doi.org/10.1093/nar/gkw1012. Acesso em: 14 jun. 2025.

Saab, K.; Natarajan, V. *et al.* (2024). *Capabilities of Gemini Models in Medicine*. arXiv. Disponível em: https://doi.org/10.48550/arXiv.2404.18416. Acesso em: 14 jun. 2025.

Singhal, K. *et al.* (2023). *Large language models encode clinical knowledge*. Nature. Disponível em: https://www.nature.com/articles/s41586-023-06291-2. Acesso em: 14 jun. 2025.

Touvron, H. *et al.* (2024). *The Llama 3 Herd of Models*. arXiv preprint arXiv:2307.09288. Disponível em: https://arxiv.org/abs/2407.21783. Acesso em: 14 jun. 2025.

Tu, T.; Schaekermann, M. *et al.* (2025). *Towards conversational diagnostic artificial intelligence*. Nature. Disponível em: https://doi.org/10.1038/s41586-025-08866-7. Acesso em: 14 jun. 2025.

Classification and Annotation of Metagenome-Assembled Genomes

João Carlos Setubal[1], Izabel de Geus Monteiro Gomes[1,2]

1. Institute of Chemistry, University of São Paulo
2. Polytechnic School of the University of São Paulo

Metagenomics has revolutionized the study of microbial diversity by enabling the reconstruction of a rapidly growing number of Metagenome-Assembled Genomes (MAGs). Despite these advances, a substantial proportion of MAGs remain taxonomically unclassified, limiting the understanding of their ecological and evolutionary significance (Parks et al., 2017). To address this challenge, we present the development of a computational tool designed to classify and annotate MAGs of prokaryotic microorganisms, providing both taxonomic assignments and contextual ecological insights. MAGs can be categorized into three major groups (Setubal, 2021): (i) SMAGs, which are classified to the species level; (ii) HMAGs, which remain unclassified and lack known orthologs; and (iii) CHMAGs, which are unclassified but possess orthologous MAGs likely belonging to the same unknown species. Two MAGs are considered orthologous if they share ≥95% average nucleotide identity (ANI) over ≥80% of the length of the smaller genome. Thus, classification depends directly on the presence and type of orthologs. Identifying and contextualizing such relationships is critical for advancing microbial ecology, as ecosystem metadata often reveals patterns of microbial cosmopolitanism (Green and Bohanna, 2006). Our pipeline, developed in Python, integrates several robust bioinformatics components, including GTDB-Tk, the Entrez database (Baxevanis, 2006), and the Genomes OnLine Database (GOLD) API (Mukherjee et al., 2023). Given a user-submitted MAG, the tool first computes ANI and alignment fraction (AF) via GTDB-Tk to classify the MAG as SMAG, CHMAG, or HMAG. For MAGs classified as SMAGs or CHMAGs, the pipeline proceeds to: (1) identify orthologous MAGs and isolate genomes; (2) retrieve taxonomic, ecological, and geographic metadata for each ortholog via Entrez and GOLD; and (3) generate interactive graphical visualizations that highlight ecological niches and biogeographic distributions. The preliminary version of the tool was tested using 60 MAGs sourced from the MetaZoo project (Braga et al., 2021). These results suggest that, in addition to its classification capabilities, our pipeline offers complementary advantages over existing taxonomic tools (Kutuzova et al., 2024; Wood, Lu, and Langmead, 2019). Unlike other pipelines, it emphasizes identifying MAGs likely belonging to the same species using conservative, well-defined criteria (Arahal, 2014; Goris et al., 2007), rather than solely aiming for complete taxonomic assignments. By integrating ecological and geographic metadata, the pipeline also enables assessment of microbial cosmopolitanism and characterization of habitats for poorly classified or uncharacterized species, providing additional ecological insights.

Keywords: bioinformatics, metagenomics, MAG classification, orthologous genomes, microbial diversity

References

Arahal, D. R. (2014). Chapter 6 – Whole-Genome Analyses: Average Nucleotide Identity. In: Goodfellow, M.; Sutcliffe, I.; Chun, J. (Eds.). *Methods in Microbiology*, v. 41 – New Approaches to Prokaryotic Systematics, Academic Press, p. 103–122.

Baxevanis, A. D. (2006). Searching the NCBI databases using Entrez. In: *Current protocols in human genetics*, v. 51(1), p. 6-10.

Braga, L. P. P. et al. (2021). Genome-resolved metagenome and metatranscriptome analyses of thermophilic composting reveal key bacterial players and their metabolic interactions. In: *BMC genomics*, v. 22, n. 1, p. 652.

Green, J. and Bohanna, B. J. M. (2006). Spatial scaling of microbial biodiversity. In: Trends in Ecology & Evolution, v. 21, n. 9, p. 501–507.

Goris, J. et al. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. In: *International Journal of Systematic and Evolutionary Microbiology*, v. 57, n. 1, p. 81–91.

Kutuzova, S. et al (2024). Taxometer: Improving taxonomic classification of metagenomics contigs. In: *Nature Communications*, v. 15, p. 8357.

Mukherjee, S. et al. (2023). Twenty-five years of Genomes OnLine Database (GOLD): data updates and new features in v. 9. In: *Nucleic acids research*, v. 51, n. D1, p. D957-D963.

Parks, D. H. et al. (2017). Recovery of MAGs expands the tree of life. In: *Nature Microbiology*, v. 2, p. 1533–1542.

Setubal, J. C. (2021). MAGs: concepts and challenges. In: *Biophysical Reviews*, v. 13, p. 905–909.

Wood, D. E., Lu, J. and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. In: *Genome Biology*, v. 20, p. 257.

CuratedGeneDis: Towards a Sentence-Level Gene–Disease Relation Extraction from Oncology Abstracts

Bryan K. S. Barbosa[1,2], Maria Julia S. Porto[3]

1. Universidade Federal de São Carlos
2. Núcleo Interinstitucional de Linguística Computacional – NILC
3. Universidade Estadual da Paraíba

Abstract. We present CuratedGeneDis, an ongoing project that builds a curated sentence-level corpus and baseline models for gene–disease relation extraction from oncology abstracts. We retrieved 1,233 PubMed abstracts (2018–2024) using a mutation-and-cancer query and segmented approximately 3,100 sentences. A high-precision filtering pipeline combined dual biomedical NER for genes/proteins and diseases, validated against HGNC/VGNC standards (Braschi *et al*., 2019), lexical triggers such as *associate*, *link* and *risk*, and disease-specificity constraints, yielding 320 candidate sentences. From these, 100 sentences were manually annotated with three mutually exclusive labels: ASSOCIATED, NOT_ASSOCIATED and UNCERTAIN, following practices in biomedical corpora such as DDI (Herrero-Zazo *et al*., 2013), COMAGC (Lee *et al*., 2013) and BioRED (Luo *et al*., 2022). We benchmarked two approaches: (i) a lightweight logistic regression baseline using frozen biomedical embeddings; and (ii) a fine-tuned domain transformer, building on pre-trained language models such as BERT (Devlin *et al*., 2019), BioBERT (Lee *et al*., 2020) and PubMedBERT (Gu *et al*., 2021). In five-fold cross-validation, the logistic model achieved macro-F1 ≈ 0.62, while fine-tuned PubMedBERT reached macro-F1 ≈ 0.76. Gains concentrated on less frequent classes: F1 improved by roughly 7–8 points for NOT_ASSOCIATED and approximately 30 points for UNCERTAIN, indicating improved handling of speculative language, a challenge also observed in related biomedical RE work (Bravo *et al*., 2015; Su *et al*., 2021). Both models performed best on ASSOCIATED instances, with main errors arising from hedging and ambiguity in UNCERTAIN cases. Given these preliminary results, we intend to make available the current version of CuratedGeneDis – including the annotated corpus, guidelines and reproducible pipeline – as an open resource to support benchmarking and downstream applications in biomedical NLP, with technical support from widely used libraries such as Biopython (Cock *et al*., 2009) and foundations in machine learning such as support-vector networks (Cortes; Vapnik, 1995). These curated relations may also serve as inputs for constructing oncology-specific gene–disease networks, supporting downstream network analysis and integration with other biomedical datasets. Future work will include extending the corpus to full-text articles, refining the annotation process, incorporating external biomedical knowledge sources (e.g., CTD, OMIM, UMLS) for feature enrichment and distant supervision, broadening the entity and relation coverage, exploring advanced modeling strategies such as joint entity–relation extraction and semi-supervised learning, and investigating the use of large language models (LLMs) for zero-shot and few-shot classification, weak supervision for data augmentation, and automated error analysis in low-resource biomedical RE scenarios.

Keywords: *Gene–Disease Relation Extraction; Biomedical NLP; Oncology; Corpus; PubMedBERT*

References

Braschi, B., Denny, P., Gray, K. A., Jones, T. E. M., Seal, R. L., Tweedie, S., Yates, B., and Bruford, E. A. (2019). Genenames.org: the hgnc and vgnc resources in 2019. Nucleic Acids Research, 47(D1):D786–D792.

Bravo, A., Piñero, J., Queralt-Rosinach, N., Rautschka, M., and Furlong, L. I. (2015). Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinformatics, 16:55.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11):1422–1423.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3):273–297.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019, pages 4171–4186.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 3(1):1–23.

Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., and Declerck, T. (2013). The ddi corpus: an annotated corpus with pharmacological substances and drug–drug interactions. Journal of Biomedical Informatics, 46(5):914–920.

Lee, H.-J., Shim, S.-H., Song, M.-R., Lee, H., and Park, J. C. (2013). Comagc: a corpus with multi-faceted annotations of gene–cancer relations. BMC Bioinformatics, 14:323.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240.

Luo, L., Lai, P., Wei, C., Arighi, C. N., and Lu, Z. (2022). BioRED: a rich biomedical relation extraction dataset. Briefings in Bioinformatics, 23(5):bbac282.

Su, J., Wu, Y., Ting, H.-F., Lam, T.-W., and Luo, R. (2021). Renet2: high-performance full-text gene–disease relation extraction with iterative training data expansion. NAR Genomics and Bioinformatics, 3(3):lqab062.

