BSB 2022

Brazilian
Symposium
on Bioinformatics
2022

Abstracts

Poster

&

Software Demonstration

Brazilian Symposium on Bioinformatics 2022
Búzios, Rio de Janeiro, Brazil

BSB 2022

The Brazilian Symposium on Bioinformatics (BSB) is an international conference with a focus on bioinformatics and computational biology, organized by the special interest group in Computational Biology (CE-BioComp) of the Brazilian Computer Society (SBC). Under the coordination of the general chair, Kele Belloze from CEFET/RJ, Brazil, the 15th BSB edition was held during September 21 - 23, 2022, at the Atlântico Búzios Convention and Resort, in the city of Armação dos Búzios (Brazil), after two years of online-only events. BSB 2022 was co-located and jointly organized with the Brazilian Databases Symposium (SBBD 2022), which took place during September 19 - 13, 2022. All BSB participants could attend the SBBD activities and vice versa.

As in previous editions, BSB 2022 had an international Program Committee, which was composed of 43 members. We received a total of 50 submissions, comprising 15 full papers, eight short papers, 25 poster abstracts, and two software demonstrations.

All submitted posters and software were presented in two exciting poster sessions. With the help of seven designated evaluators, we awarded two honorable mentions and one best poster (highlighted in the index page).

The papers have been published in the conference proceedings ***Advances in Bioinformatics and Computational Biology*** as part of the book series *Lecture Notes in Computer Science* (LNCS, volume 13523), sub series *Lecture Notes in Bioinformatics* (LNBI). All abstracts for poster presentations as well as the abstracts submitted for software demonstration have been collected in the present booklet.

Enjoy the reading!

September 2022

Nicole M. Scherer
Raquel C. de Melo-Minardi

Poster Abstracts

BSB
2022

Brazilian
Symposium
on Bioinformatics
2022

Software Demonstration Abstracts

Developing a Predictor of aggregation region in proteins with machine learning and tertiary structure

Carlos Alves Moreira, Eric Allison Philot, Ana Ligia Scott

Computational Biology and Biophysics Laboratory, Federal University of ABC - UFABC, Santo André, São Paulo, Brazil

Protein and peptide aggregation in amyloid fibrils has been associated with several clinical disorders, including Alzheimer's, Diabetes-II, and prion disease. Despite several studies, molecular mechanisms of fibril formation initiation are still unknown. Some studies showed short amino-acid segments located at amyloid precursor proteins can be a trigger to fibril development, making these regions promising targets for further studies. A current challenge in bioinformatics is the development of methods to accurately predict such regions aggregation prone. In our previous work we employed machine learning to develop a method called Magre-I, which predicts aggregation propensity regions based on primary structure. We evaluated the performances of Magre-II with proteins with experimental results available in the database Amypro. The predictor was in agreement with experimental data and displayed good/relevant performance to training and testing sets, and also for additional proteins and different intermediate structures of Prion. The Magre-II 3D approach showed to be able to distinguish between soluble and fibrillar forms considering Transferethyn, Alpha-synuclein, Insulin and Beta-amyloid protein. Our approach was also able to capture different secondary structure content present in transient prion protein conformations generated through *in silico* simulations. Another advantage of Magre-II is that we used experimental data, deposited in the Amypro data to train and validate the machine learning model of the predictor. Magre-II is open to improving the model including more experimental data from literature. We started to develop a web server where the user can submit a primary structure or a 3D coordinate structure file (PDB format) of the protein to predict aggregation prone regions.

Keywords: *Aggregation Region, Proteins, Machine Learning, Prediction*

Comparative transcriptomics for identification of key genes among xylose fermenting yeasts through interaction networks and co-expression analysis

Alexandra Cardelli [1], Gonçalo A. G. Pereira [1], Lucas M. Carvalho [1,2]

1. Department of Genetics, Evolution, Microbiology, and Immunology, Institute of Biology, Unicamp
2. Center for Computing in Engineering and Sciences, Unicamp

The use of fossil fuels to obtain energy is responsible for a large part of greenhouse gas emissions into the atmosphere. Second generation bioethanol (2G ethanol) appears as a great option because it is a cleaner and renewable source and does not compromise food security. However, there are still several bottlenecks that limit the efficiency of 2G ethanol production. Through the use of bioinformatics tools, such as differential expression analysis, co-expression and gene interaction networks of previously generated RNA-Seq and microarray transcriptomic data, we compared the metabolism of two xylose fermenting yeasts with the insertion of the XR/XDH and XI pathways under contrasting conditions of 2G ethanol fermentation (glucose vs xylose uptake). The coexpression analysis for the XR/XDH pathway was performed using three fermentation conditions: (i) only glucose medium, (ii) mixed sugar medium (including the glucose and the xylose fermentation phase), and (iii) only xylose medium. However, the coexpression analysis of the XI pathway has four collection points extracted throughout the fermentation in order to encompass all phases: (1) Predominant consumption of glucose, (2) transition between consumption of glucose and xylose, (3) predominant consumption of xylose and (4) end of fermentation. Beyond the interactomes, we found the hub genes from network topology metrics (degree, betweenness and closeness centrality) performing the union of the pairwise intersection of top 20 genes for each metric. The hub genes found in the interactome metrics analysis were then validated in the co-expression results, and it was considered a key genes linked to 2G fermentation if they had $|GS|>0.2$, $|MM|>0.5$, and were present in a module with module-sample relationship $R>0.8$ and p-value$<=0.05$. As a result, from the two possible pathways to consume xylose (XR/XDH and XI pathways), we generated both interactomes, which has a total of 479 and 571 nodes, respectively. Moreover, co-expression analysis provided us key modules related to each 2G ethanol fermentation phase of these two metabolic pathways of study (XR/XDH and XI). The enrichment of the key modules found showed processes related to response to stress, response to stimulus and nitrogen compound biosynthetic process for the XI pathway, and generation of precursor metabolites and energy, energy derivation by oxidation of organic compounds, aerobic respiration, ATP biosynthetic process, fungal-type cell wall organization, proteolysis, ribosome biogenesis, and gene expression for the XR/XDH pathway. In our analysis, for XR/XDH pathway, we found 8 key genes related to gluconeogenesis, reactive oxygen species metabolic process, protein polyubiquitination, DNA replication, recombinational repair, beta-alanine biosynthetic process, fatty acid beta-oxidation and endoplasmic reticulum to Golgi vesicle-mediated transport. Moreover, for the XI pathway, we found 10 key genes related to cellular response to oxidative stress, glutamate catabolic process, protein ubiquitination, lactate metabolic process, glucose 6-phosphate metabolic process, glycogen biosynthetic process, mitochondrial electron transport, ATP synthesis, glycolytic process, pyruvate metabolic process and proteolysis. These results can be used to understand the role of each group of genes in each 2G ethanol fermentation step through XI and XR/XDH pathways.

Keywords: *Systems Biology, Genetics, Transcriptomics, Yeast, Ethanol*

Comparative genomic and co-occurrence network analysis of plant growth-promoting genes in the genus *Paraburkholderia*

Francisnei Pedrosa-Silva[1], Thiago M. Venancio[1]

1. Universidade Estadual do Norte Fluminense Darcy Ribeiro - UENF

Paraburkholderia is a genus of versatile Gram-negative bacteria, which comprises mainly environmental isolates from different ecological niches such as water, soil, rhizosphere and legume nodules. Many *Paraburkholderia* species display several plant-beneficial properties, suggesting that the accumulation of the corresponding genes could have been selected in these genera. However, the occurrence and relationship between plant growth promoting genes (PGPG) among species remain largely unexplored. In this work, we report the comparative analysis of 235 genomes of *Paraburkholderia*, focusing on the identification and co-occurrence of genes potentially involved in plant growth promotion in light of the pangenome. Plasmids and genomic islands were predicted using Plasforest and IslandViewer4. Metrics of distance and identity between genomes were estimated with MASH and FastANI, respectively. PGPG and nodulation genes were investigated extensively using a collection of reference manually curated genes. The *Paraburkholderia* pangenome was estimated with Roary, using 75% identity threshold to determine gene clusters. CoinFinder was used to detect statistically significant associations of PGPG in the accessory genome. Phylogenetic reconstruction was performed with IQ-tree using core-genome SNPs. We found 28 phylogenetically misclassified genomes. The pangenome of *Paraburkholderia* comprises 225,192 gene clusters, with 1,884 core genes. We identified an accessory gene co-occurrence network consisting 1,194 connected components. The largest component comprises 1,279 communities of gene clusters. The community 734 consists in 212 coincident genes with 144 hypothetical proteins and 68 known genes. In this community, we identified 36 genes involved in biological nitrogen fixation (*nif*, *fix*, *mod*), auxin biosynthesis (*iaa*), ethylene stress reduction (*acdS*), nodulation (*nod*), and genes involved in plant-bacteria interactions, related with 78 genomes (36%). PGPG involved in phosphate solubilization (*pqq*) were located in most genomes (71%) of the genus, including phytopathogens. Taken together, our results contribute to the understanding of the genetic diversity of *Paraburkholderia* species and might help prospect novel strains for biotechnological applications.

Keywords: *(Bacteria, Pangenome, Bioinformatic)*

Brazilian
Symposium on
Bioinformatics 2022

Comparative genomics reveals novel species potential and niche segregation of *Campylobacter*

Sarah Henaut Jacobs[1], Hemanoel Passarelli Araújo[1,2], Thiago M. Venancio[1]

1. Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Campos dos Goytacazes, RJ, Brazil.
2. Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil.

The increase in sequenced microbial genomes has revolutionized the understanding of microbial diversity. *Campylobacter* comprises 41 validly published species currently and it is a genus of motile bacteria that can infect humans and other animals. This genus is the leading cause of many bacterial foodborne infections, especially by raw meat, contaminated water, and dairy consumption. Two *Campylobacter* species, *C. jejuni* and *C. coli*, are particularly known for their high morbidity rates, which biases the studies towards them, limiting the knowledge about the genus as a whole. Our objective was to better understand and explore genus structure and diversity. We identified 31 potentially new species. We employed Average Nucleotide Identity (ANI) values in a network analysis where genomes with 95% or more of genome identity would be linked, forming genome clusters that would represent different species of *Campylobacter*. We chose a random representative genome in each of the novel communities, so we would have a type genome in each species. We further used these representative genomes to analyze *Campylobacter*'s phylogeny through orthologous genes. Our results highlight the existence of a total of 65 *Campylobacter* genomic communities (i.e., a proxy for distinct species), which is more than the 34 validly published species available in the List of Prokaryotic names with Standing in Nomenclature by the time we downloaded the explored data, supporting the existence of 31 new *Campylobacter* species. Our results also corroborate the division of *C. concisus* in genomospecies I and II, as previously suggested. Moreover, we identified that the previously characterized *C. coli Clade III* should be split as distinct species. We also found a negative correlation between pangenome fluidity and openness, which clearly reflects different *Campylobacter* species' lifestyles. Species such as *C. coli* and *C. jejuni* were associated with higher genomic fluidity, which aligns with their generalist ecological character. Conversely, host specific species such as *C. hepaticus* showed lower pangenomic fluidity, these results show that *Campylobacter* pangenome metrics reflect their niche specification. Therefore, this study highlights that the *Campylobacter* genus division calls for a careful revision, probably with the recognition of several new species.