Exploring soil microbial diversity of PARNA Tijuca

Beatriz M. Silva[1], Danielly C. O. Mariano[2], Lucia H. B. Landim[2], Paulo M. Bisch[2], Diogo A. Tschoeke[1], Graciela M. Dias[2]

[1] Programa de Engenharia Biomédica, COPPE, Universidade Federal do Rio de Janeiro
[2] Instituto de Biofísica Carlos Chagas Filho, Centro de Ciências da Saúde, Universidade Federal do Rio de Janeiro

Urban green spaces are critical for the sustainability of cities, functioning as biodiversity reservoirs and providing multiple ecosystem services, including nutrient cycling, regulation of greenhouse gases (e.g., CO_2, CH_4, and N_2O), and organic matter decomposition. An example of such an urban green area is the Parque Nacional (PARNA) da Tijuca, a significant remnant of Atlantic Forest covering approximately 3,200 hectares, located in the city of Rio de Janeiro, Brazil. Despite its ecological relevance, the soil microbial community of PARNA Tijuca remains poorly characterized. This community plays a central role in essential ecosystem processes and represents a valuable resource for biotechnological exploration, particularly in the search for novel bioactive compounds and functional biomolecules. This study aims to characterize the microbial profile of three zones of the PARNA da Tijuca with different levels of anthropogenic disturbance: the primitive zone (ZP), the extensive use zone (ZE), which corresponds to a transition zone, and the intensive use zone (ZI). In addition, we will also correlate the physicochemical parameters, including pH, electrical conductivity, total nitrogen, phosphorus, and total organic carbon (TOC) in association with ecosystem functions, as well as to assess the impact of anthropogenic activities. For taxonomic identification of the soil microbiome (bacteria), five points were collected from each zone at a depth of 0-15 cm. The taxonomic identification was performed by sequencing the 16S ribosomal gene marker (V3-V4 region) using the MiSeq Illumina plataform and the data were processed using Qiime2 pipeline. Our preliminary results showed that the pH profile in the three zones ranged from 4 to 6.66, indicating an acidic soils, which may influence soil mineral nutrient availability. TOC, an indicator of soil quality, showed differences between the ZI and ZE (13,012mg/kg and 8,744mg/kg) and ZP (16,260mg/kg). Taxonomic identification revealed that the predominant phyla in the three zones were Acidobacteriota and Proteobacteria, totaling an average relative abundance of 56%, suggesting no significant differences in the microbial profile between the three zones. As a future perspective, we aim to perform taxonomic analysis using the ITS marker to identify fungal diversity and to correlate the microbial community with physicochemical factors and ecosystem functions.

References

Freitas, S.R., Neves, C.L., Chernicharo, P., (2006). Tijuca National Park: two pioneering restorationist initiatives in Atlantic forest in southeastern Brazil. Brazilian J. Biol. 66, 975–982. https://doi.org/10.1590/S1519-69842006000600004

Banerjee, S., van der Heijden, M.G.A., (2022). Soil microbiomes and one health. Nat. Rev. Microbiol. 6–20. https://doi.org/10.1038/s41579-022-00779-w

Delgado-Baquerizo, Manuel et al. (2021).Global homogenization of the structure and function in the soil microbiome of urban greenspaces. Science Advances, v. 7, n. 28, eabg5809, 2021.https://www.science.org/doi/10.1126/sciadv.abg5809.

SIQUEIRA, A. E. (2013) Guia de campo do Parque Nacional da Tijuca.

Exploring TabPFNv2 as a Novel Baseline for ADMET Prediction in Drug Discovery

Oroel Ipas[1], Ignacio Suárez Martín[1,2], Guillermo Gomez-Trenado[1], Isaac Triguero[1,3], Rocío Romero-Zaliz[1,3,4,5]

1. Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), Granada, Spain
2. Unidad de Excelencia de Química Aplicada a Biomedicina y Medioambiente (UEQ), Departamento de Química Física, Facultad de Ciencias, Universidad de Granada, Granada, Spain
3. Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
4. Research Center in Information and Communication Technologies (CITIC-UGR), University of Granada, Granada, Spain.
5. Instituto de Investigación Biosanitaria ibs.GRANADA, Complejo Hospitales Universitarios de Granada/Universidad de Granada, Granada, Spain

Tabular data is one of the most widely used formats in bioinformatics research. Therefore, improving algorithmic baselines for such data has important implications for a wide range of applications. One of these critical applications is the prediction of Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) properties of drugs, a key step in the early stages of drug development. Failures due to poor pharmacokinetic profiles remain a leading cause of attrition in clinical trials, highlighting the need for reliable predictive tools. In recent years, machine learning has emerged as a powerful approach to model complex ADMET behaviors, enabling faster, more cost-effective, and more ethical drug screening pipelines. While some current state-of-the-art approaches, such as MiniMol or MolE, leverage specialized models pretrained on millions of drug-like molecules, Gradient Boosted Decision Trees algorithms like XGBoost continue to serve as strong baselines for many general-purpose tasks.

The objective of this study is to explore the use of novel tabular foundation models as a new baseline for tabular data in bioinformatics, with a focus on ADMET drug prediction. To this end, we used TabPFNv2, an In-Context Learning model based on transformers that was pretrained on synthetic data. For evaluation, we employed the Therapeutic Data Commons benchmark, comprising 22 datasets that include both regression and classification tasks, and extracted the widely used set of 217 RDKit molecular descriptors.

This generic algorithm outperforms XGBoost in 19 out of 22 datasets and surpasses MiniMol in 9 out of 22, despite not relying on any prior, drug-specific knowledge. Notably, TabPFNv2 achieves the top rank in 3 tasks, surpassing specialized methods in the field. These results suggest that TabPFNv2 is a promising baseline for drug prediction, with potential applications in other bioinformatics tasks, including clinical and small omics datasets that meet TabPFNv2's size constraints. Furthermore, its independence from domain-specific pretraining and hyperparameter tuning enhances its applicability for non-expert practitioners.

Keywords: *(maximum 5 words)*

Tabular Foundation Models, ADMET Drug Prediction, Bioinformatics, Machine Learning in Drug Discovery

References

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T. and Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model, *Nature*, 637(8045), p. 319-326.

Garg, A., Ali, M., Hollmann, N., Purucker, L., Müller, S. and Hutter, F. (2025). Real-TabPFN: Improving Tabular Foundation Models via Continued Pre-training With Real-World Data, arXiv preprint arXiv:2507.03971.

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J. and Zitnik, M. (2021). Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development, *arXiv preprint arXiv:2102.09548.*

Kläser, K., Banaszewski, B., Maddrell-Mander, S., McLean, C., Müller, L., Parviz, A., Huang, S. and Fitzgibbon, A. (2024). MiniMol: A Parameter-Efficient Foundation Model for Molecular Learning, arXiv preprint arXiv:2404.14986.

Mendez-Lucio, O., Nicolaou, C. A., & Earnshaw, B. MolE: a molecular foundation model for drug discovery. In NeurIPS 2022 Workshop on Learning Meaningful Representations of Life.

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

Forecasting binding affinity of paratope-epitope surfaces with a novel decision tree framework

Shawnak Shivakumar[1]

1. Menlo-Atherton High School

Human metapneumovirus (hMPV) poses serious risks to pediatric, elderly, and immunocompromised populations. Traditional antibody discovery pipelines require 10–12 months, limiting their applicability for rapid outbreak response. This project introduces ImmunoAI, a machine learning framework that accelerates antibody discovery by predicting high-affinity candidates using gradient-boosted models trained on thermodynamic, hydrodynamic, and 3D topological interface descriptors. A dataset of 213 antibody–antigen complexes was curated to extract geometric and physicochemical features, and a LightGBM regressor was trained to predict binding affinity with high precision. The model reduced the antibody candidate search space by 89%, and fine-tuning on 117 SARS-CoV-2 binding pairs further reduced Root Mean Square Error (tRMSE) from 1.70 to 0.92. In the absence of an experimental structure for the hMPV A2.2 variant, AlphaFold2 was used to predict its 3D structure. The fine-tuned model identified two optimal antibodies with predicted picomolar affinities targeting key mutation sites (G42V and E96K), making them excellent candidates for experimental testing. In summary, ImmunoAI shortens design cycles and enables faster, structure-informed responses to viral outbreaks.

Keywords: *hMPV, AI/ML, Drug Design, Gradient-Boosted*

References

Centers for Disease Control and Prevention (2024). About human metapneumovirus. Available at: https://www.cdc.gov/human-metapneumovirus/about/index.html (Accessed: 11 April 2024).

Kelley, B. (2020). Developing therapeutic monoclonal antibodies at pandemic pace. Nature Biotechnology, 38(5), pp. 540–545.

Gu, M., Yang, W. and Liu, M. (2024). Prediction of antibody–antigen interaction based on backbone aware with invariant point attention. BMC Bioinformatics, 25(1).

Hollingsworth, S. A. and Dror, R. O. (2018). Molecular dynamics simulation for all. Neuron, 99(6), pp. 1129–1143.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), pp. 583–589.

Identification and annotation of alternative splicing of microexon genes in version 10 of the *Schistosoma mansoni* genome

Karolline Monteiro Vieira da Silva[1,2], Ana Carolina Tahira[1], Sergio Verjovski-Almeida[1,3]

[1]Laboratório de Ciclo Celular, Instituto Butantan, São Paulo 05503-900, SP, Brazil
[2]Programa Interunidades de Pós-Graduação em Bioinformática - USP, 05508-000, SP, Brazil
[3]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo 05508-000, SP, Brazil