Keywords: *Network analysis; Phylogenomics; Campylobacter diversity*

Impact of human contamination on the patterns of co-occurrence of a microbiome

PEDRO ALMEIDA[1], HENRIQUE FRAGOSO[2], ANDRÉ TORRES[1], MARCELO GOMES[1], ERNESTO CAFFARENA[1]

1. Programa de Computação Científica da Fiocruz, Oswaldo Cruz Foundation - FIOCRUZ
2. Department of Marine Biology, Fluminense Federal University - UFF

On November 5, 2015, the Fundão tailings dam collapsed in the city of Mariana, considered the biggest environmental disaster in Brazil. This disaster severely contaminated the Doce river, whose course crosses several cities, affecting a population of more than three million inhabitants who use water for supply. The impact of the Fundão dam failure on the microbial community of the river, years after the disaster, is not yet known. The microbiome of the region is extremely important for the ecological processes of the ecosystem, as it can also contain high levels of microorganisms that are pathogenic to humans. Assessment of the microbiota of these impacted regions is an efficient way to assess water quality for human health well-being and establish microbial bioindicators. Sediment and water from three lakes in the region were analyzed using the 16S rRNA marker. We use co-occurrence network analysis, which facilitates the investigation of ecological processes that shape microbial communities. The nodes in a network, with high centrality, can be interpreted as a key species, which plays important roles in the structuring of communities. The structure of the network also reveals how a community responds to environmental disturbances. The study area where water and sediment collection were carried out in a large sampling region: The Rio Doce basin region in the state of Espírito Santo, more specifically in 4 lakes in that region, together with a point considered as a control. The studied lagoons were Lagoa do Limão (LLM), Lagoa Nova (LNV), Lagoa Juparanã (LJP) and Lagoa do Areal (LAL) and the control point Rio Guandu (RGU). To create the network, the SPARCC correlation algorithm was used. The identification of key species was used the PageRank algorithm. The analysis of the networks showed that several key species are linked with a high degree of contamination of heavy metals, thus demonstrating that they can be considered bioindicators of contamination. Some of the microorganisms observed were Pirellulaceae, which have already been reported in environments with heavy metals and mining residues in river sediments (Chen *et al.*, 2018). A genus that was observed several times as a key species was the Sphingobacterium, this bacterium has already been described as a microorganism quite resistant to heavy metals (Nam *et al.*, 2015)(In-Hyun et al 2015), in human matters some species are the cause of several respiratory infections, being resistant to several antibiotics that are used in (Lambiase *et al.*, 2009)(Freney *et al.*, 1987). More time working in these places is necessary for further development of this study.

Keywords: microorganisms, co-occurrence, heavy metals, networks, 16S

References

Chen, Y. *et al.* (2018) ‘Long-term and high-concentration heavy-metal contamination strongly influences the microbiome and functional genes in Yellow River sediments.’, *The Science of the total environment*, 637–638, pp. 1400–1412. Available at: https://doi.org/10.1016/j.scitotenv.2018.05.109.

Freney, J. *et al.* (1987) ‘Septicemia caused by Sphingobacterium multivorum.’, *Journal of clinical microbiology*, 25(6), pp. 1126–1128. Available at: https://doi.org/10.1128/jcm.25.6.1126-1128.1987.

Lambiase, A. *et al.* (2009) ‘Sphingobacterium respiratory tract infection in patients with cystic fibrosis’, *BMC Research Notes*, 2(1), p. 262. Available at: https://doi.org/10.1186/1756-0500-2-262.

Nam, I.-H. *et al.* (2015) ‘Effects of Heavy Metals on Biodegradation of Fluorene by a Sphingobacterium sp. Strain (KM-02) Isolated from Polycyclic Aromatic Hydrocarbon-Contaminated Mine Soil’, *Environmental Engineering Science*, 32, p. 150825131838000. Available at: https://doi.org/10.1089/ees.2015.0037.

Mathematical Modelling for Microbial Community Induced Metabolic Diseases (MATOMIC)

The MATOMIC Consortium, https://www.sdu.dk/en/forskning/matomic

The obesity pandemic and its comorbidities cause very high societal and economic costs and treatment options are limited (Karlsson, 2013). The intestinal microbiome is altered in obese patients, but a faecal microbiome transplantation (FMT) could be an alternative treatment to obesity (Bäckhed, 2004; Zhang, 2019). To understand how the stability of the intestinal microbiota is maintained before and after transplantation it is necessary to use a mathematical model capable of rationally designing a stable microbial community and predicting the desired metabolic interactions both qualitatively and quantitatively. The SImplified HUMan Intestinal microbiota Consortium (SIHUMIx) has established the continuous cultivation of 8 species in bioreactors under continuous controlled conditions, thus allowing the examination of pH, nutrient availability, and waste products (Becker, 2011; Schäpe, 2019). Therefore, we will develop an atom-level modelling framework capable of connecting directly with the experimental evidence available through mass spectrometry (MS) experiments. We will model the chemical reactions of the SIHUMIx microbiota using directed hypergraphs, where the reactions are the hyperarcs and the molecules are the vertices. These vertices are also graphs representing the structure of the molecules, thus making it possible to track atoms across reactions, before further examination using MS. The modelling approaches developed in SIHUMIx will be applied in more complex communities derived from the human gut and, in *in vivo* systems from bariatric surgery patients, where the effects of perturbation can be studied. Finally, the results will be transferred to FMT experiments in mice. We will examine existing genomic metabolic network software for the reconstruction of the SIHUMIx species. A total of 10 publicly available software packages were chosen and are now being compared for their completeness and capability of creating community metabolic networks.

Keywords: *gene metabolic networks; microbiome; intestine; obesity*

The work presented in this poster is supported by the Novo Nordisk Foundation grant NNF21OC0066551.

References

Bäckhed, F., Ding, H., Wang, T., et al. (2004). The gut microbiota as an environmental factor that regulates fat storage. In: *Proceedings of the National Academy of Sciences of the United States of America*, 101(44), 15718–15723.

Becker, N., Kunath, J., Loh, G., Blaut, M. Human intestinal microbiota: characterization of a simplified and stable gnotobiotic rat model. (2011) In. *Gut Microbes*. 2(1):25-33

Karlsson, F., Tremaroli, V., Nielsen, J., & Bäckhed, F. (2013). Assessing the human gut microbiota in metabolic diseases. In: *Diabetes. American Diabetes Association.*

Schäpe, S.S., Krause, J.L., Engelmann, B., et al.The simplified human intestinal microbiota (SIHUMIx) shows high structural and functional resistance against changing transit times in in vitro bioreactors. (2019) In: *Microorganisms*. 7(12):641

Zhang, M., Cai, D., Slater, D., et al. (2019). Impact of Faecal Microbiota Transplantation on Obesity and Metabolic Syndrome—A Systematic Review. In: *Nutrients*, 11(10), 2291.

Taxonomic characterization of the microbial community of photovoltaic panel using Metagenomics

Jotta, V. F. M.[1], Silva, C. A.[1], Góes-Neto, A.[3], Aburjaile, F. F.[4], Ferreira, A.M.[2], Badotti, F[2].

1. Departamento de Tecnologia de Produtos e Processos, Centro Federal de Educação Tecnológica de Minas Gerais
2. Departamento de Química, Centro Federal de Educação Tecnológica de Minas Gerais
3. Instituto de Ciências Biológicas, Departamento de Microbiologia, Universidade Federal de Minas Gerais
4. *Departamento de Medicina Veterinária Preventiva, Escola de Veterinária, Universidade Federal de Minas Gerais.*

Photovoltaic panels are installed outdoors, and therefore are exposed to various weather conditions, such as winds, extreme heat and cold, rain and drought. The use of clean energy sources has increased significantly in the last decades, and at same way, the understanding of the advantages and limitations of each system has also advanced. Recent studies (SHIRAKAWA *et al.* 2015, MOURA *et al*., 2021) demonstrated considerable energy efficiency loss of panels due to the biofilm formation. Biofilm is a complex assemblage of microbial cells associated by an extracellular polymeric substance (EPS) matrix, enabling a pleasant environment that allows the survival of several species. Thus, a greater understanding of its processes should lead to novel and effective control strategies for biofilm control. Beside, extreme environments, such as photovoltaic panels are of great interest from a biotechnological perspective. The identification of the species of microorganisms comprising the biofilm and the elucidation of the defense mechanisms will certainly open up new perspectives for the industrial sector (TANNER et al., 2019). Considering the relevance of this knowledge area and the limited number of studies available, the aim of this study was the identification and characterization of the microbial community from a photovoltaic panel located in Belo Horizonte (MG). The samples were collected in different sectors of the photovoltaic panel surface (KC60 Kyocera, 751 mm X 652 mm), and were sequenced through the Illumina *HiSeq 2500* using *shotgun* metagenomics. Low quality sequences (*Phred score* ≤ 20) were trimmed using *Trimmomatic v.0.39*. After the assembling of the contigs, the data were BLAST against *SILVA database*, considering the regions of interest (16S e 18S). About 30 prokaryotic genera were identify, being *Caloramator* the prevalent (around 30%) in all the samples. *Caloramator* is known for its wide metabolic capacity at high temperatures and has attracted interest in the field of biotechnology. Other genera identified in the samples were *Rhodococcus, Arthrobacter, Mycoplasma, Pseudomonas* and *Streptosporangium*, which are known for their extremophilic characteristics or/and biofilm formation. Regarding the eukaryotic community, about 25 genera of fungi were identify, in addition to representatives of the viridiplantae and metazoa kingdoms. Some of the fungi genera identified were Peziza, *Cheilymenia* and *Varicosporina*. Our preliminary results confirmed the high microbial diversity of the photovoltaic panels, as well as its potentiality as a source of microorganisms of biotechnological interest. In addition to the taxonomic analyses it's necessary the use of functional profile analyses for better understanding of biofilm formation mechanism.

Keywords: photovoltaic panels, metagenomics, *shotgun*, and biofilm

References

Tanner, K., Mortorell, P., Genovés, S., Ramón, D., Zacarias, L., Rodrigo, M. J., Peretó, J. and Porcar, M. (2019). Bioprospecting the Solar Panel Microbiome: High-Throughput Screening for Antioxidant Bacteria in a *Caenorhabditis elegans* Model. *Frontiers in Microbiology.* https://www.frontiersin.org/articles/10.3389/fmicb.2019.00986/full. *November.*

Shirakawa, M. A., Zilles, R., Mocelin, A., Gaylard, C. C., Gorbushina, A., Heidrich, G., Giudice, M. C., Del Negro, G. M. B. and John, V. M. (2015). Microbial colonization affects the efficiency of photovoltaic panels in a tropical environment. *Journal of Environmental Manegement.* https://www.sciencedirect.com/science/article/pii/S0301479715001875?casa_token=DZ3K1O4Lu7sAAAAA:kpsweznoVU6NkFh-dQPGIDX-ZaMymqLPniPP9GsPUW2LcC-yBo9e72kUPhAFf0vHX1aWk9hKyolz0w. *December.*

Hatmaker, E. A., Klingeman, D. M., Martin, R. K., Guss, A. M.; Elikins, J. G. (2019). Complete Genome Sequence of Caloramator sp. Strain E03, a Novel Ethanologenic, Thermophilic, Obligately Anaerobic Bacterium*. Microbiology Resource Announcements.* https://journals.asm.org/doi/full/10.1128/MRA.00708-19. *November.*

Barseghyan, G. S., Solomon, P. (2011). The genus Peziza Dill. ex Fr.(Pezizales, Ascomycota) in Israel. *Ascomycete.org.* https://www.researchgate.net/profile/Gayane-Barseghyan-2/publication/230559797_The_genus_Peziza_Dill_ex_Fr_Pezizales_Ascomycota_in_Israel/links/0912f5017978bb9496000000/The-genus-Peziza-Dill-ex-Fr-Pezizales-Ascomycota-in-Israel.pdf. *July.*

Moura, J. B., Delforno, T. P., Do Prado, P. F., Duarte, I. C. (2021) Extremophilic taxa predominate in a microbial community of photovoltaic panels in a tropical region. *FEMS Microbiology Letters.* https://academic.oup.com/femsle/article/368/16/fnab105/6350555. *November.*

Analysis of Metagenome-Assembled Genomes recovered from inocula and mature compost at the composting facility at the São Paulo Zoo Park

Suzana E. S. Guima [1,2], Layla Farage Martins [1], Aline Maria da Silva [1],
João Carlos Setubal [1,2]

1. Biochemistry Department, Institute of Chemistry, University of São Paulo
2. Bioinformatics Graduate Program at the University of São Paulo

Composting is a process in which organic matter is decomposed by microbial communities resulting into a stable product used as a fertilizer. Because its substrate is mainly composed of recalcitrant lignocellulose biomass, composting material is a source of candidate microorganisms and genes for lignocellulose degradation. In order to better comprehend the microbial composition and some metabolic functions underlying the composting process, in this study, we aim to characterize the microbial profile from the composting facility at the São Paulo Zoo Park focusing on recovering the metagenome-assembled genomes (MAGs) from different composting piles and exploring potential carbohydrate-active enzymes (CAZymes) present on these MAGs. The composting at this facility occurs at high temperatures (50 °C to 75 °C), so these composting piles are a source of thermophilic organisms and enzymes. We obtained samples from seven different composting piles: samples from five inoculum composting piles (ZCI - Zoo Compost Inoculum), and two mature composting piles (ZCM - Zoo Compost Mature). All sequences were obtained by Illumina Next Generation Sequencing, with paired reads of 250 bp. We co-assembled the preprocessed reads using metaSPAdes [Nurk et al. 2017], and used MaxBin2 [Wu, Simmons, Singer 2016], MetaBAT2 [Kang et al. 2019], and Concoct [Alneberg et al. 2014] tools to get bins, a collection of sequences that is a single microbial putative genome. We consolidated the resulting bins using the bin_refinement module from the MetaWRAP pipeline [Uritskiy, DiRuggiero, and Taylor 2018]. We considered a minimum completeness of 50% and a maximum contamination of 10% according to the standards of Minimum Information about a MAG (MIMAG) for a medium-quality MAG [Bowers et al. 2017]. We used the average nucleotide identity (ANI) to compare the obtained MAGs against the ones from the previous study (ZC3 and ZC4) [Braga et al. 2021]. CAZymes were predicted for ZCI and ZCM MAGs using dbCAN2 standalone package [Zhang et al. 2018]. Taxonomic classification was done with GTDB-tk [Chaumeil et al. 2020]. We obtained a total of 166 MAGs (about two times more from ZC3 and ZC4 than in the previous study [Braga et al. 2021]). All 166 MAGs could be classified as bacteria. The phylogenetic breakdown is the following: 69 MAGs from Firmicutes, 24 from Proteobacteria, 20 from Actinobacteriota, 13 from Bacteroidota, 4 from Chloroflexota, 3 Gemmatimonadota, 2 Myxococcota, and 31 MAGs were not classified at the phylum level. All ZCI and ZCM MAGs have CAZymes annotated in the classes glycosyl transferase, carbohydrate esterase, and glycosyl hydrolase. About 30% of all ZCI MAGs and 40% of all ZCM MAGs have CAZymes in the class Auxiliary Activities. This class covers redox enzymes and ligninases. This enlarged MAG dataset will enable us to refine the understanding of the composting process at the microbial and molecular levels, and in particular, the role played by the inoculum material. This work was funded in part by FAPESP and CAPES.

Keywords: *DNA shotgun, metagenomics, lignocellulose, CAZyme genes*

References

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., ... Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature methods*, 11(11), 1144-1146.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., ... & Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature biotechnology*, 35(8), 725-731.

Braga, L. P. P., Pereira, R. V., Martins, L. F., Moura, L. M. S., Sanchez, F. B., Patané, J. S. L., ... Setubal, J. C. (2021). Genome-resolved metagenome and metatranscriptome analyses of thermophilic composting reveal key bacterial players and their metabolic interactions. *BMC genomics*, 22(1), 1-19.

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, 36(6), 1925-1927.

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7, e7359.

Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome research*, 27(5), p.824-834.

Uritskiy, G. V., DiRuggiero, J., & Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*, 6(1), 1-13.

Wu, Y. W., Simmons, B. A., Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4), 605-607.

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., ... & Yin, Y. (2018). dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic acids research*, 46(W1), W95-W101.

Survey of *Cannabis sativa* L. RNA-Seq data available in public databases

Kevelin Barbosa Xavier[1], Thiago Motta Venancio[1]

1. Universidade Estadual do Norte Fluminense Darcy Ribeiro - UENF

Cannabis sativa L., also known as hemp, cannabis, and marijuana, has been cultivated for over 10,000 years all over the world. It is one of the most versatile crops, serving as raw material for fabrics, paper, building materials, medicines, and food production. However, cannabis is historically surrounded by taboos and prejudice. This plant is considered a drug in most parts of the world due to the hallucinogenic properties of some of its secondary metabolites, (e.g. tetrahydrocannabinol, or THC). Nevertheless, only recently the scientific initiatives to study this species have started to accelerate, following a wave of legalization of cannabis cultivation in several countries, specially for medicinal uses. Following the sequencing of several cannabis genomes, we have witnessed the publication of several transcriptomic studies. Here, we performed an extensive literature mining process to gather as many as possible cannabis RNA-Seq datasets. Preliminary analyses were conducted with the R packages bears, dplyr, ggplot2, and stringr. In total, 582 raw read sequencing files were downloaded from the NCBI SRA database, being 91.75% paired-end and 8.25% single-end sequencing. This represents 35 BioProjects comprising 15 different tissue categories. Approximately 5.84% (34 of 582) of the samples lacked cultivar/genotype information. Among the other 548 samples, we found 44 different cannabis cultivar names. The cultivar Santhica 27, which a great potential as source of textile fiber and seed, represented approximately 35% (204 of 582) of the total samples, and the medicinal cultivar Cannbio-2, which a balanced CBD:THC cannabinoid ratio production, represented approximately 11% (64 of 582) of the samples. Hypocotyl was the most abundant tissue, representing approximately 24.7% (144 of 582) of the samples. Other two most abundant tissues are trichome (21.5% of the samples) and flower (15.6% of the samples), these tissues are the most important at cannabinoid production studies. The gathered data and associated information will be used for integrative studies in our lab, which hopefully will bring insights for future genetic and biotechnological applications.

Keywords: *(hemp, marijuana, biotechnology, transcriptomic, cannabinoids)*

References

Almeida-Silva, F. and Venancio, T. (2021). bears: Building Expression Atlas from RNA-Seq data. R package version 0.99.0., https://github.com/almeidasilvaf/bears, July.

Gao, S., Wang, B., Xie, S., Xu, X., Zhang, J., Pei, L., ... & Zhang, Y. (2020). A high-quality reference genome of wild Cannabis sativa. Horticulture research, 7.

Wickham, H., François, R., Henry, L. and Müller, K. (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.9., https://CRAN.R-project.org/package=dplyr,

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. https://ggplot2.tidyverse.org, July.

Wickham, H. (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0., https://CRAN.R-project.org/package=stringr, July.

Hussain, T., Jeena, G., Pitakbut, T., Vasilev, N., & Kayser, O. (2021). Cannabis sativa research trends, challenges, and new-age perspectives. Iscience, 24(12), 103391.

Bioinformatics applied to function inference of gene subfamilies related to oil biosynthesis in legumes

Dayana Kelly Turquetti-Moraes[1], Cláudio Benício Cardoso-Silva[1], Fabricio Almeida-Silva[2], Thiago Motta Venancio[1]

1. Universidade Estadual do Norte Fluminense Darcy Ribeiro - UENF
2. Ghent University

Crop legumes are remarkable for their importance in human and animal nutrition, biofuel and industry applications. The large-scale production of some legumes, particularly soybean, largely benefits from their prominent symbiotic nitrogen fixation. Soybean [*Glycine max* (L.) Merr.] is the major crop legume world wide. It has a paleopolyploid genome that was shaped by two whole genome duplications (WGDs) that occurred 59 and 13 million years ago. As a result, 75% of the soybean genes belong to multigene families. Soybean seeds store significant amounts of protein and oil. Oil content is a quantitative trait that is significantly determined by genetic and environmental factors, challenging the identification of key genes. Although the biochemical and genetic basis of lipid metabolism is clear in Arabidopsis - a model organism, there is limited information in knowledge of lipid metabolism in soybean. To address this issue, we systematically screened the predicted proteomes of 30 leguminous species. We found 48,869 orthologous groups (OGs) using Orthofinder. We used the R package Cogeqc to assess the homogeneity of the OGs when compared with gene superfamilies (50.31%) and OGs from the Plaza database (50.33%). These results were used to estimate OG sizes using a gene birth-death model available in CAFE 5. We identified 15,494 significantly expanded/contracted OGs($p<0.05$) . Within these OGs, we found 523 genes related to lipid metabolism in soybean, 12 of which with unknown functions. As a preliminary result we observe that these genes are distributed in three OGs: OG0015097 (02), OG0000966 (01) and OG0000247 (09), enriched from InterProt as a small subunit of serine palmitoyltransferase-like (SPT-like), fatty-acid desaturase domain and O-acyltransferase, respectively. Investigating the two genes from OG0015097, we found that SPT-like catalyzes the first step in sphingolipid biosynthesis. In addition to providing structural integrity to plant membranes, sphingolipids contribute to Golgi trafficking, protein organizational domains in the plasma membrane and regulation of cellular processes. We also used our published Soybean Expression Atlas (https://venanciogroup.uenf.br/cgi-bin/gmax_atlas/index.cgi) to investigate the transcription profiles of these genes. The integration of legume OGs significantly expanded on the number of members, protein domain information, gene expression data and gene structural inference assist to infer possible functions for duplicated genes in soybean.