Schistosomiasis is a disease caused by the trematode *Schistosoma mansoni*, which affects millions of people in tropical and subtropical regions and is considered a neglected tropical disease by the World Health Organization (WHO). The parasite can remain in the host for decades, evading the immune system. Currently, there are no vaccines available, and the only drug, praziquantel, is effective only against adult worms, without preventing reinfection. The efficiency of the immune evasion mechanism coordinated by the parasite suggests the possible use of alternatively-spliced microexon genes (MEGs), characterized by multiple exons of 6 to 36 nucleotides in tandem, which may be related to immune evasion mechanisms. This study aims to identify and annotate experimental evidence of alternative splicing of microexons in version 10 of the *S. mansoni* genome, using public RNA-Seq data. For this, a specific and sensitive pipeline was used to re-map RNA-Seq data to the genome (SM_V10, WormBase WBPS19) and identify microexons. The pipeline includes filtering and quality control (FastQC, fastp), followed by alignment with STAR and, in parallel, mapping with OLego. The results of both alignments are concatenated to create an annotated splicing index, which is applied in a second step with STAR, allowing the detection of both known and novel splicing sites. Parameterization tests were performed to optimize the sensitivity of the pipeline. In the first STAR pass, three groups of parameters were evaluated: “STANDARD” (least restrictive), ‘MICROEX’ (intermediate), and “MODEL” (most restrictive). At the same time, three variations of OLego were tested: with regression and annotation, with regression only, and without regression/annotation. In the second STAR pass, four groups were formed, each subdivided according to the indices. With the STANDARD parameters, an average of 204,215 junctions/sample was identified (among the three variations of the index), consistent with its lower restrictiveness and the presence of possible false positives. In the MICROEX I group, the average was 152,855 junctions/sample, and in the MICROEX II group, 155,218 junctions/sample. In the MODEL group, ~147,515 junctions/sample were detected. Transcript assembly was performed with StringTie, and the MICROEX configuration was chosen because it presented the best balance between sensitivity and efficiency. For OLego, processing using regression and annotation took 9 hours/sample, while using only regression or without regression/annotation took 2 hours/sample; thus, regression-only was chosen, considering that ~3,000 samples will be processed. Next, a filter was applied to identify MEGs, defining that: (i) maximum microexons size is 81 nt; (ii) transcripts with 3–4 internal exons must be formed only by microexons; (iii) transcripts with ≥5 internal exons must have the

largest block of tandem microexons representing at least 75% of the internal exons. Using these criteria, 70 transcripts associated with previously annotated MEGs were identified, of which 33% are new splice isoforms. The next steps include quantification with Ballgown and alternative splicing analysis with Whippet, focusing on MEGs. The work contributes to the understanding of possible molecular mechanisms of immune evasion by *S. mansoni* and points to possible new therapeutic targets.

Keywords: Microexons, MEGs, *Schistosoma mansoni*, Bioinformatics, RNA-seq, Alternative splicing

Funding: FAPESP grant 2018/23693-5; CNPq fellowship to KMVS.

References

Brasil, M. da S. (2022). *Boletim Epidemiológico - Situação epidemiológica da esquistossomose mansoni no Brasil, 2010 a 2022*. https://www.gov.br/saude/pt-br/assuntos/boletins-epidemiologicos

Brasil, M. da S. (2024). *Diretrizes técnicas -vigilância da esquistossomose mansoni*. https://bvsms.saude.gov.br/bvs/publicacoes/vigilancia_esquistossome_mansoni_diretrizes_tecnicas.pdf_1ed.pdf

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Liu, Z., Zhu, C., Steinmetz, L. M., & Wei, W. (2023). Identification and quantification of small exon-containing isoforms in long-read RNA sequencing data. *Nucleic Acids Research*, *51*(20), E104. https://doi.org/10.1093/nar/gkad810

Smith, A. D., & de Sena Brandine, G. (2021). Falco: High-speed FastQC emulation for quality control of sequencing data. *F1000Research*, 8, 1874. https://doi.org/10.12688/f1000research.21142.2

Sterne-Weiler, T., Weatheritt, R. J., Best, A. J., Ha, K. C. H., & Blencowe, B. J. (2018). Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop. *Molecular Cell*, *72*(1), 187-200.e6. https://doi.org/10.1016/j.molcel.2018.08.018

WHO. (2022). *WHO guideline on control and elimination of human schistosomiasis*. https://www.who.int/publications/i/item/9789240041608

WHO. (2023, February 1). *Schistosomiasis*. https://www.who.int/News-Room/Fact-Sheets/Detail/Schistosomiasis.

Wu, J., Anczuków, O., Krainer, A. R., Zhang, M. Q., & Zhang, C. (2013). OLego: Fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. Nucleic Acids Research, 41(10), 5149–5163. https://doi.org/10.1093/nar/gkt216

Yu, H., Li, M., Sandhu, J., Sun, G., Schnable, J. C., Walia, H., Xie, W., Yu, B., Mower, J. P., & Zhang, C. (2022). Pervasive misannotation of microexons that are evolutionarily conserved and crucial for gene function in plants. *Nature Communications*, 13(1), 820. https://doi.org/10.1038/s41467-022-28449-8

Identification of IP_3 Pathway Components in *Plasmodium falciparum* and *P. chabaudi* Using Gene Co-expression Networks

Lyang Higa Cano[1], Celia Regina da Silva Garcia[2], Ronaldo Fumio Hashimoto[1]

1. Instituto de Matemática, Estatística e Ciência da Computação, Departamento de Ciências da Computação, Universidade de São Paulo, São Paulo, Brasil.
2. Department of Clinical and Toxicological Analysis, School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil.

Plasmodium falciparum, the main causative agent of malaria, remains a global health challenge. Several *Plasmodium* species exhibit circadian rhythmicity patterns during their intraerythrocytic cycle (IEC) [1], which appears to increase parasite fitness and immune evasion [2-5]. This rhythmicity is disrupted under in vitro conditions [6], indicating that synchronization depends on signals from the host. Among the proposed host cues, melatonin, a hormone produced in a circadian manner, has been widely studied as a key regulator of parasite rhythms [7, 8]. Experimental data suggest that melatonin acts through an IP_3 signaling cascade [8], triggering intracellular calcium release and gene expression changes [8]. Although this pathway is supported by pharmacological and transcriptomic evidence [8], most of its molecular components in *Plasmodium* remain unknown [8]. In this work, we present a computational strategy to identify and investigate candidate components of the IP_3 signaling pathway. Our approach uses Gene Co-expression Networks (GCNs) [9] and a Guilt-by-Association (GBA) heuristic [10]. To reduce the false positives commonly found with GBA [11], we employ a topological strategy for threshold selection [12]. We then combined these network results with domain architecture comparisons and structural analysis. We first validated our GCN model by confirming its ability to find well-known functional modules, including components of the tubulin complex [13] and key enzymes of the glycolysis pathway [14]. We then applied the validated model to transcriptomic datasets [15, 16] from parasites either synchronized or unsynchronized with the host circadian cycle. The results suggest biologically plausible predictions, highlighting SR25 (a putative GPCR), a candidate for $G_{\alpha q}$-like protein, and MDR1 as a possible IP_3 receptor. Interestingly, the inferred pathway appears to be more closely connected to K^+-induced signaling than to direct melatonin effects [17]. To complement these results, we implemented a comparative domain analysis framework. Although no new domains were detected in the well-characterized PLC protein from *P. falciparum* [18, 19], structural comparison with the human ortholog demonstrated that the pipeline can recover biologically consistent features, supporting its application to less-characterized candidates. We also developed a method that combines our GCN model and the inferred IP_3 module to search for genes potentially related to this pathway among the 1408 genes in *P. falciparum* annotated with unknown function [20]. This is an important challenge, as about one third of the genes in the *P. falciparum* genome still lack functional annotation.

Keywords: *Plasmodium, melatonin, Ca^{2+}, IP_3 Receptor, Gene Co-expression Networks.*

References

[1] Célia R. S. Garcia, Regina P. Markus, and Luciana Madeira. “Tertian and Quartan Fevers: Temporal Regulation in Malarial Infection”. en. In: Journal of Biological Rhythms 16.5 (Oct. 2001), pp. 436–443.

[2] Philip G. McQueen and F. Ellis McKenzie. “Host Control of Malaria Infections: Constraints on Immune and Erythropoeitic Response Kinetics”. en. In: PLoS Computational Biology 4.8 (Aug. 2008). Ed. by Rob J. De Boer, e1000149.

[3] Igor M. Rouzine and F. Ellis McKenzie. “Link between immune response and parasite synchronization in malaria”. en. In: Proceedings of the National Academy of Sciences 100.6 (Mar. 2003), pp. 3473–3478.

[4] Aidan J. O’Donnell et al. “Fitness costs of disrupting circadian rhythms in malaria parasites”. en. In: Proceedings of the Royal Society B: Biological Sciences 278.1717 (Aug. 2011), pp. 2429–2436.

[5] Sarah E. Reece, Kimberley F. Prior, and Nicole Mideo. “The Life and Times of Parasites: Rhythms in Strategies for Within-host Survival and Betweenhost Transmission”. en. In: Journal of Biological Rhythms 32.6 (Dec. 2017), pp. 516–533.

[6] William Trager and James B. Jensen. “Human Malaria Parasites in Continuous Culture”. en. In: Science 193.4254 (Aug. 1976), pp. 673–675.

[7] Carlos T. Hotta et al. “Calcium-dependent modulation by melatonin of the circadian rhythm in malarial parasites.” en. In: Nature Cell Biology 2.7 (July 2000), pp. 466–468.

[8] Bárbara K. M. Dias, Abhinab Mohanty, and Célia R. S. Garcia. “Melatonin as a Circadian Marker for Plasmodium Rhythms”. en. In: International Journal of Molecular Sciences 25.14 (July 2024), p. 7815.

[9] Sipko Van Dam et al. “Gene co-expression analysis for functional classification and gene–disease predictions”. en. In: Briefings in Bioinformatics (Jan. 2017).

[10] Oliver, S. (2000). Guilt-by-association goes global. Nature, 403(6770), 601-603.

[11] Jesse Gillis and Paul Pavlidis. “ “Guilt by Association” Is the Exception Rather Than the Rule in Gene Networks”. en. In: PLoS Computational Biology 8.3 (Mar. 2012).

[12] Bhavesh R Borate et al. “Comparison of threshold selection methods for microarray gene co-expression matrices”. en. In: BMC Research Notes 2.1 (2009), p. 240.

[13] Malabika Chakrabarti, Nishant Joshi, and Geeta Kumari. “Interaction of Plasmodium falciparum apicortin with alpha- and beta-tubulin is critical for parasite growth and survival”. In: Scientific Reports 11 (2021).

[14] David D. Van Niekerk et al. “A detailed kinetic model of glycolysis in Plasmodium falciparum-infected red blood cells for antimalarial drug target identification”. en. In: Journal of Biological Chemistry 299.9 (Sept. 2023), p. 105111.

[15] Amit K. Subudhi et al. “Malaria parasites regulate intra-erythrocytic development duration via serpentine receptor 10 to coordinate with host rhythms”. en. In: Nature Communications 11.1 (June 2020), p. 2763.

[16] Filipa Rijo-Ferreira et al. “The malaria parasite has an intrinsic clock”. en. In: Science 368.6492 (May 2020), pp. 746–753.

[17] Miriam S. Moraes et al. “Plasmodium falciparum GPCR-like receptor SR25 mediates extracellular K+ sensing coupled to Ca2+ signaling and stress survival”. en. In: Scientific Reports 7.1 (Aug. 2017), p. 9545.