Keywords: *(Orthogroups, Fatty-acid, Soybean)*

References

Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*(1), 1–14. https://doi.org/10.1186/s13059-015-0721-2

Liu, N. J., Hou, L. P., Bao, J. J., Wang, L. J., & Chen, X. Y. (2021). Sphingolipid metabolism, transport, and functions in plants: Recent progress and future perspectives. *Plant Communications*, *2*(5), 100214. https://doi.org/10.1016/j.xplc.2021.100214

Machado, F. B., Moharana, K. C., Almeida-Silva, F., Gazara, R. K., Pedrosa-Silva, F., Coelho, F. S., Grativol, C., & Venancio, T. M. (2020). Systematic analysis of 1,298 RNA-Seq samples and construction of a comprehensive soybean (Glycine max) expression atlas. *The Plant Journal*, tpj.14850. https://doi.org/10.1111/tpj.14850

Mendes, F. K., Vanderpool, D., Fulton, B., & Hahn, M. W. (2020). CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics*, *36*(22–23), 5516–5518. https://doi.org/10.1093/bioinformatics/btaa1022

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H.-Y., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., … Finn, R. D. (2019). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, *47*(D1), D351–D360. https://doi.org/10.1093/nar/gky1100

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., Valliyodan, B., … Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, *463*(7278), 178–183. https://doi.org/10.1038/nature08670

Van Bel, M., Silvestri, F., Weitz, E. M., Kreft, L., Botzki, A., Coppens, F., & Vandepoele, K. (2022). PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. *Nucleic Acids Research*, *50*(D1), D1468–D1474. https://doi.org/10.1093/nar/gkab1024

Convergence, evolution and metabolism: enzymatic activities as Darwinian evolutionary units and genes encoding non-homologous isofunctional enzymes as Mendelian alleles

Fernanda Cristina de Oliveira[1], Ana Carolina Ramos Guimarães[2], Fernando Alvarez-Valín[3], Antonio Basílio de Miranda[1] & Marcos Catanho[1]

1. Instituto Oswaldo Cruz, Fiocruz, Brazil
2. Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fiocruz, Brazil
3. Unidad de Genómica Evolutiva/Sección Biomatemática, Universidad de la República del Uruguay, Uruguay

Convergent evolution is a phenomenon in which unrelated species independently evolve similar adaptations yielding similar phenotype, such as the echolocation system of bats and cetaceans, and the ability of flight in insects and birds. At a molecular level, in silico comparisons of biochemical pathways predicted in the genomes of numerous prokaryotic and eukaryotic species revealed incomplete or even absent catalytic activities. In several instances, “missing” enzymes were replaced by functional equivalent molecules, able to catalyze the same reaction but exhibiting virtually no similarity in their amino acid chains. These non-homologous isofunctional molecules, known as analogous enzymes, arose from independent evolutionary events, converging towards the same biological function in both related or unrelated phylogenetic lineages, displaying distinct catalytic mechanisms, and/or fold topologies and three dimensional structures. In this work, we demonstrate the relevance of evolutionary convergence processes in enzymatic activities, highlighting the role of analogous enzymes in the metabolism of living beings. Based on the identification of the repertoire of homologous and non-homologous isofunctional enzymes encoded in the genomes of organisms representatives of the three domains of life, and the determination of their phylogenetic profiles, we show that genes encoding analogous enzymes might behave similarly to Mendelian allele genes in a population and the enzymatic activities performed by the product of these genes might be understood as Darwinian evolutionary units.

Keywords: *convergent evolution; enzymatic activity;* evolutionary units.

Genetic variation at sites of Post-Translational Modification in the Polygenic Risk Model for Alzheimer’s disease -associated phenotypes.

Samantha L. G. Paco1,2, Michel Satya Naslavskyl1,2

1. Genetics Department, Bioscience Institute, University of São Paulo
2. Bioinformatics Graduate Program at the University of São Paulo

Alzheimer's disease (AD) is a debilitating neurological condition in which genetic variation accounts for 70% of AD cases [Breijyeh, Z. and R. Karaman 2020]. This pathology can be classified into two types: early-onset Alzheimer's disease (EOAD), with sporadic or familial cases; and late-onset Alzheimer's disease (LOAD), the most prevalent form, recently presenting a polygenic etiology [Bekris, L.M., et al. 2010]. Currently, there are dozens of AD risk loci identified by large databases of whole-genome association studies (GWAS). Nevertheless, this technique has limitations such as: difficulty in functional implications, in effect size and cellular impact of genetic variation [Visscher, P.M., et al. 2017; Insull, W., Jr 2009]. In order to overcome these issues and allow the correct distinction of variants that have neutral effects from those that induce the phenotype, this project aims to annotate genomic sites that encode amino acids linked to post-translational modifications (PTMs) related to AD. Our hypothesis is: a priori association signals can improve the risk assessment after annotation of specific functional regions of the genome. Among these regions, PTMs are the targets of this study, as they are key mechanisms in almost all physiological and biochemical processes in cells [Peng, D., et al. 2020]. In addition, this study also used a repository of genome variants of Brazilian individuals of the Brazilian Online Archive of Mutations (ABraOM),providing observations in a mixed population and analysis of the effects of different ancestries on the risk of Alzheimer's disease [Naslavsky MS, Scliar MO, Yamamoto GL, Zatz M, et al. 2022]. Hence, this proposal aimed to create a database for annotations of genes (integrative annotator) found in studies of GWAS in polygenic and monogenic forms to later investigate the effect of genetic variation in potentially target sites of PTMs in risk prediction [Ramesh, M., P. Gopinath, and T. Govindaraju 2020]. We have preliminary results on the integrative annotator, where candidate genes data were extracted from the PTMs repository (iPTMnet), transcripts and isoforms mapped, and variants in these genes obtained in databases (ABraOM and gnomAD) were annotated. In the next steps, we will expand the list to include genes identified by approaches of genomic interaction networks (BioGRID, KEGG and inBio Discover™, Encode and GTEx). The assessment of the effect on AD risk prediction will be carried out through collaborations and access to data repositories (dbGaP, EGA and UK Biobank) for comparison of Polygenic Risk Score techniques applied in the Cox regression model using the function anova() in R. The expected results are intended to provide relevant information to filter variants more strongly related to AD, revealing the consequence of the addition of the PTMs relationship factor in the PRS, PTMs types distributions as well as analyses of the differences in allele frequencies between the variants associated with PTMs and remaining variants without signal or with non-PTM signals.

Keywords: *GWAS, polygenic risk score (PRS), mixed population, integrative annotator, Alzheimer's Disease,*

References

Breijyeh, Z. and R. Karaman, Comprehensive Review on Alzheimer's Disease: Causes and Treatment. Molecules, 2020. 25(24).

Bekris, L.M., et al., Genetics of Alzheimer disease. J Geriatr Psychiatry Neurol, 2010. 23(4): p. 213- 27.

Visscher, P.M., et al., 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet, 2017. 101(1): p. 5-22.

Insull, W., Jr., The pathology of atherosclerosis: plaque development and plaque responses to medical treatment. Am J Med, 2009. 122(1 Suppl): p. S3-s14.

Peng, D., et al., PTMsnp: A Web Server for the Identification of Driver Mutations That Affect Protein Post-translational Modification. Front Cell Dev Biol, 2020. 8: p. 593661.

Naslavsky MS, Scliar MO, Yamamoto GL, Zatz M, et al., Whole-genome sequencing of 1,171 elderly admixed individuals from São Paulo, Brazil. Nat Commun. 2022 Mar 4;13(1):1004. doi: 10.1038/s41467-022-28648-3. Erratum in: Nat Commun. 2022 Mar 30;13(1):1831. PMID: 35246524; PMCID: PMC8897431..

Ramesh, M., P. Gopinath, and T. Govindaraju, Role of Post-translational Modifications in Alzheimer's Disease. Chembiochem, 2020. 21(8): p. 1052-1079.

Prediction of drug-target interactions for Brazilian herbal medicines and hallucinogens, by computational approaches

Ronald Sodre Martins[1], Marcelo Ferreira da Costa Gomes[2] and Ernesto Raul Caffarena[2]

1. Instituto Oswaldo Cruz, Oswaldo Cruz Foundation, Brazil
2. Programa de Computação Científica, Oswaldo Cruz Foundation, Brazil

The chemogenomics computational methods use chemical and biological information about drugs and therapeutic targets to predict new interactions. The independence of protein structures and the ability to explore large sets of targets and drugs simultaneously are advantages of these approaches (Ezzat et al., 2019). In addition, these methodologies allow the prediction of new interactions from a group of drugs and targets (Reker et al., 2017). There are several chemogenomics methodologies, such as techniques based on networks, machine learning, and based on matrix factorization. Thus, in the search for new therapies, it is relevant to explore the therapeutic potential of different groups of compounds and identify new drug-target interactions. In this work, we applied different chemogenomic methodologies to a group of natural products from Brazilian plants and to a group of hallucinogens. We observed differences in the results of predictions of new interactions between the methodology based on networks (Cheng et al., 2012) and the one based on matrix factorization (Zheng et al., 2013). Therefore, we proposed a heterogeneous ensemble based on the two evaluated methodologies, offering robustness to the results. As a result, we identified two new drug-target interactions from natural plant products (Martins et al., 2022). Subsequently, we applied the same strategy of the heterogeneous ensemble in the hallucinogen group, and we identified two new interactions. From the predicted interactions, it was possible to correlate the two natural products of plants with different diseases, including breast cancer, mammary tumor, and syndrome of Cushing. On the other hand, the two hallucinogens obtained from the predicted interactions were correlated with several diseases, including epilepsy and Parkinson's. Currently, we are applying the heterogeneous ensemble strategy in other drug groups to identify other potential therapeutics.

Keywords: drug-target interactions, chemogenomic methodologies, heterogeneous ensemble, network methods, matrix factorization method.

References

Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., Zhou, W., Huang, J., Tang, Y., 2012. Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference. PLoS Comput. Biol. 8, e1002503. https://doi.org/10.1371/journal.pcbi.1002503

Ezzat, A., Wu, M., Li, X.-L., Kwoh, C.-K., 2019. Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. Brief. Bioinform. 20, 1337–1357. https://doi.org/10.1093/bib/bby002

Martins, S.R., da Costa Gomes, M.F., Caffarena, R.E., 2022. Combining network-based and matrix factorization to predict novel drug-target interactions: A case study using the Brazilian natural chemical database. Curr. Bioinforma. 17, 1–1. https://doi.org/10.2174/1574893617666220820105258

Reker, D., Schneider, P., Schneider, G., Brown, J., 2017. Active learning for computational chemogenomics. Future Med. Chem. 9, 381–402. https://doi.org/10.4155/fmc-2016-0197

Zheng, X., Ding, H., Mamitsuka, H., Zhu, S., 2013. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Chicago Illinois USA, pp. 1025–1033. https://doi.org/10.1145/2487575.2487670

Multi-omic approach for characterization of tentacle and mucus composition of sea anemone *Bunodosoma caissarum*

Maria Eduarda Mazzi Esquinca[1], Leandro Mantovani de Castro [1]

1. *Biosciences Institute at UNESP – São Paulo State's Coast Campus*

Sea anemones are sessile invertebrates belonging to the Cnidaria phylum and their survival and evolutive success is highly related to the ability of producing and fast inoculating venom. Sea anemones' venom is enriched with biomolecules like polypeptides, proteins, polyamines and salts and some of these are potent toxins able to interact with sodium and potassium ion channels[1,2], blocking or reducing their sensibility, leading to diverse effects on the prey, including fast inability to move. Toxins are intensely produced in tentacles and are also frequent in the mucus[3]. *Bunodosoma caissarum* is an endemic sea anemone specie which populates the intertidal zone of the Brazil's southern coast. Some toxins were previously isolated like phospholipase A2, neurotoxins BcIII and IV and caissarolisin I.

In order to bioprospect new sequences, omic approaches provide a detailed overview of biomolecules present in all kinds of samples, so some of them were used in this study. We performed a transcriptomic analysis of tentacles and peptidome and proteome of tentacles and mucus of B. caissarum. The assembly, annotation and quality assessment of the transcriptome were performed using Trinity RNA-Seq de novo, Trinotate and BUSCO, respectively. Peptides were extracted, measured by Fluorescamine and analyzed by liquid chromatography coupled to mass espectrometry (ESI-LC/MS/MS), with eletrospray ionization (ESI)[4]. Peptidome and proteome analysis was performed using PEAKS software.

Putative toxin genes were identified with BLASTp local alignment between TransDecoder sequences and Toxprot database.Transcriptome assembly by Trinity resulted in 186.978 transcripts. Trinotate identified 23,444 annotated genes, 1% of them were characterized as toxins. BLASTp search resulted in 2055 genes with putative association with toxin activity, which were grouped in classes. Preliminary analysis from mucus and tentacle's proteome revealed 746 proteins, 26% of them uncharacterized, and validated the presence of Potassium channel type 1 and 3, Sodium channel inhibitor and phospholipase A2 toxins. Differential protein expression analysis showed a higher abundance of known toxins in mucus than in tentacles samples. In the mucus, 118 proteins were found in higher expression levels, being 16 of them toxins, whilst in the tentacles, 316 proteins were highly expressed with 6 of them identified as toxins. Furthermore, new peptide sequences with similar structure to toxins were enriched in mucus peptidome. In conclusion, integrated omics improved the understanding about tentacle and mucus composition of specie.

Keywords: Omics; sea anemone; toxins

References:

1. Diochot S, Lazdunski M. Sea anemone toxins affecting potassium channels. Prog Mol Subcell Biol. 2009; 46:99-122.

2.Moran Y, Gordon D, Gurevitz M. Sea anemone toxins affecting voltage-gated sodium channels--molecular and evolutionary features. Toxicon. 2009 Dec 15;54(8):1089-101.

3. Ramírez-Carreto S, Vera-Estrella R, Portillo-Bobadilla T, Licea-Navarro A, Bernaldez-Sarabia J, Rudiño-Piñera E, Verleyen JJ, Rodríguez E, Rodríguez-Almazán C. Transcriptomic and Proteomic Analysis of the Tentacles and Mucus of *Anthopleura dowii* Verrill, 1869. Mar Drugs. 2019 Jul 25;17(8):436.

4. Correa, Claudia Neves et al. Sample preparation and relative quantitation using reductive methylation of amines for peptidomics studies. Journal of Visualized Experiments, v. 2021, n. 177, 2021.

In embryo and *in silico* positioning of a novel gene in sensory neurons gene regulatory network

Vitória Samartin Botezelli [1], Carolina Purcell Goes [1], Ana Elisa Ribeiro Orsi [2], Chao Yun Irene Yan [1]

1. Department of Cellular and Developmental Biology, Institute of Biomedical Sciences, University of São Paulo, Brazil
2. Department of Genetics and Evolutionary Biology, Biology Institute, University of São Paulo, Brazil

The dorsal root ganglion (DRG) is a component of the peripheral nervous system that contains the cell body of sensory neurons, which are divided mainly into three subtypes: mechanoceptive, proprioceptive and nociceptive. The development of these cells depends on the sequential expression of specific genes to progress from sensory lineage commitment to subtype differentiation. This transcriptomic evolution is mediated by a complex network of transcription factors (TFs) that can act jointly, antagonistically, or in feedback loops. This network is also known as a gene cascade. Here, we focused on the individual role and relationships between TFs that are expressed during the early stages of differentiation. Specifically, we centered our studies on SCRT2, expressed only in neural postmitotic cells. This study aims to define the transcriptomic profile of SCRT2-positive cells and its role in defining the next elements in the gene cascade. We first analyzed the transcriptome of SCRT2 positive cells with the publicly available mouse DRG scRNA-seq data[1]. SCRT2 is not specifically expressed in a cell population at all the stages that we analyzed. We then searched for co-expression with other TFs (NEUROG2, ISL1, and POU4F1) whose role in specific steps of DRG development is known. The evolution of the temporal expression pattern of SCRT2 and the late commitment genes ISL1 and POU4F1 suggest a positive covariance between the three. For instance, in advanced stages, the number of SCRT2 positive cells and average single-cell expression levels decrease. The same occurs with ISL1 and POU4F1. Further, cells that express SCRT2 do not express early sensory lineage commitment genes, such as NEUROG2. Together, these data suggest that SCRT2 participates in the gene cascade that is active during late stages of sensory lineage commitment. To confirm the above co-expression results in embryo, we analyzed the spatial expression pattern of SCRT2, ISL1, POU4F1, and NEUROG2 in chick embryonic DRGs. Our *in situ* hybridization data show that SCRT2 is indeed co-expressed with ISL1 and POU4F1 and its expression domain is complementary to NEUROG2. NEUROG2, ISL1, and POU4F1 have been shown by others to participate in a neuronal genetic regulatory network, with diverse feedback connections. Thus, our next step was to verify if SCRT2 is also part of this network and could regulate the expression of NEUROG2, ISL1, and POU4F1. We then used neural tube CUT&RUN and ATAC-seq data to identify SCRT2 target sites in the genome. We identified SCRT2-target sites in the genomic region of all three genes. These sites were all located in evolutionarily conserved non-coding regions. Finally, these regions were enriched for binding sites of other TFs, suggesting that these are potential regulatory regions. Thus, with this analysis, we conclude that SCRT2 can bind to those transcription factor binding sites and regulate the other genes present in this GRN. To confirm if SCRT2 regulates the expression of ISL1 we overexpressed SCRT2 in chick embryos and counted the number of ISL1 positive cells in the DRG. Exogenous SCRT2 increased the number of ISL1 positive cells, indicating that SCRT2 indeed regulates the expression of ISL1.

Keywords: *scRNA-seq, DRG, neurons, GRN, embryo*

Funding Agencies: FAPESP and CAPES.
Ethical Committee: CEUA no 9506131021 (ICB-USP)

References

SHARMA, N. et al. The emergence of transcriptional identity in somatosensory neurons. Nature, v. 577, n. 7790, p. 392–398, 2020.

Comparative analysis and genomic diversity of the genus *Azospirillum*

Isabella Oliveira-Pinheiro[1], Francisnei Pedrosa-Silva[1], Thiago Motta Venancio[1]

1. Universidade Estadual do Norte Fluminense Darcy Ribeiro - UENF

The beneficial effects of bacteria on plant growth have led to the development of promising commercial formulations for sustainable agricultural intensification [Basu *et al*. 2021]. *Azospirillum brasilense* is one of the most well-studied and widely used bacteria in inoculants [Cassán *et al*. 2020]. However, systematic and comparative studies within the genus are remain scarce. In this work, we sought to evaluate the diversity and pathogenic potential of the genus *Azospirillum*. Genomes of *Azospirillum* spp. were selected from the evaluation of distance and identity metrics among all representatives of the *Azospirillaceae* family, using the MASH and fastANI tools. The identity between the genomes of the genus *Azospirillum* was evaluated by estimating the average nucleotide identity (ANI) with pyani. The pangenome was characterized using the Roary. The identified SNPs (single nucleotide polymorphisms) were used for phylogenetic reconstruction using IQ-TREE. The resistome and virulome of the genus were analyzed using the USEARCH tool. The evaluation with MASH and fastANI of *Azospirilaceae* allowed the reliable selection of 90 genomes of the genus *Azospirillum* and the exclusion of isolates with incorrect classification. *Azospirillum* genomes with ANI value >= 99.9% were clustered, consisting in a final set of 61 non-redundant genomes. We were able to determine 24 different phylogroups (ANI values >= 95%) for the genus, the largest group being represented by the species *A. brasilense*. The pangenome of the genus showed 72,307 genes clusters and 1,670 core genes. The virulome analysis uncovered genes related to adhesion, immune modulation, and stress resistance. The core-resistome was represented by *ceoB* gene, related with aminoglycosides and fluoroquinolones resistance. Collectively, our results constitute important steps to better understand *Azospirillum* diversity and might help prospect novel species for biotechological applications.

Keywords: virulome, resistome, biotechnology.

References

Basu, A., Prasad, P., Das, S. N., *et al.* (2021). Plant growth promoting rhizobacteria (PGPR) as green bioinoculants: recent developments, constraints, and prospects. *Sustainability*, v. 13, n. 3, p. 1140.

Cassán, F., Coniglio, A., López, G., *et al.* (2020). Everything you must know about Azospirillum and its impact on agriculture and beyond. *Biology and Fertility of Soils*, v. 56, n. 4, p. 461–479.