[18] Andreas Raabe et al. “Genetic and transcriptional analysis of phosphoinositidespecific phospholipase C in Plasmodium”. en. In: Experimental Parasitology 129.1 (Sept. 2011), pp. 75–80.

[19] Paul-Christian Burda et al. “Global analysis of putative phospholipases in Plasmodium falciparum reveals an essential role of the phosphoinositide-specific phospholipase C in parasite maturation”. en. In: mBio (July 2023).

[20] Cristina Aurrecoechea et al. “PlasmoDB: a functional genomic database for malaria parasites”. In: Nucleic Acids Research 37.suppl_1 (Jan. 2009), pp. D539–D543.

IMPACT OF ITIM1 POLYMORPHISMS ON SHP-2-LILRB2 COMPLEX

Laura Maria de Araújo Pereira[1], Silvana Giuliatti[1]
Department of Genetics, Ribeirão Preto School of Medicine, University of São Paulo, Ribeirão Preto, São Paulo, Brazil[1]

The immune system is a complex network of molecules, cells, and signaling pathways that ensure defense against pathogens and maintenance of homeostasis. Moreover, among its regulators, the receptor LILRB2 negatively modulates the immune response via immunoreceptor tyrosine-based inhibitory motifs (ITIMs), which recruit phosphatases such as SHP-2 to dephosphorylate signaling proteins and thereby inhibit cellular activation. Single-nucleotide polymorphisms (SNPs) in the LILRB2 ITIM1 region may perturb this interaction, undermining the balance of immune signaling. This study aims to analyze the structural and conformational impact of single-nucleotide polymorphisms (SNPs) in the ITIM1 region of LILRB2 and to evaluate their effects on its interaction with the phosphatase SHP-2. The full three-dimensional structure of LILRB2 was modeled using Modeller, I-TASSER, and Rosetta softwares. The tertiary structure of SHP-2 was retrieved from the Protein Data Bank (PDB ID: 2SHP). Protein–protein docking analyses were performed with HADDOCK 2.4. Complementarily, ENCoM was employed to conduct Normal Mode Analysis (NMA), and variation of Gibbs energy values (ΔΔG, in kcal·mol^{-1}·K^{-1}) were calculated via FoldX4, which also provided vibrational entropy changes (ΔS, in kcal·mol^{-1}·K^{-1}). The polymorphisms L531F (rs1182017432), L531I (rs1182017432), Y532C (rs770499234), A533T (rs367709648), A533V (rs199590986), A534D (rs2080202238), A534T (rs755507820), V535L (rs139115677) and V535M (rs139115677). Structural analyses of the LILRB2–SHP-2 complex allowed the assessment of the impact of nine polymorphisms in the ITIM1 region on protein stability and dynamics. Most mutations exhibited a neutral or slightly stabilizing effect, without significantly compromising the overall structure. However, the A534D mutation stood out as the most critical, with a highly destabilizing effect (ΔΔG = +3.13 kcal/mol) and reduced local flexibility (ΔSvib = -2.46 kcal·mol^{-1}·K^{-1}), propagating alterations to residues in the SH2 domain of SHP-2. This profile suggests impairment of LILRB2 structural integrity and its signaling potential. Among the other polymorphisms, L531I showed a neutral effect (ΔΔG = -0.45 kcal/mol), associated with loss of flexibility, while A534T was the only variant classified as slightly stabilizing (ΔΔG = -0.85 kcal/mol), although with moderate dynamic restrictions. Mutants such as L531F, Y532C, A533T, V535L, and V535M displayed ΔΔG values ranging from -1.13 to -1.40 kcal/mol, suggesting mild structural stabilization. Nevertheless, local flexibility changes, as observed for Y532C and A533T, indicate a potential modulation of the interaction with SHP-2. The combined ΔΔG and ΔSvib analyses highlight the conformational flexibility of the ITIM1 motif of LILRB2. The A534D mutation proved highly destabilizing, increasing ΔΔG and potentially impairing LILRB2–SHP-2 interaction. In contrast, mutations such as L531F, Y532C, A533T, A533V, V535L, and V535M showed negative ΔΔG values, suggesting stabilization through reduced flexibility. Since SHP-2 activation depends on phosphorylated ITIM motifs, such alterations

may compromise its recruitment and downstream signaling. Polymorphisms in LILRs have been linked to infections and autoimmune diseases, reinforcing the need for broader analyses of all ITIM motifs to clarify their role in immune regulation.

Keywords: *Coarse-Grained; polymorphisms; conformational effects; molecular docking; molecular modelling.*

Acknowledgments: CNPq, CAPES, FAEPA and SDumont Supercomputer

References

Anselmi, M. and Hub, J. S. (2020). An allosteric interaction controls the activation mechanism of SHP-2 tyrosine phosphatase. Scientific Reports, 10, 18530.

Dominguez, C., Boelens, R. and Bonvin, A. M. J. J. (2003). HADDOCK: a protein–protein docking approach based on biochemical and/or biophysical information. Journal of the American Chemical Society, 125, 1731–1737.

Frappier, V., Chartier, M. and Najmanovich, R. J. (2015). ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. Nucleic Acids Research, 43(W1), W395–W400.

Oliveira, M. L. G., Silva, F. L., Castro, M. V. and Bortolini, M. C. (2022). Genetic diversity of the LILRB1 and LILRB2 coding regions in an admixed Brazilian population sample. HLA, 100(4), 325–348.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. and Serrano, L. (2005). The FoldX web server: an online force field. Nucleic Acids Research, 33(Web Server issue), W382–W388.

Vangone, A. and Bonvin, A. M. J. J. (2015). Contact-based prediction of binding affinity in protein–protein complexes. eLife, 4, e07454.

Improving Clustering Coherence in scRNA-seq Data with Prior Knowledge and Pairwise Constraints

Davi Guimarães[1], Mateus Pereira[1,3], Kele Belloze[1], Marcel Pedroso[2], Eduardo Bezerra[1]

[1] *Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (Cefet/RJ)*
[2] *Fundação Oswaldo Cruz (Fiocruz)*
[3] *IBM Research*

The high intratumoral heterogeneity of breast cancer directly impacts patient prognosis and disease burden. Single-cell RNA sequencing (scRNA-seq) has emerged as a powerful tool to dissect this heterogeneity at the cellular level [Guo et al. 2020]. However, the large dimensionality and sparsity of scRNA-seq data demand advanced computational strategies.

In this study, we explore unsupervised and semi-supervised machine learning approaches to identify biologically relevant cellular subpopulations in breast tumors. We developed an analysis pipeline based on public scRNA-seq data (GSE75688), comprising 549 cells from 11 breast cancer patients [National Center for Biotechnology Information 2025]. The pipeline includes quality control, selection of highly variable genes (HVGs), dimensionality reduction via Principal Component Analysis (PCA), and clustering.

The experimental phase aimed to assess the impact of incorporating prior knowledge into the task of clustering single-cell gene expression data. After quality control and normalization, the top 500 HVGs were selected to capture the main signals of cellular heterogeneity. Dimensionality reduction was then performed using PCA, retaining the first 50 components to preserve the global structure of the data while mitigating the effects of high dimensionality.

Clustering was carried out by fixing the number of clusters at $k = 6$, reflecting a preliminary estimate of cellular diversity in the sample. Two algorithms were compared: standard K-Means and COP-KMeans, which incorporates supervised constraints. For COP-KMeans, 51 must-link constraints (enforcing that pairs of cells belong to the same cluster) and 75 cannot-link constraints (forcing separation of cells from different types) were used. These constraints were derived from known biological annotations, simulating a realistic semi-supervised scenario [Cai et al. 2023].

Both algorithms were executed 10 times with different random initializations to mitigate bias due to randomness and to enable a statistically robust comparison of results. Clustering quality was evaluated using the Normalized Mutual Information (NMI) met-

ric, which is widely used to quantify the similarity between algorithm-assigned labels and ground truth annotations. NMI measures the agreement between the predicted cluster assignments and the ground truth labels, accounting for chance overlap. It ranges from 0 (no mutual information) to 1 (perfect match), and is particularly suitable for evaluating unsupervised clustering methods where label alignment is not guaranteed. Higher NMI values indicate better correspondence with known cell types.

COP-KMeans achieved an average NMI of 0.4421, outperforming k-Means (average NMI = 0.4286) in alignment with known cell labels. To assess whether the COP-KMeans algorithm yields significantly higher NMI values compared to standard k-Means, we conducted a paired t-test. The results revealed a statistically significant difference between the methods ($t = 3.014$, $p = 0.0056$, one-tailed), with COP-KMeans showing superior performance. These findings support the hypothesis that incorporating constraints into the clustering process positively contributes to the quality of the resulting partitions.

We conclude that semi-supervised learning provides a promising framework for deciphering tumor heterogeneity in scRNA-seq data, offering insights into the cellular mechanisms that underlie disease progression and burden. Future analyses will investigate more powerful semi-supervised clustering algorithms, such as Deep embedded clustering [Ren et al. 2019].

Keywords: scRNA-seq, Breast cancer, Tumor heterogeneity, Semi-supervised learning.

References

[Cai et al. 2023] Cai, J., Hao, J., Yang, H., Zhao, X., and Yang, Y. (2023). A review on semi-supervised clustering. *Information Sciences*, 632:164–200.

[Guo et al. 2020] Guo, C. et al. (2020). Single-cell transcriptome analysis reveals tumor heterogeneity and tumor microenvironment in breast cancer. *arXiv preprint arXiv:2005.06692*.

[National Center for Biotechnology Information 2025] National Center for Biotechnology Information (2025). Geo accession viewer: Gse75688. `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75688`. Acesso em: 31 de julho de 2025.

[Ren et al. 2019] Ren, Y., Hu, K., Dai, X., Pan, L., Hoi, S. C., and Xu, Z. (2019). Semi-supervised deep embedded clustering. *Neurocomputing*, 325:121–130.

In silico identification of Small ORF-Encoded Peptides in long non-coding RNAs

Rafael Teixeira do Nascimento[1], Eduardo Moraes R. Reis[1], João Carlos Setubal[1].
1. Institute of Chemistry, University of São Paulo.