A multi-omic approach to agricultural pest control using *Meloidogyne incognita* as a model

Giovana Motta Oliveira[1], Maria Fernanda Zaneli Campanari[2], Marcelo Falsarella Carazzolle[2], Henrique Marques-Souza[1]

1. Brazilian Laboratory on Silencing Technologies (BLaST) - Instituto de Biologia - UNICAMP - Campinas - SP - Brazil
2. Laboratório de Genômica e bioEnergia (LGE) - Instituto de Biologia - UNICAMP - Campinas - SP - Brazil

Parasitic nematodes that attack agronomically important cultivars are responsible for immense agricultural losses worldwide. Among the different control strategies, RNA interference (RNAi) has been shown to be viable and efficient. This is possible because RNAi allows vital and essential genes to be identified, given the functional annotation of proteins and analysis of differential expression, so that the complete development of all stages of these animals' lives. Bioinformatic approaches applied to genomic and transcriptomic analyses are fundamental for the identification of essential genes. The search for target genes by transcriptomic analysis requires the development of bioinformatics pipelines. Depending on the context of the data, this development can be initiated through the de novo assembly (or reference) of the transcriptome, functional annotation of the transcripts, calculation of differential gene expression in the different experimental conditions of interest, cluster analysis of gene families in search for targets that do not have paralogs, among others. The main goal of this project is to develop a pipeline of genomics and transcriptomics data analysis for the identification of target genes for silencing. Root-knot nematodes (genus Meloidogyne) are major contributors to crop losses caused by nematodes, and reduce the productivity of crops worldwide by 5 to 10% per year (including: cotton, various vegetables, spices and coffee). These nematodes secrete effector proteins in the plant, derived from two sets of pharyngeal gland cells, to manipulate the physiology and immunity of the host. Details of how root-knot nematodes regulate transcription remain sparse. The focus of this project is the identification of essential genes in the endoparasite *Meloidogyne incognita*, through a multi-omic approach for silencing via RNAi. For this, we considered different public data of *M. incognita* genome, transcriptome of different life stages (egg, J2/3/4, female/adult, with 3 replicas) and low coverage sequencing (DNA-seq) of 11 pathogens of Brazilian cultivars from different regions of the country. Initially, we search for genes in *M. incognita* that are significantly expressed in the J2 life (the infective phase by the Meloidogyne adult female – formation of galls). We then performed phylogenomic analysis of 10 organisms phylogenetically close to *M. incognita*, to select single copy genes, to avoid function compensation after the target gene' silencing. After, were evaluated the sequence conservation of these genes among 11 Brazilian cultivars of *M. incognita* to select a set of highly conserved genes (95% of conservation). Finally, these genes were filters for essential functions in model organisms related to *M. incognita,* such as *Caenorhabditis elegans* and *Saccharomyces cerevisiae*. This analysis resulted in a list of four candidate genes (differentially expressed in J2, single copy orthologs, conserved between cultivars and essential in model organisms). We will now validate our multi-omic approach for finding essential genes as RNAi candidates by performing gene silencing assays in *C. elegans.* We believe that it is possible to apply our pipeline for agricultural pest control beyond the scope of this project, which can reduce the numerous damages caused by parasitic nematodes around the world.

Keywords: *Transcriptomic, Genomic, Agriculture, Pest, RNAi*

References

Sato, K., Kadota, Y. & Shirasu, K (2019). Plant Immune Responses to Parasitic Nematodes. *Front. Plant Sci*. 10, 1–14.

Abad, P. et al. (2008) Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. *Nat. Biotechnol.* 26, 909–915.

Basso, M. F. et al. (2020) MiDaf16-like and MiSkn1-like gene families are reliable targets to develop biotechnological tools for the control and management of Meloidogyne incognita. *Sci. Rep.* 10, 1-13.

Chaudhary, S., Dutta, T. K., Shivakumara, T. N. & Rao, U. (2019) RNAi of esophageal gland-specific gene Mi-msp-1 alters early stage infection behaviour of root-knot nematode, Meloidogyne incognita. *J. Gen. Plant Pathol.* 85, 232–242.

Maule, A. G. et al. (2011) An eye on RNAi in nematode parasites. *Trends Parasitol.* 27, 505-513 .

Perry R.N. & Moens M. (eds). 2013. Plant Nematology, Second edition. Wallingford, Oxfordshire, UK and Boston, USA, *Russian Journal of Nematology*, 2014, 22 (1), 77-82

Blanc-Mathieu, R. et al. (2017) Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes. *PLoS Genetics* vol. 13.

What can unmapped reads uncover about sugarcane during biotic stress

Leite-Junior, J.N.*; Miranda, R. P.[1]; Zerillo, M. M.; Dias[1], H. M.; Van Sluys, M. A.[1]

1. Universidade de São Paulo

RNA-Seq technology allows identifying which genes are being transcribed at a specific time point and comparing their expression levels under different conditions. Usually, the "mapping first" methodology is used to filter the reads aligned in the references, although it is possible to proceed with the "assembly first", which forms contiguous sequences with the *de novo* assembly. In present study, total RNA was extracted from a susceptible cultivar of sugarcane subjected to different biotic stresses with leaf scald causing agent. Subsequently, the transcripts were sequenced by the Illumina HIGH-SEQ platform and aligned, respectively, to the references of the study comprising: rRNA, organelas genes, and bacterial genes. After the last filtering, different outputs of unmapped reads were generated, which were used as input in Trinity de novo assembly of RNA-Seq data (Galaxy Version 2.9.1) to obtain contigs. These sequences were sent to a database with genomic sequences for identification and subsequent selection. After using different packages in R Studio, it was seen that several contigs had high TPM values under all conditions. Interestingly, alignments of these sequences were identified in Saccharum and Non-Redundant taxa, including members of the Poaceae family. In addition, similarities were identified of these sequences in sugarcane transposable elements identified in the study by (DOMINGUES et al., 2012). Furthermore, our analyzes show that there is a differential distribution of *Acinetobacter soli* aligned contigs between treatments, suggesting a relationship between their proliferation and the designed biotic stress. Further analyses are necessary for data curation and validation, but we conclude that our pipeline for unmapped reads presents to be a relevant strategy to uncover information from transcripts that would otherwise be unnoticed remain noticed, in addition to showing the microenvironment within the sugarcane and its variation between treatments.

Keywords: *RNA-Seq, sugarcane, unmapped reads, microbiome, lncRNA*

Reference

Domingues, D.S., Cruz, G.M., Metcalfe, C.J. et al. (2012). Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. In: *BMC Genomics* 13, 137.

PHEMALE-Bacteria: genome-based tool for prediction of bacteria phenotypes with machine learning

Bruno Koshin Vázquez Iha[1], Carlos Morais Piroupo[1], João Carlos Setubal[1]

1. Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, 05508-000, Brazil

Bacteria can live in a wide range of environmental conditions, from the human gut to boiling water deep in the ocean. Most bacterial species have not been cultivated in the lab. Creating the optimal conditions for some species to thrive in a lab setting may require expensive equipment, expensive materials or several tests with varying environmental conditions. This consumes time and/or resources. Nowadays, due to advances in DNA sequencing technology a bacterial DNA sequence is easily and cheaply obtained. Therefore a tool that can predict bacterial phenotypes from bacterial genomes should be of great help in the search of adequate conditions for bacterial cultivation. In addition, there is a growing number of bacterial Metagenome-Assembled Genomes (MAGs) for which we only have the genome sequence; the ability to predict phenotypes for MAGs would be helpful in understanding the role the corresponding bacteria have in their respective environments. Current tools based on machine learning techniques that can predict bacterial phenotypes based on bacterial genomes (Feldbauer et al., 2020; Weimann et al., 2016) have limited scope. They can predict phenotypic classifications (e.g. if a bacteria is Gram negative or Gram positive, or if it is thermophilic or psychrophilic) but cannot predict numerical phenotypes (e.g. range of pH of the environment that a bacteria can withstand) due to limitations of the machine learning technique chosen.

Here we present PHEMALE-Bacteria (PHEnotype MAchine LEarning - Bacteria), a tool capable of predicting both classification and numerical bacterial phenotypes based on the annotated genome sequence. The tool implements multiple machine learning techniques, ranging from linear regression to deep learning. To train the machine learning models used in the tool the phenotypic information was obtained from Madin et al. (2020), which includes data on more than 15000 species of bacteria and archaea, and 4576 bacterial genomes were obtained from the NCBI database (Benson et al., 2012).

Preliminary tests achieved precision higher than 97% of F1-macro for sporulation (whether the species sporulates or not) and higher than 0.80 of R^2 for prediction of pH ranges that the species can withstand. Based on these promising results we plan to include additional phenotypes for prediction.

This work was funded in part by a CAPES fellowship to BKVI.

Keywords: metagenomics, bacteria, phenotype, machine learning

References

Benson, D. A., Cavanaugh, M., Clark, K. et al. (2012). GenBank. Nucleic Acids Research, 41 (D1)

Feldbauer, R., Hyden, P., Lüftinger, L. (2020). Phenotrex. https://github.com/univieCUBE/phenotrex, July 21st, 2022

Madin, J.S., Nielsen, D.A., Brbic, M. et al. (2020). A synthesis of bacterial and archaeal phenotypic trait data. Scientific Data 7, Article 170

Weimann, A., Mooren, K., Frank, J. et al. (2016). From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer. mSystems, Nov-Dec 1(6)

Contact profile optimization: a novel approach for rational design of inhibitory peptides against sars-cov2

Ana Paula de Abreu[1,3] ,Frederico Chaves Carvalho[3], Adriano de Paula Sabino[2], Raquel Cardoso de Melo-Minardi[3]

1 Institute of Biological Sciences, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
2 Faculty of Pharmacy, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
3 Institute of Exact Sciences, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

The Sars-CoV-2 pandemic has prompted an urgent demand for effective therapeutics and vaccines to control and reduce this outbreak. Peptide-based therapeutics are gaining increasing attention due to their intrinsic advantages, such as lower toxicity, higher specificity, and greater ability to interfere with protein-protein interactions. Here we propose an algorithm to iteratively optimize peptides using docking protocols and contact analysis to achieve better amino acid contact profiles.

Our algorithm works by iteratively replacing amino acid residues in a given peptide to ensure favorable interactions. To do so, at each iteration, our protocol (1) performs docking simulations using Rosetta; (2) counts and classifies the contacts using VTR (http://bioinfo.dcc.ufmg.br/vtr/); (3) replaces peptides that make unfavorable interactions by residues that make favorable interactions; (4) repeats until no change can occur. Alternatively, step 1 can be performed manually by using any docking webserver such as HPEPDOCK.

Using the 6M0J structure as a case study, we optimized 18 peptides to achieve high binding specificity to the RBM of the Sars-CoV-2 spike protein. We started with ten peptides from the literature and eight peptides designed based on the ACE2 interface of the RBM. The peptides generated by our algorithm for contact optimization achieved higher scores and established more contacts with the desired binding site according to two different docking protocols: HPEPDOCK(manual) and ROSETTA (automatic). The contact profile was also more favorable, presenting more attractive contacts and less repulsive ones.

After optimization, the 18 best-performing peptides will be synthesized and have their binding affinity and ability to inhibit the Sars-CoV-2 RBD/hACE2 interaction experimentally evaluated. Currently, a web tool is being developed to ensure reproducibility.

Keywords: *Optimization algorithms, protein-protein interaction, peptides, SARS-CoV-2.*

Genome assembly of *Vellozia tubiflora* and *V. peripherica*: A story about charting unknown genomes in the era of Big Data

Felipe Eduardo Ciamponi[1], Mariana Feitosa Cavalheiro[1], Stephen Richards[3], Harris Lewin[3], Paulo Arruda[1,4], Ricardo Augusto Dante[1,2], Isabel Rodrigues Gerhardt[1,2]

1. *Genomics for Climate Change Research Center (GCCRC) - University of Campinas*
2. *Embrapa Digital Agriculture*
3. *University of California, Davis*
4. *Center for Molecular Biology and Genetic Engineering (CBMEG) - University of Campinas*

A high-quality chromosome-level genome scaffold is of paramount importance to drive innovation and new scientific knowledge for non-model organisms. However, chromosome-level scaffolding and accurate genome annotation from plant drafts is a notoriously difficult problem, especially when dealing with regions that overlap sections with many differences between haplotypes with homozygous regions. Although the use of modern algorithms, long reads, and complementary techniques such as Hi-C has greatly improved the overall accuracy and efficiency of assembly pipelines, this step remains a computationally challenging problem. A variety of techniques have been developed to solve the scaffolding problem, each with its own weaknesses and strengths, although using a combination of multiple scaffolding methods is generally better for generating chromosome-level assemblies than using a single approach. While understanding the genomic features of non-model organisms can be critical for new scientific discoveries, especially in the context of responses to stress conditions, the panorama of genome assemblies available for the Velloziaceae family is dismal. Only two publicly available genomes (one for Acanthochlamys bracteata and one for Xerophyta viscosa) are currently available to the scientific community. In this project, we scaffolded two draft genomes from different species of the genus Vellozia: Vellozia tubiflora and Vellozia peripherica. These endemic species of the Brazilian Campos Rupestres can thrive even under adverse environmental conditions, a phenotype achieved after millions of years of natural selection. By deciphering the genomic traits associated with response to various stresses, we can transfer the molecular basis of stress tolerance to other economically relevant plants. Considering that crops are among the most important commodities in Brazil, even small savings can have impacts on the order of millions of dollars per year. We not only performed a direct comparison of eight different scaffolding algorithms using reference-based and de novo strategies, but also evaluated the performance of hydride strategies using existing genome references and Hi-C data to generate high-quality chromosome-level. Our results show that by combining high-throughput experiments with the use of genomes already available in public databases, we were able to overcome the most common problems in genome assembly of novel organisms. By producing a high quality genome with structural and functional annotations, we are able to provide a solid foundation for any "omics" projects that might be undertaken by the scientific community interested in studying plant adaptation to various stresses.

Keywords: *Genomics; Vellozia; Campos rupestres; Scaffolding*

References

Bailey-Serres, J., Parker, J. E., Ainsworth, E. A., Oldroyd, G. E. D. and Schroeder, J. I. (nov 2019). Genetic strategies for improving crop yields. *Nature*, v. 575, n. 7781, p. 109–118.

Carballo, J., Santos, B. a. C. M., Zappacosta, D., et al. (15 jul 2019). A high-quality genome of Eragrostis curvula grass provides insights into Poaceae evolution and supports new strategies to enhance forage quality. *Scientific Reports*, v. 9, n. 1, p. 10250.

Costa, M.-C. D., Artur, M. A. S., Maia, J., et al. (27 mar 2017). A footprint of desiccation tolerance in the genome of Xerophyta viscosa. *Nature Plants*, v. 3, p. 17038.

Gao, Z.-Y., Li, Z.-H., Lin, D.-L. and Jin, X.-H. (3 aug 2021). Chromosome-Scale Genome Assembly of the Resurrection Plant Acanthochlamys bracteata (Velloziaceae). *Genome Biology and Evolution*, v. 13, n. 8, p. evab147.

Ghurye, J. and Pop, M. (5 jun 2019). Modern technologies and algorithms for scaffolding assembled genomes. *PLOS Computational Biology*, v. 15, n. 6, p. e1006994.

Istace, B., Belser, C., Falentin, C., et al. (30 jul 2021). Sequencing and Chromosome-Scale Assembly of Plant Genomes, Brassica rapa as a Use Case. *Biology*, v. 10, n. 8, p. 732.

Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D. and Strömvik, M. V. (2018). Current Strategies of Polyploid Plant Genome Sequence Assembly. *Frontiers in Plant Science*, v. 9.

Luo, J., Wei, Y., Lyu, M., et al. (1 sep 2021). A comprehensive review of scaffolding methods in genome assembly. *Briefings in Bioinformatics*, v. 22, n. 5, p. bbab033.

Montagu, M. V. (20 dec 2019). The future of plant biotechnology in a globalized and environmentally endangered world. *Genetics and Molecular Biology*, v. 43.

Pucker, B., Irisarri, I., Vries, J. De and Xu, B. (ed 2022). Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. *Quantitative Plant Biology*, v. 3.

Silveira, F. A. O., Negreiros, D., Barbosa, N. P. U., et al. (1 jun 2016). Ecology and evolution of plant diversity in the endangered campo rupestre: a neglected conservation priority. *Plant and Soil*, v. 403, n. 1, p. 129–152.

Identification and characterization of OSCA gene family in *Vellozia intermedia* and *V. nivea*

Mariana Feitosa Cavalheiro[1], Lucas Eduardo Canesin[1], Diego Maurício Riaño-Pachón[2], Paulo Arruda[1,3], Ricardo Augusto Dante[3,4], Isabel Rodrigues Gerhardt[3,4]

1. *Genomics for Climate Change Research Center (GCCRC) - University of Campinas*
2. *Center for Nuclear Energy in Agriculture - University of São Paulo*
3. *Center for Molecular Biology and Genetic Engineering (CBMEG) - University of Campinas*
4. *Embrapa Digital Agriculture*

The Brazilian campos rupestres, a known hotspot of biodiversity, is an ecoregion that suffers from long periods of drought. Plants that thrive in this environment have evolved different strategies to cope with hydric stress, notably those from the Velloziacea family. In order to understand mechanisms involved in different strategies associated with the perception of drought stress, we selected two hallmark species from the *Vellozia* genus: *V. nivea*, which is desiccant-tolerant, and *V. intermedia*, an evergreen plant. To explore the molecular landscape of these phenotypes, we decided to study and characterize in each species the OSCA channels, which are mechanosensitive channels responsible for the [Ca2+]i increase induced by osmotic stress in plants. Differences in OSCA family composition could indicate that different drought tolerance strategies emanate from drought perception. The conserved OSCA domain (PF02714) was extracted from the Pfam database to create a hidden Markov model profile and scanned against our protein dataset for each species using HMMER. All sequences found were manually curated using InterProScan to confirm the presence of the PF02714 domain. We found 18 proteins encoded by 16 genes in *V. intermedia* and 16 proteins encoded by 15 genes in *V. nivea*. After the OSCA channels are distributed among four different clades, to classify the OSCAs sequences we found in the previous step, we tested and selected an evolutionary model and constructed a phylogenetic tree with our sequences and the sequences classified as OSCAs for *Arabidopsis thaliana*, *Zea mays*, *Sorghum bicolor*, and *Oryza sativa* using the IQ-TREE. According to this approach, both *V. nivea* and *V. intermedia* have five genes encoding OSCA1 proteins, seven genes encoding OSCA2, two genes encoding OSCA3, but with respect to OSCA4, while *V. nivea* has one gene, *V. intermedia* has two. By applying state-of-the-art comparative genomics strategies, we were able to trace the evolutionary path of this gene family across the studied species and identify the points of origin of the OSCA family, especially those belonging to clades 1 and 2. Deciphering the genomic composition of the OSCA family can provide new insights into how wild plants respond to drought, which is a hot topic in a world where climate change is pushing the environment towards increasingly hostile conditions for crops.

Keywords: *Comparative genomics; osmosensor channels; drought tolerance;*

Cosmopolitism Analysis of Metagenome-Assembled Genomes from São Paulo Zoo environments

Robson Pontes de Oliveira[1,2], Fabio Beltrame Sanchez[2], Layla Farage Martins[1], Aline Maria da Silva[1,2], João Carlos Setubal[1,2]

1. Biochemistry Department, Institute of Chemistry, University of São Paulo
2. Bioinformatics Graduate Program at the University of São Paulo

Metagenome-Assembled Genomes (MAGs) have been important objects of study in the understanding of entire microbial communities in specific environments. According to Setubal (2021), MAGs can be classified in two ways: Species MAG (SMAG) – a MAG classified at the species level and for which there is at least one genome of isolate as a reference – or Hypothetical MAG (HMAG) – a MAG classified at less specific taxonomic levels and without a genome of an isolate as a reference. Furthermore, for HMAGs, if there is at least one other sufficiently similar genome from a study in a different geographic location, we say that it is a Conserved Hypothetical MAG (CHMAG). Given the growing number of publicly available genomes, as well as the publication of MAG catalogs, in this work we aimed to characterize SMAGs and CHMAGs recovered from three different environments at the São Paulo Zoo (composting [Braga et al. 2021], howler monkey feces [Franco et al., unpublished], and reservoir water [Barbosa et al., unpublished]), comparing them to sufficiently similar genomes, which we call “orthologous genomes”. We classified these MAGs as SMAGs or HMAGs based on results obtained with GTDB-tk (Chaumeil, 2020). Then, using Average Nucleotide Identity (ANI) as a metric, our SMAGs were compared with genomes deposited in GenBank and JGI, and with MAGs from the Genomes from Earth’s Microbiomes (GEM) (Nayfach, 2021) and Non-Human Primates Gut Microbiome (NHPGM) (Manara, 2019) catalogs. Here we consider orthologous genomes those pairs of genomes for which ANI ≥ 95%. The HMAGs were also compared with the genomes of the catalogs, so we were able to determine that, of the 150 HMAGs in the three environments, 34 are actually CHMAGs. Many orthologous genomes found for MAGs from howler monkey feces are from the human digestive system, while others were recovered from the gut microbiota of both Old World and New World monkeys. For water reservoir MAGs, all orthologous genomes come from freshwater samples, except ZLKRG40 (Burkholderiaceae), which has an ortholog recovered from a saline aquatic environment. Most orthologous genomes of the composting samples were also obtained from composting in other locations or from environments where biomass degradation takes place. We compared orthologous genomes in a set to each other using the MAGset tool, which uses the concept of Genomic Region of Interest (GRI), which is a region of at least 5,000 bp with differential presence between a target genome (our MAG) and one or more reference orthologous genomes. The GRIs existing only in target SMAGs were analyzed to find genes that could indicate adaptations to their environment. Currently, seven SMAGs are being analyzed: ZC3RG05 (*Thermobifida fusca*), ZC3RG08 (*Planifilum fulgidum*), ZC4RG04 (*Thermobispora bispora)*, ZC4RG08 (*Pseudomonas thermotolerans*), ZC4RG13 (*Rhodothermus marinus*), ZC4RG43 (*Mycobacterium hassiacum*) and ZC4RG45 (*Thermocrispum agreste*). All of them are from composting. As an example of this analysis, one GRI gene in ZC4RG04 encodes a GH3 CAZyme family protein that performs a range of functions including cellulosic biomass degradation. Acknowledgments to FAPEMA, which funded part of this work.