From the human genome are transcribed more than 60,000 lncRNAs: RNA sequences at least 200 nucleotides in length. Most of these transcripts have been annotated as non-coding RNAs, justifying their name. However, it has been shown that many of these transcripts contain unannotated open reading frames and are actually translated, with the resulting products capable of regulating biomolecular processes or producing bioactive peptides capable of controlling various cellular functions. Many small ORFs (sORFs), with 100 codons or less in length, have been identified. It has been shown that some sORF-encoded peptides (SEPs) play an essential regulatory role in various physiopathological processes and are primarily translated from lncRNAs. Confirming that sORFs in lncRNAs result in translated products is challenging since it is non-trivial to identify bona fide functional peptides. The combination of predictive techniques and nucleotide sequence-based metrics and other variables calculated from samples obtained by polysome profiling may yield more robust predictions. This work aims to discover lncRNAs encoding SEPs in diverse cancer tissue samples using machine learning techniques. The positive set used for training the models was created by combining sORF sequences associated with mass spectrometry–found peptides, obtained from the MetamORF, SPENCER, and SmProt repositories. To obtain the negative set, human lncRNA sequences from RNAcentral were classified using CPAT. ORFs predicted as non-coding were subjected to a 1% FDR filter and then BLASTed (BLASTX) against the non-redundant protein database (nr). Sequences with no matches to any protein were then selected to compose the negative set. As variables, we chose sequence-based and mathematical features, the latter calculated using the MathFeatures Python library. Preliminary results suggest that the pipeline can effectively identify SEPs in curated lncRNA transcript datasets, such as GENCODE, with good accuracy. Our curated positive datasets incorporate experimental evidence (validated peptides/sORFs) and focus specifically on ORFs with lengths equal to or less than 303 nucleotides, with extended flanking sequences to provide additional genomic context. This targeted design allows the models to capture sORF-specific features rather than whole-transcript characteristics. Preliminary results show that SVM was the best model for sORF prediction (99% AUROC, 97% accuracy, and 97% precision). A first classification test using sORFs extracted from the lncRNA set of GENCODE v.47 revealed that 51.5% of them were classified as potentially coding. The next step involves benchmarking tools designed to assess coding potential, such as CPAT and RNAsamba. Future efforts must address a critical challenge: implementing translation efficiency and ribosome occupancy–based features, in order to better characterize translated ORFs, thereby increasing the experimental basis and the reliability of the predictions. Success in these areas could establish a valuable pipeline for sORF discovery while revealing important insights into the relationship between sequence features and translation activity in putative non-coding regions.

Keywords: *non-coding RNAs, small ORF-encoded peptide, small ORF, polysome profiling.*

Acknowledgment: Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq

References

[1] Hartford, Corrine Corrina R.; Lal, Ashish. When long noncoding becomes protein coding. , In: Molecular and Cellular Biology, v. 40, n. 6, p. e00528-19, 2020.

[2] Chen Y, Ho L, Tergaonkar V. sORF-Encoded MicroPeptides: New players in inflammation, metabolism, and precision medicine. , In: Cancer Lett. 2021 Mar 1;500:263-270. doi: 10.1016/j.canlet.2020.10.038. Epub 2020 Nov 4. PMID: 33157158.

[3] Kute Pm, Soukarieh O, Tjeldnes H, Trégouët D-A And Valen E, (2022). Small Open Reading Frames, How to Find Them and Determine Their Function. , In: Front. Genet. 12:796060. doi: 10.3389/fgene.2021.796060.

[4] Sebastien A Choteau, Audrey Wagner, Philippe Pierre, Lionel Spinelli, Christine Brun, MetamORF: a repository of unique short open reading frames identified by both experimental and computational approaches for gene and metagene analyses. , In: Database, Volume 2021, 2021, baab032, https://doi.org/10.1093/database/baab032

[5] Luo X, Huang Y, Li H, Luo Y, Zuo Z, Ren J, Xie Y. Spencer: a comprehensive database for small peptides encoded by noncoding RNAs in cancer patients. , In: Nucleic Acids Res. 2022 Jan 7;50(D1):D1373-D1381. doi: 10.1093/nar/gkab822. PMID: 34570216; PMCID: PMC8728293.

[6] Li Y, Zhou H, Chen X, Zheng Y, Kang Q, Hao D, Zhang L, Song T, Luo H, Hao Y, Chen R, Zhang P, He S. SmProt: A Reliable Repository with Comprehensive Annotation of Small Proteins Identified from Ribosome Profiling. , In: Genomics Proteomics Bioinformatics. 2021 Aug;19(4):602-610. doi: 10.1016/j.gpb.2021.09.002. Epub 2021 Sep 15. PMID: 34536568; PMCID: PMC9039559.

[7] Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P., & Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. , In: Nucleic Acids Research, 41(6), e74. doi:10.1093/nar/gkt006.

[8] Robson P Bonidia, Douglas S Domingues, Danilo S Sanches, André C P L F de Carvalho, MathFeature: feature extraction package for DNA, RNA and protein sequences based on mathematical descriptors. , In: Briefings in Bioinformatics, Volume 23, Issue 1, January 2022, bbab434, https://doi.org/10.1093/bib/bbab434.

Mesoscopic model application to 2-O-methylribose/RNA hybrids reveals base fraying and salt concentration effects

Daniel Batista de Jesus[1], Gerald Weber[1]

1. Departamento de Física- Universidade Federal de Minas Gerais

The 2'-O-methylribose (Nm) is a natural post-transcriptional modification of RNA, which results from the addition of a methyl group at the 2' position of ribose. The Nm modification has structural effect in ribose and shifts the C2'/C3'-endo equilibrium, on which the increased stability could depend. X-ray diffraction indicates that the structure of the hybrids is similar to that of RNA/RNA A-type helices. This type of oligonucleotide has been used in RNA probing due to the higher stability and faster hybridization of Nm/RNA in comparison to unmodified DNA/RNA and RNA/RNA hybrids. The application of mesoscopic models can be useful to evaluate intermolecular interactions and predict quantities of experimental interest, such as the melting temperature. We applied the Peyrard-Bishop (Peyrard, 1989) non-linear mesoscopic model with a computational approach to obtain the parametrization of Nm/RNA hybrid sequences using published melting temperatures under low (LS) and high salt (HS) conditions. The Nelder-Mead simplex downhill minimization algorithm was used for the parameter optimization. To reduce the number of parameters, rules of nearest-neighbour equivalence were applied to the sequence set. Several steps of nested minimizations were performed to guarantee the statistical significance of the results. The panel used is composed of 70 sequences of 5 to 9 base pairs long in LS and 38 in from 5 to 9 base pairs long in HS condition (Kierzek et al 2005, 2006, 2009). The TM predictions obtained for the training set achieved high accuracy, with χ^2 function values of 59 and 77 for LS and HS, and corresponding quadratic mean deviations of 0.92 and 1.46, respectively. The predictions made for the test set with NM model parameters support the success of the optimization showing high accuracy, as well. The comparison to nearest-neighbor models prediction for RNA/RNA show a non-trivial Tm behavior captured by our results. Our results show a clear base fraying effect, that is, the pronounced base pair separation at the duplex termini. Unlike the RNA(Ferreira et al, 2021) the r(A)/Nm(U) has more stable hydrogen bonds in the terminal position than in the internal counterpart under LS conditions, which was not observed under HS conditions. Our results also show an important effect of salt concentration, with a similar hydrogen bond strength for methylation in purine or pyrimidine internal base pairs under high salt conditions. We discuss the Nm on end fraying and the predictions to internal positions across ionic strengths. The software is available in TfReg and VarPar package on-line (Gerald Weber - TfReg: Calculates melting temperatures of DNA and RNA with mesoscopic models), and the data in the cited references.

Funding agencies: CAPES, CNPq and Fapemig.

Keywords: Peyrard-Bishop model, 2’-O-methylribose, Statistical Physics, Simplex downhill algorithm, Epigenomics.

References

Ferreira, I., Amarante, T. D., & Weber, G. (2021). Salt dependent mesoscopic model for RNA at multiple strand concentrations. *Biophysical Chemistry*, *271*, 106551.

Kierzek, E., Ciesielska, A., Pasternak, K., Mathews, D. H., Turner, D. H., & Kierzek, R. (2005). The influence of locked nucleic acid residues on the thermodynamic properties of 2′-O-methyl RNA/RNA heteroduplexes. *Nucleic acids research*, *33*(16), 5082-5093.

Kierzek, E., Mathews, D. H., Ciesielska, A., Turner, D. H., & Kierzek, R. (2006). Nearest neighbor parameters for Watson–Crick complementary heteroduplexes formed between 2′-O-methyl RNA and RNA oligonucleotides. *Nucleic acids research*, *34*(13), 3609-3614.

Kierzek, E., Kierzek, R., Turner, D. H., & Catrina, I. E. (2006). Facilitating RNA structure prediction with microarrays. *Biochemistry*, *45*(2), 581-593.

Kierzek, E., Pasternak, A., Pasternak, K., Gdaniec, Z., Yildirim, I., Turner, D. H., & Kierzek, R. (2009). Contributions of stacking, preorganization, and hydrogen bonding to the thermodynamic stability of duplexes between RNA and 2′-O-methyl RNA with locked nucleic acids. *Biochemistry*, *48*(20), 4377-4387.

Peyrard, M., & Bishop, A. R. (1989). Statistical mechanics of a nonlinear model for DNA denaturation. *Physical review letters*, *62*(23), 2755.