Keywords*: metagenome, orthologous genomes, composting, howler monkeys, reservoir water*

References

Braga, L., R.V. Pereira, L.F. Martins, L. Moura, F.B. Sanchez, J.S.P. Patané, A.M. da Silva, J.C. Setubal. Genome-Resolved Metagenome And Metatranscriptome Analyses Of Thermophilic Composting Reveal Key Bacterial Players And Their Metabolic Interactions. In *BMC Genomics*, 22:652, 2021.

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., Parks, D. H. (2020). GTDB-Tk: A Toolkit To Classify Genomes With The Genome Taxonomy Database. In *Bioinformatics*, pages 1925-1927.

Manara, S., Asnicar, F., Beghini, F., Bazzani, D., Cumbo, F., Zolfo, M., Nigro, E., Karcher, N., Manghi, P., Metzger, M. I., Pasolli, E., & Segata, N. (2019) Microbial Genomes From Non-Human Primate Gut Metagenomes Expand The Primate-Associated Bacterial Tree Of Life With Over 1000 Novel Species. In *Genome Biology*, pages 1-16.

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I. M., Huntemann, M., Palaniappan, K., Ladau, J., Mukherjee, S., Reddy, T. B. K., Nielsen, T., Kirton, E., Faria, J. P., Edirisinghe, J. N., Henry, C. S., Eloe-Fadrosh, E. A. (2021). A Genomic Catalog Of Earth's Microbiomes. In *Nature Biotechnology*, pages 499-509.

Setubal, J.C. (2021). Metagenome-Assembled Genomes: Concepts, Analogies, And Challenges. In *Biophysical Reviews*, pages 905-909. Publishing Press.

Bacterial 2′-Deoxyguanosine Riboswitch Classes as Potential Targets for Antibiotics: A Structure and Dynamics Study

Deborah Antunes[1], Lucianna Helene Santos[,2], Ernesto Caffarena3, Ana Carolina Guimarães[1]

1. Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz
2. Biomolecular Simulation Group, Institut Pasteur de Montevideo
3. Grupo de Biofísica Computacional e Modelagem Molecular, Programa de Computação Científica, Fiocruz

The proliferation of germs that are resistant to antibiotics poses a significant risk to public health. The majority of antibiotics used now only affect a small number of metabolic pathways, which promotes antibiotic resistance. As a direct result of this, innovative regulatory inhibitory methods are required. Riboswitches are an interesting potential target for the development of antibacterial medicines. Purine riboswitches are an intriguing topic to discuss due to the crucial functions that they play in the genetic regulation of metabolic processes in bacteria. Class I (2′-dG-I) and class II (2′-dG-II) riboswitches are examples of the several types of 2′-deoxyguanosine (2′-dG) riboswitches that are involved in the regulation of deoxyguanosine metabolism. On the other hand, a high affinity for nucleosides requires either local or distal changes around the ligand-binding pocket, and this varies from class to class. Therefore, it is absolutely necessary to get an understanding of the recognition mechanisms utilized by these riboswitches while acting as antibiotic targets. In this study, we investigated the structure, dynamics, and energy landscape of both classes of 2′-dG that were attached to the nucleoside ligands, 2′-deoxyguanosine and riboguanosine, using a variety of computational biophysics techniques. According to the findings of this study, a higher nucleoside ligand affinity is associated with enhanced stability as well as an increase in the number of contacts that take place in the three-way junction where 2′-dG riboswitches are located. Additionally, alterations in the structure of the 2′-dG-II aptamers make it possible for improved intramolecular communication. In general, the fact that the 2′-dG-II riboswitch is able to identify cognate as well as noncognate ligands makes it a potentially useful target for the design of drugs.

Keywords: *2′-deoxyguanosine riboswitch; purine riboswitch; molecular dynamics simulation; ligand binding mechanism*

BSB 2022
Brazilian
Symposium on
Bioinformatics 2022

Creating PERCI data repository of non-coding RNAs involved in colorectal cancer with search application of BERT.

Jaqueline Gutierri Coelho[1], Maria Emília Machado Telles Walter[1,2], Li Weigang[1],João Batista de Sousa[1], Maristela Terto de Holanda[1], Lucas Maciel Vieiras[1,2]

1. Universidade de Brasília
2. Universität Leipzig

Collecting information about colorectal cancer (CRC) and visiting it in a centralized and reliable way can help CRC research, especially help researchers to understand the mechanisms that may help them understand the cellular mechanisms that contribute to tumor appearance, cancer progression, or even promote its prognosis. In this background, this study aims to collect available information about ncRNAs related tocolorectal cancer and provide it to the public for consultation. We implemented a PERCI data repository online (https://www.percidatabase.com) provides information on five CRC subtypes and three specific types of ncRNAs. The types of cancer include colorectal cancer, colon cancer, adenocarcinoma, tumor and hepatic colorectalcancer metastasis. The three types of ncRNAs involved in CRC are long ncRNAs (lncRNAs), long circular ncRNAs (circ nc-RNAs) and micrornas (miRNAs), as well assome genomic characteristics. "Which lncRNAs are associated with adenocarcinoma?", "Given a specific circRNA, which cancer subtypes are involved in this circRNA?", or other questions can be answered by usingour database. This project reviews the applications of machine learning (such as BERT) in optimization research, and puts forward some suggestions for the PERCI database. In the successful development of Perci database, we intend to continue to complete the cancer genome sequences, including clinical data of cancer patients. It also intends to include ceRNAs networks, which link lncRNAs, miRNAs and specific, cancer-linked proteins in the database for finding and proposing ceRNAs in silico, to be further proven in the laboratory. Further more, we propose to apply artificial intelligence (specifically Transformer- based machine learning method as BERT) into the workings of the database using MER so that it can find more efficient searches and faster, more agile database models.

Keywords: *database, artificial intelligence, bioinformatics, long non-coding RNAs, BERT*

Mini-Docking – A simple script to fast execute the Virtual Screening Process

Lucas Rolim Medaglia[1], Tiago Alves de Oliveira[1,2,*], Eduardo Habib Bechelane Maia[2], Alisson Marques da Silva[2], Alex Gutterres Taranto[1]

1. Department of Bioengineering, Federal University of São João del-Rei - UFSJ, Praça Dom Helvécio, 74, Fábricas, São João del-Rei, MG, Brazil
2. Department of Informatics, Management and Design, Federal Center for Technological Education of Minas Gerais - CEFET-MG, Campus Divinópolis, Rua Álvares de Azevedo, 400, Bela Vista, Divinópolis, MG, Brazil
** Software Demonstration Presenter*

Virtual Screening (VS) is one of the latest advances supporting new drug discoveries. The *in-silico* technique allows a better selection of molecules with the desired chemical characteristics, reducing characteristics, and reducing the number of experiments, including in animals, to search for the most probable drug candidates among an extensive library of compounds. Generally, proprietary software for VS has a high cost, which makes its use unfeasible in some scenarios. On the other hand, the free tools used for this purpose are often complicated and do not even have a graphical interface. In this context, this study presents Mini-Docking, free software with a simple and intuitive interface that automates VS experiments. Mini-Docking integrates the following tools: Dock6, MGL Tools, Chimera, and Autodock Vina. In contrast to other platforms, Mini-Docking can simulate fitting an unlimited number of compounds into a set of molecular targets. The main advantages of Mini-Docking over Dock6, MGL Tools, Chimera, and Autodock Vina are its automation, ease of use, speed, and error reduction. However, a human operator must manage the four programs mentioned if Mini-Docking is not used. Furthermore, the output of one program must be used as input to the next program, which frequently requires user action, introducing delays in the VS process and the possibility of user-generated errors. Therefore, there is also a need to check for human mistakes at every step that requires user action. In addition, the user needs to perform user action steps for each target ligand combination. However, in Mini-Docking, as soon as one of the programs completes, the next one runs automatically without user intervention. Consequently, Mini-Docking reduces the possibility of user-generated error because it reduces human interactions. In the Mini-Docking workflow, the Chimera is used to prepare molecules. In the next step, the MGL Tools determine the rotation ligations and bonding and assign Gasteiger-Marsili net atomic charges. Then, the Autodock Vina performs the first molecular docking, with your output used as input of Dock6. So, Dock 6 executes molecular docking with the production of Dock6. The results are computed and presented to the user in the last step. In conclusion, the Mini-Docking has a simple workflow that requires minimal user action to perform in the VS process since the molecular preparation for the docking process. Therefore, it helps the user automate the process and minimizes errors during the process. It also avoids the need to know diverse file formats and forms of execution of the programs used in Mini-Docking. At last, Mini-Docking is available for download free of charge at https://www.drugdiscovery.com.br/software/.

Link: https://www.drugdiscovery.com.br/software/

Keywords: *Virtual Screening, Docking, Molecular Modeling, Structure Based-Drug Design.*

OLATCG: expanding the frontiers for the use of Bioinformatics

Luiz Miguel Viana Barbosa[1], Anna Carolina de Oliveira Mendes[2], Guilherme Inocêncio Matos[3] and Kele Teixeira Belloze[3]

1. Federal University of Rio de Janeiro (UFRJ)
2. Osorio Foundation
3. Federal Center for Technological Education of Rio de Janeiro (CEFET/RJ)

The Education area is constantly concerned about exploring and offering new teaching modalities to stimulate students even more towards understanding new content and skills' domain. In an increasingly technologized world, using computational tools such as software or games has become the formal academic curriculum topic since elementary years. Particularly in Biology science, new contents using Bioinformatics tools are growingly important in current scientific production and must be incorporated into the school context. Nevertheless, its use is somehow limited, mainly due to the languages adopted for its dissemination. On the one hand, scientific language relies on many technical terms. On the other, programming languages require minimum knowledge to execute such tools. These kinds of languages are not easily understood by high school students nor school-age youths in general. Therefore, the challenge is establishing mechanisms to allow greater dissemination of bioinformatic tools, narrowing the gap in their understanding. Given the above, this project aims to enable data understanding and the processes in the Biology field by developing a teaching platform designed for high school students. Within this context, we present "OLATCG" (https://olatcg.herokuapp.com), a didactic tool to serve as a base for a teaching strategy in Biology Classes. The platform allows teachers to work with problem-based learning activities applied to different Biology fields: Genetics, Evolution, Taxonomy, and others. This tool has controlled data sets to facilitate teacher's interaction in problem-solving situations, thus instigating classroom discussions. "OLATCG" has been able to perform alignments between genetic sequences and phylogenetic analyses on its first version. All content is presented through user-friendly interfaces and explanations in easy-to-understand everyday language for students. The platform also provides introductory concepts on Bioinformatics and a tutorial snippet on how to use all its resources facilities. We intend the platform to provide not only the expansion of Bioinformatics at the secondary level of education but also the strengthening of a developing network to foster Bioinformatics teaching. This network has several national and international research and educational institutions, such as the "Helmholtz for Environmental Research (UFZ)" Center in the German city of Leipzig, focusing on multidisciplinary environmental research. As a result, OLATCG is expected to encourage the collective construction of content for the popularization, appreciation, and didactics of Bioinformatics related areas.

Keywords: *Bioinformatics; popularization; teaching; software tool*