Molecular Modeling as a Strategy to Specific Antibody Design for ApoE4's Carries

Victor H. O. Andrade[1], Silvana Giuliatti[1]

1. Department of Genetics, Faculty of Medicine of Ribeirao Preto, University of Sao Paulo, USP, Ribeirao Preto, Brazil.

Alzheimer's disease (AD) is a neurodegenerative condition characterized by the accumulation of amyloid-beta (Aβ) plaques in the brain (Scheltens et al., 2021). Apolipoprotein E (ApoE) has an important physiological function, helping lipid transport across cells and organs (Blumenfeld et al., 2024). ApoE promotes Aβ clearance by forming complexes with it and promoting its removal from the brain's extracellular space. This process occurs primarily through interactions with blood-brain barrier (BBB) receptors, such as LRP1 (Low-Density Lipoprotein Receptor-Related Protein 1). In particular, the ApoE4 isoform critically influences Aβ aggregation and clearance dynamics and is the strongest genetic risk factor for sporadic AD (Blumenfeld et al., 2024). Studies suggest that the nonlipidated form of ApoE4 is associated with a greater predisposition to AD, making it a preferential target for therapeutic antibodies like HAE-4, which showed significant reduction in plaque Aß deposition in mice models with APPPS1-21/ApoE4 phenotype (Liao et al., 2018; Poblano et al., 2024). Given the need for more selective and effective immunotherapies, this research looks to find unique conformational epitopes of nonlipidated ApoE4, that are distinct from those in ApoE3 and lipidated ApoE4, for the development of novel therapeutic antibodies. Targeting nonlipidated ApoE4 is strategically justified by its documented structural instability, increased propensity to form pathogenic aggregates, and enhanced affinity for Aβ. This study will employ structural bioinformatics approaches such as comparative modeling, molecular dynamics simulations, and conformational epitope prediction, with physicochemical accessibility and immunogenicity analysis to show novel conformational epitope candidates, advancing our understanding of ApoE4's pathological mechanisms and supporting the development of targeted antibody therapies for AD. In structural modeling, the ApoE3 and ApoE4 proteins were constructed by comparative modeling using the MODELLER software (Webb and Sali, 2016), while the Amyloid Beta 42 peptide was modeled by an ab initio approach using QUARK Web Server (Xu and Zhang, 2012). The ApoE3 model was generated using the structure number 1 of the 2L7B PDB model as a template. For ApoE4 modeling, both the newly constructed ApoE3 model and the 1GS9 PDB structure served as templates in the comparative modeling approach. These new structures were sent to stereochemical analysis for quality evaluation in PROCHECK (Laskowski et al., 1996), VERIDY3D (Luthy et al., 1992) and ERRAT (Colovos and Yeates, 1993) platforms. For comparison, the structure 2L7B was used as a quality reference for ApoE models. The models showed superior stereochemical quality compared to reference, with over 86% of residues found in the most favored regions of the Ramachandran diagram (Ramachandran et al., 1963). About the modeling of Aβ, the peptide adopts a conformation rich in β-sheets and a short α-helix, consistent with the structural characteristics reported by (Gillet, 2021). The next steps of this research involve modeling a VLDL-mimetic

lipid nanoparticle (LNP), generating ApoE/Aβ and ApoE/Aβ/LNP complexes with different isoforms, and performing molecular dynamics simulations with them. It is expected that this will lead to discover structural epitopes candidates to antibody ApoE4's specific design, contributing for the treatment of DA to ApoE4 carriers.

Keywords: neurogenerative disease, apolipoprotein E, molecular dynamics, molecular modeling, antibody design

References

Blumenfeld, J., Yip, O., Kim, M. J., and Huang, Y. (2024). Cell type-specific roles of APOE4 in Alzheimer disease. *Nature Reviews Neuroscience* |, *25*, 91–110. https://doi.org/10.1038/s41583-023-00776-9

Colovos, C., and Yeates, T. O. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Science : A Publication of the Protein Society*, *2*(9), 1511–1519. https://doi.org/10.1002/PRO.5560020916

Gillet, J.-N. (2021). Alzheimer's disease: unraveling APOE4 binding to amyloid-beta peptide and lipids with molecular dynamics and quantum mechanics. *Journal of Biomolecular Structure and Dynamics*, *39*(14), 5026–5032. https://doi.org/10.1080/07391102.2020.1784287

Laskowski, R. A., Rullmann, J. A. C., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996). AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *Journal of Biomolecular NMR*, *8*(4), 477–486. https://doi.org/10.1007/BF00228148

Liao, F., Li, A., Xiong, M., Bien-Ly, N., Jiang, H., Zhang, Y., Finn, M. B., Hoyle, R., Keyser, J., Lefton, K. B., Robinson, G. O., Serrano, J. R., Silverman, A. P., Guo, J. L., Getz, J., Henne, K., Leyns, C. E. G., Gallardo, G., Ulrich, J. D., … Holtzman, D. M. (2018). Targeting of nonlipidated, aggregated apoE with antibodies inhibits amyloid accumulation. *The Journal of Clinical Investigation*, *128*(5), 2144–2155. https://doi.org/10.1172/JCI96429

Luthy, R., Bowie, J. U., and Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, *359*, 83–85.

Poblano, J., Castillo-Tobías, I., Berlanga, L., Tamayo-Ordoñez, M. C., del Carmen Rodríguez-Salazar, M., Silva-Belmares, S. Y., Aguayo-Morales, H., and Cobos-Puc, L. E. (2024). Drugs targeting APOE4 that regulate beta-amyloid aggregation in the brain: Therapeutic potential for Alzheimer's disease. *Basic and Clinical Pharmacology and Toxicology*, *135*(3), 237–249. https://doi.org/10.1111/BCPT.14055

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, *7*(1), 95–99. https://doi.org/10.1016/S0022-2836(63)80023-6

Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., Teunissen, C. E., Cummings, J., and van der Flier, W. M. (2021). Alzheimer's disease. *Lancet (London, England)*, *397*(10284), 1577–1590. https://doi.org/10.1016/S0140-6736(20)32205-4

Webb, B., and Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics*, *54*, 5.6.1-5.6.37. https://doi.org/10.1002/CPBI.3

Xu, D., and Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics*, *80*(7), 1715–1735. https://doi.org/10.1002/prot.24065

Multivariate conformational patterns may distinguish HLA-DRB1 alleles associated with Multiple Sclerosis susceptibility

Levy Bueno Alves [1], Silvana Giuliatti [1]

1. Department of Genetics, Faculty of Medicine of Ribeirão Preto, University of São Paulo, USP, Ribeirão Preto, Brazil.

Multiple sclerosis (MS) is an autoimmune disease mediated by T-cell responses to central nervous system autoantigens. Genetic factors play a significant role in its etiology, with the HLA locus standing out as the primary region associated with susceptibility. Certain HLA-DRB1 alleles are correlated with both risk and protection, although the mechanisms underlying this divergence remain poorly understood. Polymorphic residues in the DRB1 binding groove are crucial for peptide specificity and recognition by T-cell receptors, but the impact of conformational variations in this groove remains poorly explored. In this study, molecular dynamics (MD) simulations were conducted to investigate the intrinsic structural plasticity of HLA-DR allelic variants in the peptide-free form, seeking to identify topologies capable of differentiating the dynamic behavior of predisposing and protective alleles. To perform the MD simulations, the methods involved obtaining the crystallographic structure of the DRB1*15:01 allele, modeling missing residues, and constructing five additional allelic isoforms of DRB1, in addition to DRB5*01:01. The resulting seven heterodimers were inserted into lipid bilayers using the CHARMM-GUI server and subjected to 500 ns simulations with GROMACS. The trajectories were analyzed through RMSD, RMSF, molecular volume, free energy landscape (FEL), and binding energy calculations using the MM/PBSA method. The complexes showed structural stability, with an average RMSD of 2.2 Å and approximate volumes of 28,622 Å³. The ΔRMSF, calculated for the low-risk allele DRB1*01:01, showed greater mobility in the risk-associated variants, especially DRB1*15:01 and DRB1*15:03, whose positive fluctuation areas of the polymorphic chain were higher. However, the heterogeneous distribution of fluctuations along the chains points to the existence of complex and multivariate dynamic patterns. FEL analysis indicated that protective alleles share low-energy conformational regions, suggesting functional convergence. In contrast, DRB1*15:01 exhibited a more dispersed energy landscape, sharing no metastable states with DRB1*15:03, but partially overlapping with DRB5*01:01, present in the same haplotype. Binding energy calculations revealed comparable affinities between DRB1 variants (≈ –162 kcal/mol), indicating similar stability at the interface with the non-polymorphic chain. Taken together, the results suggest that greater structural flexibility and conformational diversity characterize the alleles associated with MS risk, possibly contributing to a greater capacity for antigen self-presentation or escape from central tolerance. Conversely, the conformational convergence observed among the protective alleles may reflect functional constraints that limit the activation of autoimmune responses. Furthermore, the heterogeneous distribution of fluctuations observed in the ΔRMSF curves, particularly in polymorphic regions, highlights distinct patterns of mobility between risk and protective alleles. This variability, coupled with subtle differences

in local dynamics, suggests the existence of conformational patterns, the elucidation of which requires approaches capable of integrating multiple structural and temporal variables extracted from the trajectories. Given the observed conformational complexity, supervised machine learning models are being developed to identify dynamic signatures that differentiate risk and protective alleles. The results of this research are expected to contribute to the advancement of the mechanistic understanding of the structural bases of susceptibility to MS, in addition to providing support for future investigations into the mechanisms of antigen presentation and recognition by T cells.

Keywords: Autoimmunity, HLA-DRB1, polymorphisms, structural bioinformatics.

Acknowledgements: CNPq, CAPES, FAEPA, and Santos Dumont Supercomputer.

References

De Silvestri, A. et al. (2019). The Involvement of HLA Class II Alleles in Multiple Sclerosis: A Systematic Review with Meta-analysis. In *Disease Markers*, pages 1-8 (Vol. 2019).

Goris, A. et al. (2022). Genetics of multiple sclerosis: lessons from polygenicity. In *The Lancet Neurology*, pages 830-842 (Vol. 21, Issue 9).

Hollenbach, J. A. and Oksenberg, J. R. (2015). The immunogenetics of multiple sclerosis: A comprehensive review. In *Journal of Autoimmunity*, pages 13-25 (Vol. 64).

Optimization of genomic prediction in young nellore bulls: balancing production and resilience for tropical cattle production

Leonardo Augusto Kohara Melchior[1]

1. Programa de Pós-Graduação em Ciências da Saúde na Amazônia Ocidental, Centro de Ciências Biológicas e da Natureza, Universidade Federal do Acre.

Introduction: Brazil's beef cattle industry, sustained mainly by the Nellore breed, has achieved significant productivity gains, largely due to genetic selection and tools such as genomic selection (GS). However, selection with a heavy emphasis on production traits, such as weight gain and carcass yield, can result in animals that are efficient under ideal conditions but poorly adapted to the real challenges of production systems. Although the Nellore breed is known for its hardiness, adaptation is a spectrum, not a binary state. In other words, genetic variations exist within the breed that confer varying degrees of thermotolerance or better utilization of local forages to certain groups of individuals. **Objective**: The study aims to develop and validate a biologically informed, weighted genomic prediction model (weighted G-BLUP) to increase the accuracy of selecting young Nellore bulls for a genetic merit that balances both production and resilience. **Materials and Methods**: The computational methodology will be executed in two phases. In Phase 1 (Data Mining), a panel of adaptation markers will be created from the reanalysis of public genomic data. Population differentiation statistics, such as FST (Fixation Index) and XP-EHH (Cross-Population Extended Haplotype Homozygosity), will be used to compare the genomes of *Bos indicus* and *Bos taurus*, identifying regions under strong divergent selection associated with tropical adaptation. In Phase 2 (Model Implementation and Validation), the identified adaptation markers will be assigned a higher weight in the wG-BLUP prediction model, reflecting their greater biological relevance. To test the approach's effectiveness, the predictive accuracy of the weighted model (wG-BLUP) will be compared to that of the standard model (G-BLUP), which does not use weighting. This comparison will be performed using the k-fold cross-validation methodology on a Nellore reference dataset containing genotypic and phenotypic information from thousands of animals. The correlation between the predicted and observed genetic merit will be the primary criterion for determining the model's superiority. **Expected Results and Impact**: It is expected that the weighted model will demonstrate significantly higher predictive accuracy, especially for a selection index that includes resilience traits. The validation of this methodology will provide the sector with a more robust and intelligent tool to identify genetically superior young sires, not only in terms of productive potential but also in adaptation and resilience. This breakthrough represents a fundamental step towards developing a more sustainable, efficient, and adapted beef cattle industry, especially for challenging biomes such as the Amazonian.

Keywords: Adaptation, *Bos indicus*, Genomic Selection, SNPs, wG-BLUP.

References

Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4), 1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Neves, H. H., Carvalheiro, R., O'Brien, A. M., Utsunomiya, Y. T., do Carmo, A. S., Schenkel, F. S., Sölkner, J., McEwan, J. C., Van Tassell, C. P., Cole, J. B., da Silva, M. V., Queiroz, S. A., Sonstegard, T. S., & Garcia, J. F. (2014). Accuracy of genomic predictions in Bos indicus (Nellore) cattle. Genetics, selection, evolution : GSE, 46(1), 17. https://doi.org/10.1186/1297-9686-46-17

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., … Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature, 449(7164), 913–918. https://doi.org/10.1038/nature06250

Lopez, B. I., Lee, S. H., Park, J. E., Shin, D. H., Oh, J. D., de Las Heras-Saldana, S., van der Werf, J., Chai, H. H., Park, W., & Lim, D. (2019). Weighted Genomic Best Linear Unbiased Prediction for Carcass Traits in Hanwoo Cattle. Genes, 10(12), 1019. https://doi.org/10.3390/genes10121019

Albuquerque, L. G., & Meyer, K. (2001). Estimates of direct and maternal genetic effects for weights from birth to 600 days of age in Nelore cattle. Journal of Animal Breeding and Genetics, v. 118, n. 2, p. 83-92.

Simple yet robust annotation of homologous and non-homologous isofunctional enzymes

Emanuel Umbelino [2], Alexander da Franca Fernandes [1], Sérgio Lifschitz [2], Edward H. Haeusler [2], Ana Carolina Ramos Guimarães [1], Marcos Catanho [1] & Antonio Basílio de Miranda [1]

1. Instituto Oswaldo Cruz, Fiocruz, Brazil
2. Pontifícia Universidade Católica do Rio de Janeiro, Brazil

Convergent evolution - the independent emergence of similar traits in unrelated lineages - has been treated as a rare exception to the "no reinvention of the wheel" principle, often relegated to the periphery of evolutionary thinking, requiring more rigorous standards of evidence and frequently viewed with scepticism. However, convergence is not an anomaly but rather a pervasive phenomenon (Wu et al., 2020), reflecting the constrained nature of biochemical solutions to life's functional demands. Among the most compelling examples of molecular convergence are non-homologous isofunctional enzymes (NISE), distinct proteins with no detectable common ancestry that catalyze the same biochemical reaction. These enzymes differ in structure, domain architecture, or mechanism, yet fulfil equivalent roles within metabolic networks. When homologous enzymes (HISE) exist alongside non-homologous counterparts, the functional landscape becomes even more complex, with multiple genetic routes to the same metabolic endpoint. In this work (still in progress), we present a simple yet robust approach to identify and characterize HISE and NISE in protein sequence databases in which entries have assigned enzymatic activity (EC) and SUPERFAMILY (SF) domain. The SF database (Pandurangan et al., 2019) uses Hidden Markov Models to assign protein domains from SCOP superfamilies (Andreeva et al., 2020) to sequences across many completely sequenced genomes. Each SCOP superfamily represents a group of proteins that are believed to have a common evolutionary origin, even if their sequences have diverged significantly, based on structural data. Indeed, SF classification has been proved effective in discriminating between NISE and HISE even in the absence of 3D-structure information (Piergiorge et al., 2017). As a proof of concept, we applied a preliminary Perl implementation of this method to identify NISE and HISE in the UniProtKB reviewed (Swiss-Prot) database (UniProt Consortium, 2025). We found enzymatic activities carried out by structurally distinct proteins across 3,234 organisms in all three domains of life, including viruses - Archaea=116; Bacteria=1,417; Eukaryota=1,233; Viruses=467, and 338 enzymatic activities performed by structurally unrelated enzymes, with distinct SF domain architectures, comprising 32,895 sequences (oxidoreductases=3,305; transferases=10,937; hydrolases=11,463; lyases=3,876; isomerases=1,896; ligases=835; translocases=583) in 846 clusters representing the total number of independent origins within the 338 enzymatic activities. The workflow consists in a few steps yielding structured, parseable outputs: monomeric and multimeric (single and multiple SF domain, respectively) NISE and HISE, among others. The code is being reimplemented in the Python programming language, and a stand-alone application will soon be freely available to the scientific community.

Keywords: *convergence, enzymatic activity, non-homologous isofunctional enzymes*

References

Andreeva A, Kulesha E, Gough J, Murzin AG. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res. 2020 Jan 8;48(D1):D376-D382. doi: 10.1093/nar/gkz1064. PMID: 31724711; PMCID: PMC7139981.

Pandurangan AP, Stahlhacke J, Oates ME, Smithers B, Gough J. The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. Nucleic Acids Res. 2019 Jan 8;47(D1):D490-D494. doi: 10.1093/nar/gky1130. PMID: 30445555; PMCID: PMC6324026.

Piergiorge RM, de Miranda AB, Guimarães AC, Catanho M. Functional Analogy in Human Metabolism: Enzymes with Different Biological Roles or Functional Redundancy? Genome Biol Evol. 2017 Jun 1;9(6):1624-1636. doi: 10.1093/gbe/evx119. PMID: 28854631; PMCID: PMC5737724.

UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res. 2025 Jan 6;53(D1):D609-D617. doi: 10.1093/nar/gkae1010. PMID: 39552041; PMCID: PMC11701636.

Wu CI, Wang GD, Xu S. Convergent adaptive evolution-how common, or how rare? Natl Sci Rev. 2020 Jun;7(6):945-946. doi: 10.1093/nsr/nwaa081. Epub 2020 May 12. PMID: 34692115; PMCID: PMC8288862.

SynDRA: Synonym Drug Repurposing Alignment

Tolga Corbaci [1,2,3], Katjuša Koler [4], Erlend Skaga [5], Pourya Naderi Yeganeh [1,2], Winston Hide [1,2]

1. *Department of Pathology, Beth Israel Deaconess Medical Center, Boston MA 02115, USA.*
2. *Harvard Medical School, Boston, MA 02115, USA.*
3. *School of Medicine, Bahçeşehir University, Istanbul, Turkey*
4. *University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom*
5. *Department of Neurosurgery, Vilhelm Magnus Laboratory for Neurosurgical Research, Oslo University Hospital, Oslo, Norway*

Mapping drug names consistently across biomedical datasets remains a major obstacle in computational drug repurposing. Variations in naming conventions, abbreviations, and chemical descriptors cause mismatches that reduce model accuracy and hinder integration between resources. This bottleneck is especially problematic when linking literature-derived or AI-predicted hits to transcriptomic data from platforms such as the Library of Integrated Network-Based Cellular Signatures (LINCS) and its Connectivity Map (CMap) resource. LINCS is a large-scale resource that catalogs gene expression responses to small-molecule perturbations across human cell lines. CMap relies on Broad Institute's standardized internal identifiers (BRD_IDs) to enable meaningful comparisons between transcriptional signatures, yet many drug names from external databases fail to map cleanly to these identifiers.

We present **SynDRA (Synonym Mapping for Alignment of Repurposing Therapeutics)**, a unified drug synonym mapping system designed to bridge drug names across multiple widely-used platforms. SynDRA integrates and harmonizes four major resources: the Broad Institute's Drug Repurposing Hub, TTD (Therapeutic Target Database), PRISM, and LINCS2020 metadata into a deduplicated synonym-to-ID dictionary. The pipeline performs normalization, cleaning, synonym explosion, and outer joins to propagate identifiers across shared names. Entries lacking valid BRD_IDs were excluded to maintain relevance for transcriptomic perturbation studies.

The final dataset contains **193,113 unique synonyms across 33,858 BRD_IDs**, with mappings to **2,775 TTD_IDs** and **950 PubChem CIDs**. To test practical coverage, we benchmarked our mapping approach against a curated set of 527 cancer drugs from the FIMM FO5A collection, a widely used drug screening library described in *a* 2024 *Nature Protocols* publication by Chen et al. Direct matching with LINCS compound names resulted in 309 successful matches (58.63%). When using SynDRA, the match rate increased to 354 out of 527 drugs (67.17%), enabling broader downstream integration with CMap analyses and target-based drug repurposing workflows.

SynDRA enhances compatibility between external drug sources and LINCS-CMap, streamlining repurposing studies and improving the interpretability of transcriptomic

signatures. The SynDRA database, source code, and Shiny app are available at: https://github.com/hidelab/SynDRA.

Keywords: **Drug repurposing, Drug Synonyms, LINCS Connectivity Map (CMap), Transcriptomic signatures, Synonym mapping**

References

Chen, Y., He, L., Ianevski, A., Ayuda-Durán, P., Potdar, S., Saarela, J., Miettinen, J. J., Kytölä, S., Miettinen, S., Manninen, M., Heckman, C. A., Enserink, J. M., Wennerberg, K. and Aittokallio, T. (2024). Robust scoring of selective drug responses for patient-tailored therapy selection. *Nature Protocols*, 19(1), p. 60–82.

Keenan, A. B., Jenkins, S. L., Jagodnik, K. M., Koplev, S., He, E., Torre, D., Wang, Z., Dohlman, A. B., Silverstein, M. C., Lachmann, A., Kuleshov, M. V., Ma'ayan, A., et al. (2018). The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. *Cell Systems*, 6(1), p. 13–24.

Koler, K. (2020). *A systematic pathway-based network approach for in silico drug repositioning*. Doctoral dissertation, University of Sheffield.

Zhou, Y., Zhang, Y., Zhao, D., Yu, X., Shen, X., Zhou, Y., Wang, S., Qiu, Y., Chen, Y. and Zhu, F. (2024). TTD: Therapeutic Target Database describing target druggability information. *Nucleic Acids Research*, 52(D1), p. D1465–D1477.

Towards a Roadmap for Translational Bioinformatics in the Era of AI and Personal Data Governance

Luana F. Alvarenga[1], Leonardo C. T. Bezerra[2], Renan C. Moioli[1], César Rennó-Costa[1]

[1]Centro Multiusuário de Bioinformática (BIOME) Instituto Metrópole Digital (IMD), Universidade Federal do Rio Grande do Norte (UFRN), Natal, RN, Brazil
[2]Division of Computing Science and Mathematics, University of Stirling, Cottrell Building, Stirling FK9 4LA, United Kingdom

The translation of bioinformatics innovations, particularly those leveraging Artificial Intelligence (AI), from academic research to clinical application faces significant hurdles. These barriers stem from the convergence of rapid AI advancements and an increasingly complex and fragmented tripartite regulatory landscape, which encompasses medical device regulations, emerging AI-specific laws, and stringent data protection mandates, such as the GDPR. This environment creates substantial uncertainty, heightening the risks of non-compliance, project failure, and ethical oversights, ultimately slowing the delivery of beneficial technologies to patients.

This work addresses these challenges by proposing a strategic roadmap designed to mitigate risks and accelerate the transition from bench to bedside for AI-driven bioinformatics. The proposed roadmap is built on three core pillars. First, it advocates for a "Regulation-by-Design" approach, which mandates the integration of regulatory strategy into the earliest stages of research and development, treating compliance as a core design principle rather than a final hurdle. Second, it introduces the "FAIR-T" framework, an extension of the FAIR data principles (Findable, Accessible, Interoperable, Reusable) by adding "T" for "Trustworthy." This framework enriches datasets with essential metadata on provenance, consent, bias, and ethical review, creating a foundation for robust and ethically sound AI models. Third, the roadmap calls for the cultivation of the "Translational Bioinformatician," a new professional profile with expertise spanning computational science, regulatory affairs, data ethics, and commercialization, requiring a fundamental redesign of academic curricula.

By implementing this integrated strategy, the bioinformatics community can navigate the modern innovation landscape more effectively. This approach aims to foster a more predictable and responsible ecosystem, ensuring that the transformative potential of AI in healthcare is realized safely, ethically, and efficiently, ultimately bridging the gap between scientific discovery and tangible clinical impact.

Keywords: *Translational Bioinformatics, Artificial Intelligence (AI), Regulation, Health Innovation, Data Governance*

Type VI Secretion System in *Pseudomonas aeruginosa* – distribution and comparative analysis

Hadassa Loth de Oliveira[1], Graciela Maria Dias[2], Bianca Cruz Neves[1]

1. Instituto de Química, Universidade Federal do Rio de Janeiro, Rio de Janeiro/RJ
2. Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro, Rio de Janeiro/RJ

Bacteria possess a variety of protein secretion systems that are crucial for interaction with the environment, with host organisms and other members of the microbiota. Recently, much attention has been focused on the Type VI Secretion System (T6SS), which is widely distributed in Proteobacteria. There is a wide range of processes associated with them, grouped into T6SS subclasses, involved in specific functions. Some T6SSs mediate competitions by secreting toxins, which reach specific target cells, or even provide effectors that promote survival in adverse environments by acquiring metal ions. *Pseudomonas aeruginosa*, considered a model in T6SS studies, is a ubiquitous microorganism due to its metabolic versatility and wide adaptability to different conditions. Studies on microbial communities in extreme environments, such as oil industry systems, are essential for the control of many technical, operational and environmental issues. In this perspective, we highlight the PA1-Petro strain, isolated from oil production water in Northeast Brazil, as well as other *P. aeruginosa* strains of environmental origin. The objective of this work is to better understand the gene repertoire of *P. aeruginosa* strains, focusing on their differentials and T6SS, through the pangenome of 203 *P. aeruginosa* strains of environmental classification, according to the IPCD database, in 2019. The pangenome was generated with the GET_HOMOLOGUES v3.4 program, its algorithms and tools, the results were analyzed with COG, 2020 and with SecReT6 v3.0. The pangenome totalized 18,406 clusters. The most abundant COG categories in the pangenome comprises transcription, signal transduction, transport and metabolism. The largest repertoire of T6SS proteins (structural, regulatory and effector) are found in the cloud partition, where proteins are present in up to two genomes. The most present T6SS and associated proteins were ClpV, LadS, TagS, pqsA, pqsE, hcp2, VgrG2b, VgrG1b and Tle3. Given the results, we can conclude that T6SS is structurally conserved in *P. aeruginosa*, differing in the amount of orphan proteins in each genome, results that add to the high genomic plasticity, versatility and metabolic robustness of the species. This intraspecific diversity of the T6SS in *P. aeruginosa* raises important evidence about the complexity of its interactions with other bacteria and with the human host.

Keywords: *Pseudomonas aeruginosa*, Type VI Secretion System, Comparative Genomic, Pangenome, Bioinformatics.

Uncovering the hidden structure of subspecies in large metagenomic datasets

Arthur Henrique Barrios Solano [1], João Carlos Setubal [1, 2]

1. Inter-Departmental Graduate Program in Bioinformatics, USP
2. Department of Biochemistry, Institute of Chemistry, USP

We introduce the concept of species subgroups as a means to better capture taxonomic classification of genomes, contigs and reads below the species level. Subgroups are clusters of genomes that share a minimum of 98% of similarity on average, a threshold higher than the 95% used for the species level definition [Jain et al. 2018]. Subgroups are obtained by the metagenomic classification pipeline BH [Solano & Setubal 2024], which uses the model of Maximal Independent Set to generate a reduced and non redundant set of genomes as a Genomes Reference Set (GRS) for investigation of species from a target genus via BLASTn comparisons. We demonstrate the subgroup approach using *Acinetobacter baumannii* as an example. *A. baumannii* strains are one of the most worrisome human pathogens due to the fact that these strains have become resistant to antibiotics, posing a major public health concern. The GRS we have generated for this species based on 5.435 RefSeq genomes classified as *A. baumannii* contains 25 subgroups with 20 genomes or more. For these 25 subgroups, we show their positioning and structure in a clustermap based on the average distance of 1-ANI. Aiming to relate the subgroups with the Clonal Complexes (CCs) concept commonly used to obtain subspecies resolution for *A. baumannii* isolates [Müller et al. 2023], we have applied a Multi-locus Sequence Typing analysis with MLST (https://github.com/tseemann/mlst) and the Pasteur scheme [Diancourt et al. 2010], based on the allelic profiling of seven housekeeping genes, resulting in a map between each subgroup and an exclusive set of CCs, with few exceptions. We then applied the BH tool along with the *A. baumannii* GRS on three global metagenomic datasets: MetaSUB [Danko et al. 2021], cFMD [Carlino et al. 2024], and Global Sewage [Hendriksen et al. 2019]. The most abundant subgroups were *A. baumannii* - 270, *A. baumannii* - 289 and *A. baumannii* - 287. In the MetaSUB dataset, these three subgroups were found in 34%, 23% and 78% of the cities sampled, respectively; in the cFMD dataset, these subgroups were found in 36%, 23% and 41% of the countries sampled, respectively; lastly, in the Global Sewage dataset, these subgroups were found in 45%, 71% and 89% of the countries sampled, respectively. These results suggest that *A. baumannii* - 289 is an worldwide disseminated subgroup, while the other two subgroups were highly abundant but less disseminated. Finally, using Abricate (https://github.com/tseemann/abricate) we searched for genes that encode OXA-type carbapenemases, a group of enzymes associated with the resistance to carbapenems [Müller et al 2023], in each of the most abundant subgroups. All the genomes in the subgroup 287 harbored *blaOXA-51*; 50% of the genomes in the subgroup 289 harbored both *blaOXA-58* and *blaOXA-23*, while 51% of the genomes in subgroup 270 conserved the *blaOXA-23.*

Acknowledgements: This work was made possible in part by a FAPESP PhD fellowship to A.H.B.S. (award #2024/01729-9) and by a grant from CNPq (award #440230/2022-5). *Correspondence: arthur.barrios@usp.br*

Keywords: *Subgroups, Acinetobacter baumannii, carbapenems resistance, metagenomics*

References

Carlino, Niccolò, et al. "Unexplored microbial diversity from 2,500 food metagenomes and links with the human microbiome." Cell 187.20 (2024): 5775-5795.

Danko, D., Bezdan, D., Afshin, E. E., Ahsanuddin, S., Bhattacharya, C., Butler, D. J., Chng, K. R., Donnellan, D., Hecht, J., Jackson, K., et al. (2021). A global metagenomic map of urban microbiomes and antimicrobial resistance. Cell, 184(13):3376–3393.

Diancourt, L., Passet, V., Nemec, A., Dijkshoorn, L. & Brisse, S. The population structure of Acinetobacter baumannii: expanding multiresistant clones from an ancestral susceptible genetic pool. PLoS ONE 5, e10034 (2010).

Hendriksen, R. S. et al. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nat. Commun. 10, 1124 (2019).

Müller C, Reuter S, Wille J, Xanthopoulou K, Stefanik D, Grundmann H, Higgins PG, Seifert H.2023.A global view on carbapenem-resistant Acinetobacter baumannii. mBio14:e02260-23.https://doi.org/10.1128/mbio.02260-23

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ani analysis of 90k prokaryotic genomes reveals clear species bound-aries. Nature communications, 9(1):5114.

Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. 2018 Sep 24;3:124. doi: 10.12688/wellcomeopenres.14826.1. PMID: 30345391; PMCID: PMC6192448.

Solano, A., & Setubal, J. (2024). A computational pipeline for species- and strain-level classification of metagenomic sequences. In Proceedings of the 17th Brazilian Symposium on Bioinformatics, pp. 155-166. Porto Alegre: SBC.